High-Frequency Trading and Market Stability*

Dion Bongaerts†and Mark Van Achter‡

March 2015

Abstract

In recent years, technological innovations and changes in financial regulation induced a new set of liquidity providers to arise on financial markets: high-frequency traders (HFTs). HFTs differ most notably from traditional market participants in the fact that they combine speed and information processing. We compare a setting with HFTs to settings with traders that only have speed technology or only information processing technology available. Speed technology by itself will only be adopted when socially efficient. Information processing technology by itself will only generate mild inefficiencies due to a lemons problem. The combination of the two, however, can lead to the implementation of inefficient speed technology or the amplification of the lemons problem. In the latter case, liquidity evaporates when it is most needed and markets can freeze altogether for periods of time. We also discuss how regulation can prevent such sudden drops of liquidity and how the market may recover after a freeze.

JEL Codes: D53, G01, G10, G18

Keywords: High-Frequency Trading, Limit Order Book, Market Freeze, Market Stability

*We would like to thank Jean-Edouard Colliard, Hans Degryse, Jerôme Dugast, Frank de Jong, Thierry Foucault, Nicolae Garleanu, Terry Hendershott, Johan Hombert, Katya Malinova, Sophie Moinas, Christine Parlour, Ioanid Roşu, conference participants at the 2014 FIRS annual meeting, the 2014 EFA annual meeting and seminar participants at Erasmus University Rotterdam and HEC Paris for helpful comments and suggestions. Mark Van Achter gratefully acknowledges financial support from Trustfonds Erasmus University Rotterdam.

†Rotterdam School of Management, Erasmus University, Department of Finance, Burgemeester Oudlaan 50, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. E-mail: *dbongaerts@rsm.nl.*

‡Rotterdam School of Management, Erasmus University, Department of Finance, Burgemeester Oudlaan 50, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. E-mail: *mvanachter@rsm.nl.*

1 Introduction

In recent years, technological innovations and changes in financial regulation (e.g. Regulation NMS in the United States and MiFiD in Europe) have induced trading to become more automated. This development has drastically altered the nature of liquidity provision on financial markets. More specifically, traditional intermediaries have been complemented or even replaced by a new set of liquidity providers: high-frequency traders (HFTs). HFTs invest heavily in trading technology allowing them to benefit from a combination of low-latency access to the financial market (i.e., "speed") and superior information processing.[1] In particular, they use automated algorithms to scan (order book) information at an extremely fast rate and instantly form trading decisions. Co-location near the market server assures these decisions are transferred to the market in microseconds. In order to exploit their speed advantage as much as possible, HFTs compete for low latency amongst each other (e.g. for an optimal co-location near the market server). In parallel, trading venues have been very active in setting up policies to attract HFTs (e.g. through offering beneficial pricing policies, co-location opportunities or privileged information access mechanisms) in order to increase turnover. For 2012, HFTs were involved in an estimated 55 percent of all daily US equity trading volume and 45 percent of all daily European equity trading volume (Tabb Group, 2012).

Meanwhile, the massive participation of these new "middlemen" in trades across the globe spurred an intense public debate on the desirability of HFTs. This debate was fueled further by the May 2010 "flash crash" which featured an unprecedented vicious liquidity spiral causing US equity markets to instantly dry up and the major index to temporarily decrease by more than 9% (corresponding to \$1 trillion in market value evaporating).[2] In recent years, markets allegedly have become more susceptible to technology-related incidents. Especially the increasing incidence rate of "mini flash crashes" has been linked by many market observers to the emergence of HFTs.[3] Hence, policy makers and regulators have become increasingly concerned that HFT-based liquidity provision could come at the expense of an evaporation of liquidity when it is most

[1]Latency refers to the total reaction time to a change in the state of the market, and can be decomposed into the time needed to acquire, process, and trade upon upon new information (see e.g. Hasbrouck and Saar, 2010).

[2]Although HFTs did not trigger the flash crash, their highly-correlated responses to an initial shock contributed considerably to the severity of the drop. Furthermore, HFTs did not lose money during this crash, but in fact seem to have made more profits than on previous days. In contrast, traditional intermediaries (i.e., market makers, pension funds and mutual funds) incurred significant losses (Kirilenko, Kyle, Samadi and Tuzun, 2011). See CFTC-SEC (2010), Menkveld and Yueshen (2011), and Easley, Lopèz de Prado and O'Hara (2012) for further in-depth analyses of the flash crash.

[3]Mini flash crashes are abrupt and severe price changes that occur in an extremely short period. Recently-reported examples include the shares of Google on 4/22/2013 (Russolillo, 2013), of Symantec on 4/30/2013 (Vlastelica, 2013) and of Anadarko on 5/17/2013 (Nanex, 2013). Another notable example is the BATS IPO on 3/23/2012 (Beucke, 2012). See Dugast and Foucault (2013), Golub, Keane and Poon (2012) and Johnson et al. (2012) for analyses on the linkage between HFT and mini flash crashes.

needed (see e.g. CFTC-SEC (2010) and Niederauer (2012)).

This paper addresses exactly this concern. As such, it analyzes whether or not HFTs (i) can destabilize financial markets, (ii) contribute to efficiently financing the economy in the long run, and (iii) should be regulated (and if so how). To do so, we construct a novel model of HFT liquidity provision in which potentially informed order flow arrives to a limit order market.[4] Initially, liquidity in this market is provided by a homogeneous set of relatively slow liquidity providers (i.e., low-frequency traders, or LFTs), such as traditional market makers or institutional investors. In line with reality, we then give traders the option to become technologically more advanced by investing upfront in speed and/or superior information processing technology. Nowadays, the simultaneous investment in both technological advances (which is the setup closest to real-life HFTs) generates synergy benefits for the HFTs. Historically, such benefits have been much smaller or even non-existent.[5]

To show the significance and the impact of these synergy benefits, we proceed along the following three steps. We first give traders the option to only invest in speed technology which allows to monitor the market at a lower cost. We show that if this technology is competitive enough (i.e., if the installation cost is low enough compared to the speed advantage it yields), the fast liquidity providers take over the whole market, while nobody adopts the new technology if it is too expensive. In a second step, we assume that instead of speed technology, only superior information processing technology is available. This technology allows its users to spot the typical indications of order flow stemming from better-informed traders (e.g. informed trade clustering as documented in Admati and Pfleiderer (1988)) better and faster. These users can use this information to avoid providing liquidity to incoming informed order flow, which will then end up with the non-users. As such, LFTs bear disproportionally large adverse selection losses when providing liquidity to "toxic order flow" (see also Biais, Foucault and Moinas (2013) and Easley, Lopèz de Prado and O'Hara (2012)). As compared to the first setting, we find that some traders will indeed invest in this technology. Interestingly though, not all LFTs will do so in equilibrium. The reason is that if too many traders adopt this technology, LFTs will leave the market. As a result, there will be no liquidity deman-

[4]Our focus on HFT liquidity provision is supported by Kirilenko, Kyle, Samadi and Tuzun (2010) who find that 78% of the HFT orders in their sample (trades in the E-mini futures S&P500) are limit orders. Jovanovic and Menkveld (2011) find that the HFT they are focusing on is on the passive side of the transaction in about 78% (respectively 74%) of the transactions on which it is involved on Chi-X (respectively Euronext).

[5]Consider the NYSE specialist from the past as an example. If anything, analyzing data from several sources would slow down rather than speed up his market making operations. For a liquidity-providing HFT, hardware upgrades offer computing power, memory and low latency that are useful for both information processing as well as fast order routing (e.g. multi-core processing). Co-location would again yield benefits for both speedy order routing as well as superior (in this case earlier) information processing. Moreover, modern day IT infrastructure allows for unprecedented communication speeds between the information processing and trading functions of the system.

ders to absorb informed order flow and costly market freezes would arise. These freezes would prevent informationally advanced traders from realizing informed trading profits. Therefore, the adoption rate of such technology is limited such that freezes do not occur in equilibrium. As a third step, we explore a setting in which traders can opt to invest in technology that combines speed and information superiority. The overall effect of this setting depends on whether speed technology is efficient or not.[6] If speed technology is inefficient, synergy benefits between speed and information technology increase the adoption likelihood as profits from informational superiority may cross-subsidize the high speed costs. In this case, market freezes do not occur for the same reason as indicated in the second step. However, if speed technology is very efficient, the gains from speed superiority may create an allowance for the costs resulting from market freezes. As a consequence, negative externalities from market freezes may prevent efficient market organization.

The resulting main insights can be summarized as follows. First, allowing LFTs to invest in speed technology only yields efficient outcomes: if the technology is too expensive, it will not be adopted and vice versa. Second, providing LFTs the option to invest in information processing technology may trigger information asymmetry problems. Yet, the severeness of these problems is limited as market freezes cannot materialize. Third, if LFTs are allowed to purchase speed and information processing technology simultaneously (i.e., become HFTs), the overall impact hinges on the efficiency (i.e. cost per unit of speed improvement) of the speed technology. If speed technology is inefficient, cross-subsidization from informational gains can nonetheless lead to its adoption. Market freezes in this case do not materialize. If, in turn, speed technology is efficient enough, adoption rates can grow so large that costly and inefficient market freezes can occur in equilibrium.

Our results indicate that in the absence or with low levels of informed trading, HFTs can improve liquidity. More and faster HFTs reduce average transaction costs, and cause quotes to converge faster to the efficient price. These findings indeed concur with the existing empirical results that the presence of HFTs improves market quality (see e.g. Brogaard, Hendershott and Riordan (2013), Hasbrouck and Saar (2012), Hendershott, Jones and Menkveld (2011), and Malinova, Park and Riordan (2013)). However, a different storyline unfolds when suspicions of informed trading are high. In such situations, HFTs will shun the market, even when these suspicions are ex-post unfounded/incorrect (e.g. if they were induced by a fat-finger error triggering a series of market orders). In those scenarios, only the LFTs can keep the market going. If, however, LFTs have been largely pushed out of the market as described above, trading will be thin, liquidity will be low, price discovery will be slow and markets can even stop functioning altogether.

[6] With efficient we mean that the ratio of speed over technology cost is more favorable for advanced traders than for LFTs.

As such, our model captures the potential systemic risk HFT activity brings to financial markets. While an increase in HFTs' market share improves liquidity and price discovery under some market conditions, it induces market freezes to arise in equilibrium with increasing frequency under other conditions.[7]

Our model provides insights on how financial markets should be optimally organized and regulated in the presence of HFTs. In particular, we provide insights on the impact of allowing market participants to adopt advanced speed and/or information processing technology. Moreover, we assess the effectiveness of several proposed (or implemented) regulatory measures to manage HFT activity. These include imposing a financial transaction tax, minimum latency requirements, the introduction of (contingent) make-take fees, and affirmative liquidity provisions. Those measures are shown to affect the equilibrium number of HFTs and LFTs (and as such, the aforementioned trade-off between high liquidity and low systematic risk) in different ways.

While the baseline version of the model provides valuable insights, it takes a very simplified approach to information production technology. In an extension of the model, we show how the results of our baseline model translate to a more realistic setting. To this end, we introduce a dynamic setting in which advanced traders can learn about informed trading in the recent past by observing the order book. If informed trading shows persistence, this information is useful in forecasting the likelihood of informed trading in the current period.

To our knowledge, no papers exist analyzing the effect of the introduction of HFTs on market stability. Taking a wider perspective, our paper is related to different sets of literature. First, our model contributes to the emerging theoretical HFT literature (e.g. Aït-Sahalia and Saglam (2013), Bernales and Daoud (2013), Biais, Foucault and Moinas (2013), Biais, Hombert and Weill (2010), Budish, Cramton and Shim (2013), Foucault, Hombert and Roşu (2013), Hoffmann (2013), Jovanovic and Menkveld (2011), Martinez and Roşu (2011), Pagnotta (2010), and Pagnotta and Philippon (2012)). In particular, our model is the first to focus on the systemic risk potentially brought to the financial market by HFT activity. That is, it allows to endogenously generate (and analyze) market freezes and relate their occurrence to the degree of speed and information-processing advantage that an investment in technology can generate. In addition, we obtain our findings in a novel framework featuring both liquidity shocks and adverse selection.

Second, our model fits into the literature modeling dynamic trading in financial markets through limit order books (e.g. Foucault, (1999), Goettler, Parlour and Rajan (2005, 2009), Foucault, Kadan and Kandel (2005), Parlour (1998) and Roşu (2009)). The

[7] This finding puts forward a new channel through which the evidence on crashes and high-frequency trading reported in Sornette and von der Becke (2011) could be understood. Moreover, it could be seen as an additional negative outcome of the HFT arms' race documented in Biais, Foucault and Moinas (2013) and Budish, Cramton and Shim (2013).

limit order book setting we construct is most closely related to Cordella and Foucault (1999) who consider two symmetric dealers competing for uninformed order-flow. We add to this paper, and to the theoretical limit order book literature, by introducing endogenous liquidity provision by multiple liquidity providers which can be either fast (AT) or slow (LFT). Moreover, we incorporate potentially informed incoming order flow. The few existing dynamic limit order book models which are solvable in closed-form (i.e., Foucault (1999), Foucault, Kadan and Kandel (2005), and Roşu (2009)) all abstract from informed trading.

The remainder of the paper is structured as follows. Section 2 introduces the setup of our model. Section 3 presents a formal definition of the market equilibrium, and Sections 4 and 5 analyze the equilibria arising under different informational settings. Section 6 provides extensions of the model, while Section 7 presents an analysis of some regulatory measures. Section 8 concludes. Proofs are relegated to an appendix. For the reader's convenience, a notational summary is included towards the end of the paper in Appendix C.

2 Setup

Consider a limit order book for a security with payoff $\widetilde{V}$. Given the available *public* information on this asset, the fundamental value of the asset equals μ. The set of possible quotes at which liquidity could be provided is discrete. The grid on which traders can post their prices is characterized by the size of the minimum price variation (or tick size), δ. Note that a larger δ implies a finer grid. On the grid, we denote by $\langle p\rangle^-$ the highest price which is strictly lower than p. In a similar way, $\langle p\rangle^+$ is the lowest price which is greater than or equal to p. The set of possible prices on the grid is $Q = \{..., p(-i), ..., p(0), ..., p(i), ...\}$, with $p(i) = \langle\mu\rangle^- + i\cdot\delta$ and $p(-i) = \langle\mu\rangle^- - i\cdot\delta$, $i \in \mathbb{N}$. We assume that $\mu - \langle\mu\rangle^- = \langle\mu\rangle^+ - \mu = \frac{\delta}{2}$; that is, the position of the expected asset value is halfway between ticks. In the remainder of the paper, we will focus on traders posting sell limit orders on the ask side.[8] We call $p(1)$ the "competitive price". This is the first price on the grid above μ. Furthermore, time and price priority hold on this market, and by assumption the sell limit orders expire upon being undercut.

Over time, which is continuous and is indexed by $t \in [0, +\infty]$, market participants arrive to the market. At a random time $\widetilde{T}$ within the trading game, a liquidity demander submits a market order which reflects her reservation price. This liquidity-demanding trader can be either trading out of liquidity needs, or because she has private information. Let us denote the type of liquidity demander that enters the market as a state of nature $\zeta \in \{liq, inf\}$, where liq and inf denote the liquidity induced and the private

[8]The analysis for the bid side is completely symmetric.

information induced type, respectively. The unconditional probabilities of ending up in states with $\zeta = inf$ and $\zeta = liq$ are given by $\bar{\pi}$ and $1 - \bar{\pi}$ respectively. If $\zeta = liq$, the liquidity-demanding trader arriving is assumed to have a rectangular demand, that is, she purchases 1 unit of the asset if the best ask price is lower than or equal to her reservation price p_{liq}. By assumption, p_{liq} is positioned on the price grid. We assume that $\widetilde{T}$ is exponentially distributed with parameter ν_{liq}. In turn, if $\zeta = inf$, with intensity ν_{inf} an informed trader arrives to the market at some point and submits a market order to buy the asset. She has accurate private fundamental information that $\widetilde{V} = \mu_{inf}$, where $\mu_{inf} > p_{liq}$. By assumption p_{liq} is also her reservation price for buying the security.[9] As such, liquidity providers in this market always run adverse selection risk, because they cannot provide liquidity at a quote at which only the traders buying for liquidity reasons are interested. If a liquidity demander ever arrives to an empty order book, the state of nature stays the same and the liquidity demander will re-visit the market at a later time again according to the same intensity. Importantly, none of the liquidity providers can observe whether a liquidity demander has already sent a market order to the order book when it was still empty. When the trade occurs, the game ends and the asset payoff $\widetilde{V}$ is realized.

There is a unit mass of risk neutral agents in the market that can choose to invest in liquidity provision technology before trading starts. These agents can choose to become either of two types of liquidity providers: (i) advanced traders (ATs) that can be fast, smart or both, and (ii) low frequency traders (i.e., LFTs). In our model the fraction of agents that becomes AT is denoted by $m \in [0, 1]$ and the fraction that becomes LFTs is denoted by $n \in [0, 1 - m]$. Before the trading game starts, ATs and LFTs need to make fixed cost investments. More specifically, the masses of ATs and LFTs need to make an investment mC_A and nC_L, respectively, which are borne equally by all constituents in each respective group. Hence, individual ATs and LFTs face cost densities of C_A and C_L respectively.[10] These costs could be seen as annualized costs of IT infrastructure, fees for keeping trading accounts or fees for co-location at the exchange and are incurred ex-ante. Once endogenously determined, m and n are assumed to remain constant over time throughout the trading game. Note that when $m + n < 1$, some traders simply choose not to participate.[11] Moreover, as will be further explained below, during the trading game the four trader types differ in two other respects: (i) the magnitude of their

[9] There can be several reasons why informed traders have a reservation price that strictly falls short of the private value. One can think about limited market capacity and staged trading with price impact as in Kyle (1985), the need to recoup information production costs, having noisy information in combination with risk aversion, etc.

[10] We consider a setting with a continuum of liquidity providers for tractability reasons. It can be derived as the limit of a discrete case where the numbers of LFTs and ATs are large.

[11] We will assume that the total mass of players eligible to be liquidity provider is so large that the upper bound of 1 never binds. This ensures that for m and n we either have a boundary solution at 0 or an interior solution.

monitoring cost (which determines the frequency at which they are able to access the market), and (ii) their processing capacity of real-time order-book information. When ATs are only fast, they have lower monitoring costs and are therefore faster, but not better informed than the LFTs. When ATs are only smart, they have superior ability to process order book information and are therefore better informed, but are not faster than LFTs. Finally, ATs that are both fast and smart are what we would classify as HFTs in today's limit order markets. Those traders are faster and better at processing information than LFTs by the virtue of their superior hardware and co-location.

Over time, liquidity providers arrive randomly and post sell limit orders. In particular, traders arrive to the market following a Poisson process. To capture the speed advantage of advanced traders relative to LFTs, we assume that ATs have technology to monitor the market γ times as often as LFTs. As a result, aggregate LFT market arrival intensity equals $n\lambda$, whereas the aggregate AT market arrival intensity is given by $m\gamma\lambda$. By assumption, $\gamma > 1$ for fast and HFT advanced trader types and $\gamma = 1$ for smart ATs. This setup reflects the higher frequency with which fast traders and HFTs monitor the market, and also captures the greater competition for exposure if γ and/or m increase. Furthermore, we assume that smart ATs and HFTs have superior abilities to process information compared to LFTs. These divergences in monitoring capacities are captured in different information sets ψ_k available to the liquidity providers of type k. In particular, for smart ATs and HFTs, ψ_{AT} contains a noisy but informative signal $s \in \{inf, liq\}$ available about the state of nature. Signals $s = liq$ and $s = inf$ are correct with probabilities $\phi_1 \in (0.5, 1]$ and $\phi_2 \in (0.5, 1]$, respectively. Let us for tractability reasons also assume that the unconditional probability of a signal $s = inf$ equals $\bar{\pi}$ such that signals are unbiased.[12]

The information asymmetry among liquidity providers may lead to a lemons problem that is so severe that markets freeze. We assume that such freezes are particularly costly for ATs.[13] In particular, every time the market freezes, the mass of advanced traders incurs a cost mC_M, to be split equally among all constituents. Hence, upon the occurrence of a freeze, ATs face an additional cost density of C_M.[14] The expected freeze costs are assumed to at least offset any information advantage an AT may have (i.e. $C_M \geq \phi_2(\mu_{inf} - p_{liq})$). In the base case, we do not make any assumptions as to how the market unfreezes again.[15]

[12]In Subsection 6.4, we extend the model to a dynamic setting where ATs learn by observing past order flow. If states are persistent, observing past order flow allows them to forecast the current state of nature in a rather accurate way. The assumption $P(s = inf) = \bar{\pi}$ is also consistent with this framework.

[13]Among others, this is motivated by the fact that advanced traders such as HFTs are very thinly capitalized and therefore very sensitive to increasing volatilities, margins and holding periods.

[14]We normalize freeze costs for LFTs to zero.

[15]In Subsection 6.2 we put forward some mechanisms for the market to unfreeze again.

3 Equilibrium

The aim of this section is to provide a formal definition of the equilibrium. First, ATs and LFT limit order placement strategies are characterized. Such a strategy is a mapping $R_k(\cdot)$, with $k \in [LFT, AT]$, from the set of possible states of the order book (i.e. standing best quote) into the set of possible offers Q. The reaction function $R_k(\cdot)$ provides the new price posted by a trader given the state of the order book upon arrival. If a trader is indifferent between two limit orders with different prices, we assume that she submits the limit order creating the larger spread. In a next step, we define an equilibrium of the trading game, which is a pair of order placement strategies (i.e., R^*_{LFT} and R^*_{AT}) such that each trader's strategy is optimal given the strategies of all other traders. Finally, [conditions for] the equilibrium number of AT and LFT traders, set in the initial participation stage, is [are] derived.

3.1 Traders' Order Placement Strategies

We analyze trader k's order placement strategy given a standing best ask quote $\widehat{a}$ positioned on the price grid upon arrival at time τ.[16] Assuming the time of arrival τ is earlier than the time of arrival of the market order and given the information set ψ_k, trader k's expected profit of posting a limit order at quote a could be depicted as follows:

$$\Pi_k(a, \widehat{a}) = \begin{cases} 0 & \text{if } a \geq \widehat{a} \\ E\left(\Phi\left(a, \psi_k\right) \cdot (a - \widetilde{V}) | \psi_k\right) & \text{if } a = \widehat{a} - i \cdot \delta \end{cases} \tag{1}$$

where $i \in \mathbb{N}^+$, $\Phi\left(a, \psi_k\right)$ is the trader's expected execution probability corresponding to quote a, and $E(\cdot|\psi_k)$ is the trader's expectation over states of nature conditional on her information set. In particular, the asset value may equal μ or μ_{inf}, and traders make assessments of this value and execution probabilities based upon the information set they have upon their arrival. For both trader types, submitting an ask quote a which is less or equally aggressive than the best quote upon arrival yields a zero expected execution probability and therefore a zero expected profit. In turn, submitting a quote which improves the best quote upon arrival by i ticks features a positive expected execution probability hinging on future arriving traders' strategies. Noteworthy, when $\zeta = liq$, undercutting to the competitive quote $p(1)$ yields $p(1) - \mu$ with certainty (i.e., $\Phi(p(1), \psi_k) = 1$), as this quote can never be profitably undercut by any liquidity provider. As such, upon arrival, the traders commonly face a trade-off between a higher execution price and a higher expected execution probability.

[16] As by assumption all backlying sell limit orders expire upon being undercut by an order at $\widehat{a}$, the order placement strategies depend only on this quote (and not on all the orders submitted at less aggressive quotes). In section XX we sketch a repeated version of the model in which liquidity providers can actively choose to cancel quotes or not when a new iteration starts.

3.2 Equilibrium Definition

Let $V_k(\widehat{a})$, with $k \in \{AT, LFT\}$, be trader k's expected profit given that the current best quote is $\widehat{a}$ and the trader is about to react. $V_k(\widehat{a})$ can be expressed as:

$$V_k(\widehat{a}) = \max_{R_k \in Q} \Pi_k(R_k, \widehat{a}) \tag{2}$$

where all traders behave according to R^*_{LFT} and R^*_{AT}. Thus, both trader types account for the expected profit of their current action only (i.e., $\Pi_k(R_k, \widehat{a})$). As players are atomistic, the probability of arriving to the market again, given arrival now is zero.[17]

The solutions of these dynamic programming relationships yield the optimal placement strategies, R^*_{AT} and R^*_{LFT}. The expected execution probabilities of both trader types are computed assuming that traders follow these strategies. Traders' optimal order placement strategies hinge on the expected execution probabilities. The expected execution probabilities are in turn determined by traders' order placement strategies. The type of equilibrium we are looking for is a Nash equilibrium.

3.3 Initial Participation Stage

The equilibrium definition of the trading stage in subsection 3.2 starts from given masses of ATs and LFTs, m and n, respectively. However, with fixed participation cost parameters C_A and C_L, participation may not be optimal for any masses of ATs and LFTs. Therefore, as highlighted in the setup, the model starts off with a pre-trade participation stage which allows to solve for the equilibrium participation masses, m^* and n^*. As agents are rational and we consider a market with perfectly competitive entry, ex-ante expected equilibrium profits must be positive and will mostly equal zero. Hence, we need to find a pair $\{m^*, n^*\}$ with $m^*, n^* \geq 0$ such that for both player types marginal utility of participation is positive but as close to zero as possible:

$$n^* = \begin{cases} 0 \;\text{ if } E_{\widehat{a}} \left(\Pi_{LFT}(R^*_{LFT}(\widehat{a}), \widehat{a}) | m^*, n\right) < C_L \;\; \forall n, \\ \arg\min_n E_{\widehat{a}} \left(\Pi_{LFT}(R^*_{LFT}(\widehat{a}), \widehat{a}) | m^*, n\right) - C_L \geq 0 \text{ otherwise.} \end{cases} \tag{3}$$

and similarly

$$m^* = \begin{cases} 0 \;\text{ if } E_{\widehat{a}} \left(\Pi_{AT}(R^*_{AT}(\widehat{a}), \widehat{a}) | m, n^*\right) < C_A \;\; \forall m, \\ \arg\min_m E_{\widehat{a}} \left(\Pi_{AT}(R^*_{AT}(\widehat{a}), \widehat{a}) | m, n^*\right) - C_A - I^F \bar{\pi} c_M \geq 0 \text{ otherwise,} \end{cases} \tag{4}$$

[17]It is possible to set up the model with a discrete number of LFTs and HFTs and allow for re-entering the market. This hardly affects the results and comes with a substantial loss of tractability.

where I^F is an indicator function that equals one in case of a market freeze and 0 otherwise.

3.4 Market liquidity

The derivation of the equilibrium number of ATs and LFTs in the previous subsection is closely related to the average liquidity level z in the book (that is, the average effective spread). If all expected revenues are exactly offset by investments in the most liquidity enhancing technology, we would obtain a 'first best' spread level z^*. However, we may have that endogenous barriers to entry allow for rents. These are revenues not spent on technology (in expectation). Hence, these rents increase spreads. Moreover, there may be allocative inefficiency in equilibrium, leading to investments in inefficient technology and therefore, lower undercutting speed and higher spreads. Finally, information technology may become so widely adopted that a substantial fraction of all informed trades can be avoided altogether. However, this would mean that markets freeze every now and then, leading to revenue losses on false positives and freeze costs. The net of those would be deadweight loss and hence lead to an underinvestment in technology and thereby to increased spreads. Let us call the expected spread markups due to rents, allocative inefficiency and net freeze costs z^r, z^{ineff} and z^{freeze} Hence, we have:

$$E(z) = z^* + z^r + z^{ineff} + z^{freeze}. \tag{5}$$

We will explicitly refer to these components in the different equilibria we analyze.

4 Quote Dynamics and Trading Costs

In this section, we characterize the equilibrium order placement strategies for cases with (i) LFTs and fast, but equally uninformed ATs, (ii) LFTs and smart, but slow ATs, and (iii) LFTs and smart and fast ATs (i.e HFTs). However, we first derive equilibrium strategies for what we call the uninformed trading case where the informed state of nature never materializes. The uninformed case is illustrative for our model setup and an important building block for our more general case with informed trading. Moreover, one can show that the equilibrium with fast ATs in the presence of informed liquidity demanders can be derived from a simple transformation of the uninformed case. Next, we develop the informed trading case. To maintain tractability, we look at an informed case with certain parameter restrictions.[18] The main features and trade-offs put forward in this paper will largely extend to the unrestricted version of the informed case.

[18] A general informed case can be derived but has very low tractability.

4.1 Uninformed Trading Case

The uninformed case is characterized in the model by setting $\bar{\pi} = 0$. This parameter restriction is maintained throughout section 4.1. As divergences in information processing capacities do not matter in this uninformed case, we can abstract from the information sets ψ_k. As a result, each AT is in the trading stage of the game equivalent to γ LFTs.

As we will see later, if m and n are endogenous, the most cost efficient type of liquidity provider will dominate the whole market. As the uninformed case is a building block for the restricted informed case where LFTs and ATs can co-exist, we derive optimal strategies for LFTs and ATs when they compete with one another.

4.1.1 Equilibrium Strategies

Consider a time τ (assumed earlier than the time of arrival of the uninformed market order) at which a trader k arrives to the market. Let us assume that the standing best price in the market upon arrival $\widehat{a}$ is strictly above $p(1)$. Joining the queue at the standing best quote or reverting to a backlying quote upon arrival yields this trader a zero execution probability, and thus zero profit. In contrast, undercutting to the competitive quote $p(1)$ yields a positive expected profit of $(p(1) - \mu)$ with certainty. As such, queue-joining or reverting strategies are always strictly dominated by an undercutting strategy in terms of expected payoffs, and hence will never be played (see also Subsection 3.1). Furthermore, as traders are atomistic, there is a zero probability of arriving in the market again and observing a self submitted standing best quote.

In case the standing best price in the market upon arrival $\widehat{a}$ equals $p(1)$, the competitive price is reached. This implies that it is no longer possible to play a profitable undercutting strategy. We assume arriving traders observing this quote upon arrival choose to join this best queue. This allows us to establish the following properties of the equilibrium order placement strategies and consequently of the expected equilibrium execution probabilities.

Lemma 1 *(Monotonicity). Consider equilibrium order placement strategies $R^*_{LFT}(\cdot)$ and $R^*_{AT}(\cdot)$ with $\bar{\pi} = 0$. For all parameter values, these functions have the following properties:*

- ***(P1)*** *$R^*_k(\widehat{a}) < \widehat{a}$ if $\widehat{a} \geq p(2)$; and*
- ***(P2)*** *$R^*_k(p(1)) = p(1)$.*

As a result, the expected execution probability of a limit order undercutting the standing best quote $\widehat{a}$ is derived as follows:

- *For limit orders undercutting to a quote which is strictly larger than $p(1)$, submitted by an AT and LFT, respectively, we have*

$$\Phi(R^*_{AT}(\widehat{a})) = \Phi(R^*_{LFT}(\widehat{a})) = \frac{\nu_{liq}}{\nu_{liq} + \lambda(\gamma m + n)} \equiv \Phi. \tag{6}$$

- *For a limit order undercutting to $p\,(1)$, we have*

$$\Phi(R^*_{AT}(\widehat{a})) = \Phi(R^*_{LFT}(\widehat{a})) = 1. \tag{7}$$

Proof. See appendix. ■

Summarizing, Lemma 1 is important for two reasons. First, ***(P1)*** states that in equilibrium, the best ask quote must decrease as long as it is strictly greater than the competitive price $p(1)$. Undercutting is thus the unique possible evolution for the best ask quote. Second, ***(P2)*** claims that, with time priority, the unique focal price is the competitive price.[19] These results imply that there necessarily exists a price $\tilde{p}^* \in (p(1), p_{liq}]$, such that when the best quote reaches $\tilde{p}^*$, the arriving trader without execution priority finds it optimal to post $p(1)$ and thus secure execution. The next proposition characterizes the unique price at which the "jump" to the competitive price occurs. It also provides traders' order placement strategies in equilibrium.

Proposition 1 *(Equilibrium Order Placement Strategies). With time and price priority enforced, any market participant $k \in \{LFT, AT\}$ follows the following strategy when observing quote $\widehat{a}$ upon arrival:*

$$R_k = \begin{cases} p_{liq} & \text{if } \widehat{a} - \delta \geq p_{liq} \\ \widehat{a} - \delta & \text{if } p_{liq} > \widehat{a} - \delta \geq \tilde{p}^* \\ p(1) & \text{if } \widehat{a} - \delta < \tilde{p}^* \end{cases}, \tag{8}$$

where

$$\tilde{p}^* = \left\langle \mu + \frac{\delta}{2\Phi} \right\rangle^+ = p(1) + \left\lfloor \left\lfloor \frac{1-\Phi}{2\Phi} \right\rfloor \right\rfloor \cdot \delta \tag{9}$$

with $\lfloor\lfloor x \rfloor\rfloor$ denoting the greatest integer strictly lower than x.

Proof. See Appendix. ■

The intuition for Proposition 1 is as follows. Consider a trader k arriving in the market at time τ, observing a standing limit order at quote $\widehat{a}$ which is smaller or equal to

[19] Following Maskin and Tirole (1988), we call a focal price a price p on the equilibrium path such that $R_k(p) = p$. If there exists a focal price, once it is reached, the traders keep posting this price until the arrival of the market order.

the incoming market order trader's reservation price p_{liq}. This trader faces the following trade-off. If she quotes the competitive price, she secures execution and obtains with certainty a profit equal to $p(1) - \mu = \frac{\delta}{2}$. If instead she undercuts $\hat{a}$ by only one tick, she obtains a larger profit (i.e., $\widehat{a} - \delta - \mu$) in case of execution. Yet, she then runs the risk of being undercut by a subsequently arriving trader before the market order has arrived. Hence, the payoff of this limit order accounts for the corresponding execution probability (see Lemma 1). When $\tilde{p}^*$ is reached in the sequential undercutting process, traders switch strategies from tick-by-tick undercutting to quoting $p(1)$ immediately. To get an idea of how the undercutting patterns look like, one could have a look at Figure 1. The undercutting starts at p_{liq} and continues with all players undercutting each other. When $\tilde{p}^*$ is reached, all traders jump to $p(1)$, which is the quote at which execution will later take place when the liquidity demander arrives (here at time 190).

Previous empirical literature has found that ATs in general improve market liquidity. Lemma 1 and Proposition 1 provide insight into how ATs improve market liquidity absent information asymmetry. In this setting, more liquidity providers are beneficial for market liquidity for two reasons. First, with more liquidity providers, the arrival frequency of liquidity providers to the market is higher, leading to faster undercutting and therefore lower effective spreads. Second, the increased competition for order flow will also induce more aggressive strategies from liquidity providers, inducing them to jump to $p(1)$ earlier (i.e. higher $\tilde{p}^*$). Holding constant the total mass of liquidity providers, both effects are stronger with ATs, because those have $\gamma \geq 1$.

4.1.2 Expected Trading Profits

In order to calculate the equilibrium masses of ATs and LFTs, m^* and n^*, respectively, we need to calculate the expected profit densities $E(\sum_{\widehat{a}} \Pi_{AT}(R^*_{LFT}(\widehat{a})))$ and $E(\sum_{\widehat{a}} \Pi_{LFT}(R^*_{LFT}(\widehat{a})))$. If, conditional on m and n, the strategies R^*_{AT} and R^*_{LFT} are played, we can distinguish two regions along the equilibrium path. In the first region from p_{liq} down to $\tilde{p}^*$ inclusive, denoted "*UC*", both ATs and LFTs undercut the standing best quote tick-by-tick when upon arrival to the market. In the second region, denoted "*comp*", each liquidity provider that accesses the market will post a quote at $p(1)$. Figure 1 depicts these two regions graphically.

Next, let us first define $\bar{\lambda} = (n + \gamma m)\lambda$, the overall arrival intensity of liquidity providers. Moreover, let us define Z as the number of ticks from p_{liq} up to $\tilde{p}^*$ inclusive. Proposition 2 then presents the unconditional expected profits for both trader types.

Proposition 2 *For an LFT and an AT, the unconditional expected profit densities are* *given by, respectively*

$$E\left(\sum_{\widehat{a}} \Pi_{AT}(R^*_{LFT}(\widehat{a}))\right) = (1 - f_{LFT})m^{-1}(E(\Pi^{UC} + \Pi^{comp})), \tag{10}$$

$$E\left(\sum_{\widehat{a}} \Pi_{LFT}(R^*_{LFT}(\widehat{a}))\right) = f_{LFT}n^{-1}(E(\Pi^{UC} + \Pi^{comp})), \tag{11}$$

where

$$E(\Pi^{UC}) = \sum_{i=0}^{Z} \frac{\nu_{liq}\bar{\lambda}^i}{(\nu_{liq} + \bar{\lambda})^{i+1}}(p_{liq} - i \cdot \delta - \mu), \tag{12}$$

$$E(\Pi^{comp}) = (1 - P_{UC})(p(1) - \mu). \tag{13}$$

$$P_{UC} = \sum_{i=0}^{Z} \frac{\nu_{liq}\bar{\lambda}^i}{(\nu_{liq} + \bar{\lambda})^{i+1}}, \tag{14}$$

$$f_{LFT} = \frac{n}{n + \gamma m}. \tag{15}$$

Proof. See appendix. ■

The interpretation of the expressions in Proposition 2 is as follows. ATs and LFTs share in the aggregate expected surplus according to their relative presence in the market given by f_{LFT}. The aggregate expected profits in the UC region is given by the probability weighted average trading profit at each tick in this range (where weights can sum to less than one). The aggregate expected profit in the *comp* region is given by the probability of reaching it times the guaranteed profit of half a tick.

With the expressions in Proposition 2, we can look for the equilibrium number of ATs and LFTs. As expected profits for both LFTs and ATs are monotonically decreasing in m and n and cost densities are constant, it is always possible to find an equilibrium with a strictly positive mass of at least one type of liquidity providers.

At this point, we can apply a trick to facilitate our analysis. Due to the assumption of exponentially distributed arrival times, aggregate liquidity provider arrival intensities are linear in m and n with coefficients γ and 1 respectively. Total costs for liquidity provision are also linear in m and n with the same coefficients. Therefore, one AT with speed γ and cost C_A is equivalent to γ ATs with speed 1 and cost $\frac{C_A}{\gamma}$. We state the following lemma without proof:

Lemma 2 *The original problem is equivalent to a modified problem in which each AT has speed 1, cost density* $\frac{C_A}{\gamma}$ *and where the mass of ATs is* γ *times as large. This result holds in the uninformed and informed setting.*

Lemma 2 simplifies our analyses considerably. The equilibrium masses of ATs and LFTs can now be derived in a straightforward way. We have a competitive market with free entry for a homogeneous product. Therefore, prices in equilibrium must equal production costs *of the most efficient producer of liquidity provision services.* As liquidity provision at those expected revenues is not profitable for the least efficient liquidity provider, the most efficient liquidity providers must dominate the market. Hence, if $\frac{C_A}{\gamma} < C_L$ we will only have ATs in equilibrium and if $\frac{C_A}{\gamma} > C_L$, we only have LFTs.

Proposition 3 *In the uninformed case, liquidity provision is conducted in equilibrium by ATs when $\frac{C_A}{\gamma} \leq C_L$, and by LFTs otherwise.*

Proof. See appendix. ■

Due to proposition 3 allocation is always efficient. Moreover, as entry into the market is free, liquidity providers cannot make positive profits in expectation and hence, expected spreads must be at their first best level z^*.

4.2 Informed Trading Case

In this subsection, we work out the model with information asymmetry. Within the uninformed trading case, the market would be dominated by either ATs or LFTs, depending on the cost of speed. In the setting with information asymmetry, we can have that LFTs and ATs both participate in equilibrium. Smart ATs and HFTs have the benefit that they can process information better than LFTs. This allows them to forward toxic order flow to LFTs, hence draining LFT profits and increasing their own. However, this information superiority can lead to a lemons problem that results in costly market freezes which will be further analyzed in Section 5. The possibility of such market freezes can form entry barriers for ATs. As a result, equilibria may be possible with both LFTs and ATs.

To facilitate exposition and tractability, we assume infinitely impatient informed liquidity demanders, that is $\nu_{inf} = \infty$.[20] One could think about this assumption as having a large informed trader that has a substantial volume to trade and sequentially splits this in smaller blocks (as for instance documented in Admati and Pfleiderer (1988)). The informed trader will monitor the market constantly in order to push through the volume as quickly as possible (for example because information may be perishable). The main advantage to this way of modeling is that informed trading is immediately disclosed as soon as a limit order is put into the book. This makes the inference for LFTs that arrive to a non-empty order book trivial: there is no informed trading. Therefore, if a quote survives, the trading game reduces immediately to the uninformed case. Hence,

[20]The model can be extended to allow for more patient informed liquidity demanders, at the expense of reduced tractability and increased notational complexity. The main results will be largely unaffected.

it is sufficient to solve for the opening bid of the trading game only and all uncertainty is resolved right at the beginning of the stage game.

Below, we first show how under this impatience assumption, the equilibrium with fast ATs is equivalent to the uninformed case with a parameter transformation. Next, we develop trading equilibria in the presence of smart ATs and HFTs.

4.2.1 Only speed matters: equilibria with fast ATs

The uninformed case is easy to derive and offers high tractability. However, to do a full comparison among the different settings with the different types of ATs, we need to have a setting with fast ATs and informed trading. In this subsection, we show that under mild conditions the equilibrium with fast ATs can easily be obtained from the uninformed case. To see this, one should realize that informed trading generates unavoidable losses for ATs and LFTs alike, since none of them can use any conditioning information. Therefore, these expected losses when entering an opening quote in the book can be considered as exogenous as long as they do not exceed the expected profits from providing liquidity to uninformed liquidity demanders. Therefore, the expected losses (and somewhat lower expected income) can be seen as an additional fixed cost. Hence, quote posting strategies are identical to those in the uninformed case. The only difference is in the participation stage, where participation is more costly. Therefore, the equilibrium strategies must be the same as the equilibrium strategies arising from the uninformed case with the following modifications to participation cost densities:

$$\tilde{C}_L = \frac{C_L + \bar{\pi}\frac{1}{n+\gamma m}(\mu_{inf} - p_{liq})}{1-\bar{\pi}}, \qquad \tilde{C}_A = \frac{C_A + \bar{\pi}\frac{\gamma}{n+\gamma m}(\mu_{inf} - p_{liq})}{1-\bar{\pi}}. \tag{16}$$

4.2.2 Information processing matters: equilibria with smart ATs and HFTs

To derive the optimal quote posting strategies for ATs and LFTs, with $\nu_{inf} = \infty$ it suffices to analyze their respective strategies upon arrival to an empty book. When an AT arrives to an empty book, it will only add a quote p_{liq} when the expected profits from posting an initial quote outweigh the expected losses from doing so. Expected freeze losses do not contribute to this decision, as those are infinitely small for an individual AT. In contrast, adverse selection losses can be substantial on an individual basis. Intuitively, this could be seen as a traditional commons problem in which no AT individually internalizes the general freeze cost. Therefore, it is optimal to post an initial quote when the expected gain of providing liquidity to uninformed order flow exceeds the expected loss due to liquidity provision to informed order flow:

$$(p_{liq} - \mu)\hat{P}(\zeta = liq|\psi_{AT})\Phi(\zeta = liq) \geq (\mu_{inf} - p_{liq})\hat{P}(\zeta = inf|\psi_{AT})\Phi(\zeta_l = inf) \tag{17}$$

where $\hat{P}(\zeta = inf|\psi_{AT})$ and $\hat{P}(\zeta = liq|\psi_{AT})$ are the posterior probabilities for the AT of having an informed or uninformed trader as the first liquidity demander to come to the market, respectively. We have that

$$\hat{P}(\zeta = inf|\psi_{AT}) = \begin{cases} \phi_2 \text{ if } s = inf, \\ 1 - \phi_1 \text{ if } s = liq. \end{cases} \tag{18}$$

The execution probabilities are also completely defined, because in the case of informed trading execution is guaranteed and immediate, while in the case of uninformed trading, the game reduces after the first stage to the uninformed trading game. Hence, we have

$$\Phi(\zeta = inf) = 1, \qquad \Phi(\zeta = liq) = \Phi. \tag{19}$$

Substituting these expressions and (6) into (17) and rewriting indicates that an AT will never post a quote to an empty book at all if

$$(\gamma m + n) > \frac{\nu_{liq}(p_{liq} - \mu)\phi_1}{\lambda(\mu_{inf} - p_{liq})(1 - \phi_1)}. \tag{20}$$

Note that if it is not profitable for ATs to post in an empty book, the same must be true for LFTs, as ATs have superior information over LFTs.

On the other hand, an AT will always post a quote in an empty book if

$$(\gamma m + n) \leq \frac{\nu_{liq}(p_{liq} - \mu)(1 - \phi_2)}{\lambda(\mu_{inf} - p_{liq})\phi_2}. \tag{21}$$

In all other cases ATs will post upon a signal $s = liq$ and will not post upon signal $s = inf$.

In turn, for the LFT, there is a similar profitability condition to be met. In order to post a quote to an empty book the expected gains from liquidity provision to uninformed order flow must exceed the expected loss from providing liquidity to informed order flow:

$$(p_{liq} - \mu)\hat{P}(\zeta = liq|\psi_{LFT})\Phi(\zeta = liq) \geq (\mu_{inf} - p_{liq})\hat{P}(\zeta = inf|\psi_{LFT})\Phi(\zeta = inf), \tag{22}$$

Naturally, this inequality is more likely to be violated when the posterior probability of informed trading is larger, informed trading losses are larger, uninformed trading gains are lower and uninformed trading execution probabilities are lower.

Proposition 4 *LFTs leave the market when informed trading losses are large, uninformed trading gains are low, uninformed trading execution probabilities are low and the* *posterior probability of informed trading conditional on arrival to an empty order book is high. This posterior probability is increasing in: (i) the fraction of ATs (m), (ii) the ATs' speed advantage (γ), (iii) the probability that the informed trading signal is correct (ϕ_2), and (iv) ATs conditioning on information. It is decreasing in: (i) the fraction of LFTs (n), and (ii) the unconditional probability of ending up in an informed state ($\overline{\pi}$).*

5 Profitability, Participation and Market Failure

Having established optimal strategies of the different players in this economy, we can analyze the costs and benefits of having market participants with advanced technology available. In line with previous literature, we find that the availability of speed technology in itself is good. If it is inefficient (i.e. too expensive), it will not be adopted and vice versa. Competition among liquidity providers assures that the lower costs of providing liquidity benefits society as a whole in the form of more liquid markets.

Proposition 5 *If LFTs can only choose to adopt speed technology, the availability of this technology never reduces liquidity. If it is efficient enough, it takes over the whole market and market liquidity improves.*

Proof. See Appendix. ■

The availability of information processing technology on the other hand may trigger information asymmetry problems. To analyze those, let us first define expected profit functions for LFTs and ATs with information technology conditional on their optimal quote posting strategies. First let $b = \gamma m$. Due to Lemma 2 this transformation is without loss of generality. Next, let us define

$$g(b+n|\tilde{p}_{liq}=q) = \frac{E(\Pi^{UC}+\Pi^{comp})}{n+b}. \tag{23}$$

One can easily verify that $g'(\cdot) < 0$.

If in equilibrium (21) is satisfied, ATs always quote at an empty book as in the speed-only case. In this case, information technology is irrelevant and the most efficient liquidity provider dominates the market.

Proposition 6 *Information technology is irrelevant if participation costs are high, information technology is inaccurate, uninformed trading is intense and informed trading losses are small. This is the case when*

$$g^{-1}\left(\min\left(\frac{\tilde{C}_A}{\gamma},\tilde{C}_L\right)\middle| p_{liq}\right) \le \frac{\nu_{liq}(p_{liq}-\mu)(1-\phi_2)}{\lambda(\mu_{inf}-p_{liq})\phi_2}. \tag{24}$$

Under this condition, liquidity provision is efficient.

Proof. See Appendix. ■

Let us now consider what happens if (24) is violated. Define

$$g_A(b+n) = \\ (1-\bar{\pi})\left((1-\Phi)g(b+n|\tilde{p}_{liq}=p_{liq}-\delta)+\frac{\phi_1(p_{liq}-\mu)\Phi-(1-\phi_1)(\mu_{inf}-p_{liq})}{n+b}\right). \quad (25)$$

Whenever (24) is violated and (22) is satisfied, $g_A(b+n)$ is the marginal profit from trading for ATs. Whenever (24) and (22) are violated, the marginal profit from trading for ATs is given by

$$h_A(b,n) = \\ (1-\bar{\pi})\left((1-\Phi)g(b+n|\tilde{p}_{liq}=p_{liq}-\delta)+\frac{\phi_1(p_{liq}-\mu)\Phi-(1-\phi_1)(\mu_{inf}-p_{liq})}{b}\right). \quad (26)$$

One can verify that $g'_A(\cdot)<0$ and $h'_A(\cdot)<0$ are negative on their domains.

Let us also define

$$g_L(b,n) = g_A(n+b)+\bar{\pi}\frac{(1-\phi_2)(p_{liq-\mu})\Phi-\phi_2(\mu_{inf}-p_{liq})}{n}. \quad (27)$$

Whenever (22) is satisfied and (24) is violated, $g_L(b,n)$ is the marginal profit from trading for LFTs. One can verify that $g_L(0,n)=C_L$ iff $g(n)=\tilde{C}_L$. Whenever (22) and (24) are violated, marginal profit from trading for LFTs is given by $(1-\bar{\pi})(1-\Phi)g(b+n|p_{liq}-\delta)$.

The marginal cost density for LFTs is given by C_L. For ATs, the marginal cost density equals $\frac{C_A}{\gamma}$ if (22) is satisfied and equals $\frac{C_A}{\gamma}+\bar{\pi}c_M$ if (22) is violated.

Finally, let us define the pair $(\bar{b},\bar{n})$ as the point at which $(1-\bar{\pi})(1-\Phi)g(b+n|p_{liq}-\delta)=C_L$ and (22) binds exactly. One should note that participation in an empty book yields zero expected profit for LFTs in $(\bar{b},\bar{n})$. Therefore, marginal expected profits from trading for LFTs are continuous in this point.

We can now formulate the main proposition of our paper:

Proposition 7 *Assume (24) is violated. We have that*

1. *Liquidity is exclusively provided by LFTs if $g_L(g_A^{-1}(\frac{C_A}{\gamma}),0)>C_L$. This is efficient.*

2. *ATs and LFTs co-exist and jointly provide liquidity with masses $\bar{b}$ and $\bar{n}$ respectively if $g_A(g^{-1}(\tilde{C}_L))>\frac{C_A}{\gamma}$, $g_A(\bar{b}+\bar{n})>0$ and $h_A(\bar{b},\bar{n})<0$. AT receive rents in this case, but freezes do not occur.*

3. *Liquidity is exclusively provided by ATs if $g_A(\bar{b}+\bar{n})>0$ and $h_A(\bar{b},\bar{n})>0$. In this case, a market freeze takes place with probability $\bar{\pi}$. ATs break even, but this outcome is inefficient.*

Proof. See Appendix. ■

Proposition 7 helps us in assessing the social value of speed and information technology and in particular the bundled package.

If ATs only have access to information technology (i.e. $\gamma = 1$), the outcome depends on parameters. If $C_A >> C_L$, information technology is likely to be too expensive and scenario 1 materializes. If $C_A = C_L + \epsilon$, where ϵ is small and c_M is large, ATs will need to prevent freezes as those are excessively costly. However, due to their informational advantage, it is profitable for them to enter when it is not for LFTs. As marginal profits for both LFTs and ATs are strictly declining in m and n, entry by ATs must reduce the number of LFTs, which in turn allows for more AT entry. At some point however, (22) binds, ATs still have strictly positive average and marginal profit, but beyond this point marginal profit for ATs drops below 0. Hence, in this equilibrium (22) binds, ATs have strictly positive profits and no freezes occur. As a result, we will have that $z^r > 0$ and $z^{ineff} > 0$.

If ATs possess both speed and information technology (i.e., are HFTs), all three scenarios are possible. If HFT technology is excessively expensive, (i.e. $\frac{C_A}{\gamma} >> C_L$), only LFTs provide liquidity. If $\frac{C_A}{\gamma}$ is close to C_L, ATs co-exist with LFTs. One should note that this could lead to the implementation of inefficient speed technology (as with smart but slow ATs), or to under-investment in efficient speed technology (when $\frac{C_A}{\gamma}$ is slightly lower than C_L and c_M is large). Finally, the last option could materialize if $C_L >> \frac{C_A}{\gamma}$. In this case, cost savings due to a much more efficient technology create an allowance for freezes. Allocation in this case is efficient, as all liquidity is provided by the most efficient providers, but still falls short of first best. The reason is that the freeze losses create welfare costs and hence lead to an under-investment in this technology. As a result $z^{freeze} > 0$.

In itself, the restricted version of the model featuring $\nu_{inf} = \infty$ is sufficient to illustrate the main insight of the paper, namely the emergence of market freezes accompanying increases in activity of HFTs. Starting from a more general version of the model in which quote cancelations are impossible, complete market freezes are found only to arise at the beginning of the undercutting sequence, as is the case in the stylized version of the model.[21] In that case, the main difference between the stylized and the general model would be that the undercutting speed in the general model would be lower, but that posting the first quote of the sequence would be less risky.

[21] If quotes are not cancelable, a standing best quote can survive in the book for very long when informed trading suspicions are high, but it cannot disappear. Hence, the only way to have a freeze is to not have a quote posted in the first place.

6 Extensions and Practical Considerations

In this section, we address several practicalities and possible extensions to our model. Most of these extensions, except for the last, we address informally in the interest of tractability and complexity. The first extension relates to the fact that market participants in practice can have dual roles. We argue that an endogenous choice for using market or limit orders could incentivize LFTs to leave the market even more easily. The other three extensions address issues that play up in repeated versions of our model. In a repeated version of the model, frozen markets need to be defrosted, preferably in an endogenous way. Section 6.2 outlines the mechanics behind one possible unfreezing mechanism. In a repeated version of the model, assumptions on the expiry or survival and the cancelability of quotes play an important role. In section 6.3, we sketch how such assumptions in a repeated version of the model would affect our results. Finally, in section 6.4, we develop a dynamic model in which the generation of signals s is endogenously derived rather than exogenously assumed.

6.1 Dual Roles in Limit Order Markets

One of the features that crucially characterize a limit order market is that participants can trade either using limit orders or using market orders. In our setting, the freezes can arise because LFTs start to exit the market. In reality, LFTs with a trading need and a moderate tolerance for execution uncertainty (i.e. reasonable degree of patience) may provide liquidity in order to generate extra revenues and save transaction costs (note that in our setup, the degree of impatience for a certain trade is captured by C_L). Hence they trade off execution uncertainty with transaction costs. We argue that the dual role makes freezes only more likely in the presence of HFTs. After all, while revenues from liquidity provision deteriorate, the alternative of using market orders to conduct their planned trades becomes cheaper due to more intensive competition from HFTs.

6.2 Unfreezing Markets

One of the main causes of the market freezes in the presence of HFTs is that in the model uninformed liquidity demanders do not update their reservation values during market failures. After all, if p_{liq} were to adjusts upwards while μ_{inf} stays constant, expected losses due to providing liquidity to informed traders go down, while expected gains from providing liquidity to uninformed order flow go up. Hence, we can let markets unfreeze by letting p_{liq} increase after a while.

Would it be reasonable for this to happen in practice? We argue that it is. After all, the uninformed liquidity demanders depend on transaction prices for their information to base their reservation prices on. When markets freeze, the last information available to them would imply a value of μ. Only after a while, they might realize that the market has not moved for a long time and rationally increase their reservation value. After all, there is a relatively larger (posterior) probability of a higher valuation then.

For informed liquidity demanders it would also be optimal to increase reservation values in similar fashion to those of the uninformed liquidity demanders. First, this way they keep mimicking the uninformed liquidity demanders and make inference by liquidity providers harder. Second, after a while, it is likely that information would start to perish. In a frozen market, the probability of capitalizing on information is very small. Therefore, informed liquidity demanders would after a while be willing to settle for lower informed trading losses.

6.3 Quote expiry and cancelation

For the sake of simplicity and tractability, we present in our paper a one-period model. If this model were extended to a repeated version covering subsequent trades, assumptions would have to be made on quote expiry and the possibility to cancel quotes. If all quotes in the book are to expire upon the transaction taking place (and hence the end of a stage game), the one-period results we derive applies to each stage game. However, if quotes can survive to the next stage game, results may change somewhat. To get an idea about effects playing up, we sketch a two-period version of the model with non-expiring quotes below.[22]

To analyze the effect of non-expiring quotes, consider a two period version of the game where quotes carry over to the next stage game. Moreover, let us assume that quotes can be canceled upon the start of the second stage game. Finally, assume that ATs with information technology get a noisy but informative signal s upon the start of the stage game *that they can use to base their cancelation decision on.* The cancelation decision in this setting is very similar to the book opening of the one-period model. The only difference is that given n and m, the expected profitability of the two periods can be different. In particular, the opening quote in the second period is never higher and often lower than in the first. Therefore, the risk of informed trading is higher in the second period than in the first as expected gains from uninformed trading are lower and expected losses from informed trading are higher. As a consequence, LFTs will be more reluctant to provide liquidity in the second period, making freezes more likely. If the difference in profitability between the first and second opening is large, positive informed trading profits for ATs in the first period opening can be so large that an allowance for freeze losses in the second period opening is created. Hence, for limited parameter ranges and in the presence of information technology, we may have freezes in the second period that were not there in the one-period model.

[22] For tractability reasons, we limit ourselves here to a sketch of the model.

6.4 A dynamic setting

So far, the information production technology in our model has been exogenously given. If one extends the model to a fully dynamic model, then information production can be made explicit and endogenized in the model. To this end, let us consider an infinitely repeated version of our trading game. In every stage game l, a state of nature ζ_l is drawn according to a Markov Switching process with transition matrix

$$\begin{bmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{bmatrix},$$

where α and β denote the probabilities of continued liquidity trading and continued informed trading, respectively. In turn, $1-\alpha$ and $1-\beta$ denote the switching probabilities from liquidity to informed and from informed to liquidity trading, respectively. Unconditional steady state probabilities are then given by $\bar{\pi} = \frac{1-\alpha}{2-\beta-\alpha}$ and $1-\bar{\pi} = \frac{1-\beta}{2-\beta-\alpha}$.

This setup allows to capture the clustering of informed trades as further documented below. [23]

We now assume that informed order flow is less patient than uninformed order flow (i.e. $\nu_{inf} > \nu_{liq}$) and that smart ATs and HFTs can perfectly observe the historical evolution of the order book.

The difference in patience between informed and uninformed liquidity demanders allows for inference about trading types in previous periods by ATs. This information is particularly useful when $\beta \neq 1-\alpha$, because information about the previous liquidity-demanding trader type will then help to better forecast the current trader type. In particular, when $\nu_{inf} = \infty$, ATs can perfectly infer the state of nature of the previous stage game. In that case, we get a perfect Bayesian equilibrium. The signal accuracy parameters are then given by

$$\phi_1 = \alpha, \qquad \phi_2 = \beta. \tag{28}$$

For the fully dynamic setting some conditions need to be satisfied for LFTs to be unable to learn and for the learning of ATs from order flow to be rational and internally consistent. In particular, we need to have that signals are indeed informative of future price moves, while LFTs cannot learn anything from price moves. One can achieve this by letting prices react to public information releases and set conditions on the news release process. These conditions on public information releases and price processes are described and derived in Appendix B.

[23] Informed trade clustering may for instance arise because at some times there is more private information available than at others, or because a single informed trader slices his trading volume into smaller trades and feeds them consecutively to the market (see e.g. Admati and Pfleiderer (1988)).

7 Effectiveness of HFT Regulatory Measures

In the previous section, we have shown that liquidity provision by HFTs can lead to market freezes, mainly as a result of a lack of liquidity providers willing to absorb potentially toxic order flow. Several measures have been introduced or suggested recently for regulators to get more grip on HFTs. These include

- Transaction taxes,
- Latency restrictions,
- Make-take fees,
- Affirmative liquidity provision.

The framework introduced here helps to analyze the effectiveness of each of those proposals.

First, let us have a look at transaction taxes. Imposing an exogenous unavoidable transaction tax would be equivalent to having a larger participation cost. Obviously, if transaction taxes are only levied on HFTs, as is the case in some proposals, the cost of being an HFT goes up and being fast may not be efficient anymore. Hence, it is possible that instead of freezes, we get inefficient adoption of speed technology. It is also possible that the larger costs do not make HFTs inefficient, but merely limit the cross-subsidization from speed to freezes and hence helps to avoid freezes by leaving market share to LFTs. Even if transaction taxes are uniformly applied, HFTs will suffer relatively more if speed technology is efficient. To see this, one should realize that the relative increase in costs is higher for HFTs than LFTs as before taxes, HFT costs per unit of speed are lower (and tax costs add linearly). As a final note, one should realize that liquidity is bound to go down due to two effects. First, the competitive price $p(1)$ will not be quoted anymore as it is very likely to be loss-making in the presence of transaction taxes. Hence, the taxes will at least partially be forwarded to liquidity demanders. Second, as gains from trade are lower, there is less surplus that liquidity providers can capture and therefore, the funds available to invest in liquidity providing facilities is reduced. As a result, undercutting slows down and average spreads increase.

Second, policymakers have suggested to impose latency restrictions on HFTs. Depending on the exact form these latency restrictions take, HFTs could become more like smart ATs. Benefits from superior speed would in that case disappear, but so would the costly market freezes (assuming speed technology is efficient before latency restrictions are introduced).

Third, several exchanges by now have introduced make-take fees as an incentive scheme for liquidity providers to provide liquidity. In our model, static make-take fees would resort little effect. Such fees would lower the reservation prices of liquidity demanders, but also allow liquidity providers to continue undercutting to levels even below the fundamental value μ. Hence, static fees would merely resort a level-shift rather than substantially different behavior from market participants. One could however introduce a 'dynamic make-take fee' that becomes particularly high when markets freeze or become very illiquid. This would effectively be a tax on informed liquidity demand to benefit liquidity provision in dire times. As a consequence, expected informed trading losses are reduced (p_{liq} in the model is effectively increased) and liquidity providers are more quickly inclined to re-launch markets again.

Finally, we can have a look at affirmative liquidity provision. Affirmative liquidity provision in its strictest sense means that a liquidity provider is forced to provide liquidity at all times. However, in reality this is unrealistic. In extreme market circumstances, liquidity providers will simply refuse to provide liquidity to avoid 'catching a falling knife'. A more realistic version is that the failure to provide liquidity to the market at reasonable spreads would be met with fines.[24] Such a situation is incorporated in our model. The freeze cost parameter c_M would now also account for the severity of such fines. As c_M increases, being in a freeze becomes more expensive, which creates an incentive to reduce HFT entry, keep LFTs in the market and avoid freezes altogether. Hence, affirmative liquidity provision can help to avoid the most damaging market impact of HFTs on market freezes at the expense of liquidity in normal times. Further gains can be made if the proceeds of these fines are used to subsidize liquidity provision in a freeze as with the dynamic make fees.[25]

8 Conclusion

In this paper, we analyze the consequences of the emergence of high-frequency traders (HFTs), complementing or replacing the traditional liquidity providers on financial markets. Our framework of analysis is a dynamic limit order book model in which HFTs compete for incoming uninformed and informed order flow with low-frequency traders (LFTs), such as traditional market makers or institutional investors. HFTs are modeled to be superior over LFTs in two dimensions (which correspond to practice). First, HFTs have a speed advantage, enabling them to submit limit orders at higher frequencies than LFTs. Secondly, only HFTs possess the information-processing technology to make real-time inferences on "hard information" (such as transaction times).

[24]Note that this type of affirmative liquidity provision was also practiced in the past with the NYSE specialist.

[25]The main practical difficulty may be that, when affirmative liquidity provision is introduced on a market, HFTs and the majority of the trading may move to less regulated venues. Therefore, in order for this approach to be effective it is crucial that such legislation is introduced in a coordinated way.

Our findings indicate that an increase in the number/speed of HFTs improves market liquidity in the absence or with low levels of informed trading, which is in line with the early empirical literature on HFTs. Yet, the synergy between the speed and the information-processing technologies which is naturally inherent to HFTs, can make market liquidity less stable over time. Interestingly, it is speed superiority, the feature that has the largest potential benefit for improving market liquidity, that amplifies asymmetric information problems to the point where markets stop functioning when suspicions of informed trading are high. As such, HFTs can trigger periods of market failure that could not take place when market participants were only fast or possessed only superior information processing technology. Only LFTs could keep the market going, yet they have been largely pushed out of the maket for liquidity provision. As such, our model captures the potential systemic risk HFT activity brings to financial markets. Temporary market freezes could arise with increasing frequency in equilibrium as HFTs gain a larger market share and get access to more efficient technology. Our framework also allows to verify the effectiveness of several proposed (or implemented) regulatory measures to manage HFT activity in practice (such as financial transaction taxes, minimum latency requirements, make-take fees, and affirmative liquidity provisions).

The selection of the starting point of our investigation (i.e., how the HFT emergence affects liquidity provision by traditional market makers or institutional investors) is driven by the general concern that HFTs are consistently front-running slower LFTs. The LFTs are thus forced to also make costly investments to lower their latency and improve their information-processing capacity, or move out of the market for liquidity provision as evidenced by our model. In a broader perspective, and beyond the specific scope of our model, in itself this may entail other repercussions for market stability in the short run. In particular, during periods of market stress, long-term institutional investors typically function as market stabilizers withstanding short-term volatility, and the business model of traditional market makers allows easier cross-subsidization between periods of calm and stress. HFTs on the other hand, are reluctant to carry risky inventory positions for longer than some minutes as they are thinly-capitalized (Kirilenko, Kyle, Samadi and Tuzun, 2011). Moreover, they have no affirmative obligation to make markets over time and tend to retract in bad times as evidenced by the flash crash (CFTC-SEC, 2010).[26] Furthermore, in the long run, LFTs might also experience reduced profitability through other channels, as they are hampered in their portfolio choice and face more systemic risk in the markets. As such, LFTs may be hindered in their role as long-term risk takers in the mobilization of savings (e.g. pension funds

[26]Notably, this is precisely what exacerbated the vicious liquidity spiral during the May 2010 flash crash. After having swallowed an unusually large initial liquidity shock, HFTs were still lacking sufficient demand from fundamental buyers or cross-market arbitrageurs, and started rapidly buying and reselling future contracts to each other. In turn, this created broader contagion effects causing equity markets to instantly dry up.

dealing with the aging of society) and in the financing of the economy.

References

Admati, A. and P. Pfleiderer, 1988, A Theory of Intraday Patterns: Volume and Price Variability, Review of Financial Studies 1, pp. 3-40.

Aït-Sahalia, Y. and M. Saglam, 2013, High Frequency Traders: Taking Advantage of Speed, NBER Working Paper.

Angel, J. and D. McCabe, 2010, Fairness in Financial Markets: The Case of High Frequency Trading, Working Paper.

Bernales, A. and J. Daoud, 2013, Algorithmic and High Frequency Trading in Dynamic Limit Order Markets, Working Paper.

Beucke, D., 2012, BATS: The Epic Fail of the Worst IPO Ever, Bloomberg Businessweek: Markets & Finance, March 23, 2012.

Biais, B., Foucault,T. and S. Moinas, 2013, Equilibrium Fast Trading, Working Paper.

Biais, B., Hillion, P., and Spatt, C. (1995). An empirical analysis of the limit order book and the order flow in the Paris Bourse, Journal of Finance 50, 1655–1689.

Biais, B., Hombert, J. and P.-O. Weill, 2010, Trading and Liquidity with Limited Cognition, Working Paper.

Biais, B. and P. Woolley, 2011, High Frequency Trading, Working Paper.

Brogaard, J., Hendershott, T. and R. Riordan, 2013, High Frequency Trading and Price, Discovery, Forthcoming Review of Financial Studies.

Budish, E., Cramton, P. and J. Shim, 2013, The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response, Working Paper.

CFTC and SEC, 2010, Commodity and Futures Trading Commission and Securities and Exchange Commission, Findings Regarding the Market Events of May 6, 2010, Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues (September 30, 2010).

Cordella, T. and T. Foucault, 1999, Minimum Price Variations, Time Priority, and Quote Dynamics, Journal of Financial Intermediation 8, pp. 141-173.

Dugast, J. and T. Foucault, 2013, False News, Informational Efficiency, and Mini Flash Crashes, Working Paper.

Easley, D., Lopèz de Prado, M. and M. O'Hara, 2012, Flow Toxicity and Liquidity in a High Frequency World, Review of Financial Studies 25, pp. 1457-1493.

ESMA, 2011, Final Report - Guidelines on Systems and Controls in an Automated Trading Environment for Trading Platforms, Investment Firms and Competent Authorities.

European Commission MEMO/11/716, 2011, New Rules for More Efficient, Resilient and Transparent Financial Markets in Europe.

Foucault, T., 1999, Order Flow Composition and Trading Costs in a Dynamic Limit Order Market, Journal of Financial Markets 2, pp. 99-134.

Foucault, T., 2012, Algorithmic Trading: Issues and Preliminary Evidence, Chapter 1 in Abergel, F., Bouchaud, J.-P., Foucault, T., Lehalle, C.-A. and M. Rosenbaum (eds.), Market Microstructure: Confronting Many Viewpoints, Wiley.

Foucault, T., Hombert, J. and I. Roşu, 2013, News Trading and Speed, Working Paper.

Foucault, T., Kadan, O. and E. Kandel, 2005, Limit Order Book as a Market for Liquidity, Review of Financial Studies 18, pp. 1171-1217.

Goettler, R., Parlour, C. and U. Rajan, 2005, Equilibrium in a Dynamic Limit Order Market, Journal of Finance 60, pp. 2149-2192.

Goettler, R., Parlour, C. and U. Rajan, 2009, Informed Traders and Limit Order Markets, Journal of Financial Economics 93, pp. 67-87.

Golub, A., Keane, J. and S.-H. Poon, 2012, High Frequency Trading and Mini Flash Crashes, Working Paper.

Haldane, A.G., 2011, The race to zero, Speech by Mr Andrew G Haldane, Executive Director, Financial Stability, of the Bank of England, at the International Economic Association Sixteenth World Congress, Beijing, July 8th, 2011.

Hasbrouck, J. and G. Saar, 2012, Low-Latency Trading, Working Paper.

Hendershott, T., 2011, High Frequency Trading and Price Efficiency, UK Government Foresight Driver Review 12.

Hendershott T., Jones, C. and A. Menkveld, 2011, Does Algorithmic Trading Improve Liquidity, Journal of Finance 66, pp. 1-33.

Hoffmann, P., 2013, A Dynamic Limit Order Market with Fast and Slow Traders, Forthcoming Journal of Financial Economics.

Johnson, N., Zhao, G., Hunsader, E., Meng, J., Ravindar A., Carran, S., and B. Tivnan, B., 2012, Financial Black Swans Driven by Ultrafast Machine Ecology, Working Paper.

Jovanovic, B. and A. Menkveld, 2011, Middlemen in Limit-Order Markets, Working Paper.

Kirilenko, A., Kyle, A., Samadi, M. and T. Tuzun, 2011, The Flash Crash: The Impact of High Frequency Trading on an Electronic Market, Working Paper.

Kyle, A., 1985, Continuous Auctions and Insider Trading, Econometrica 53, 1315–1335.

Malinova, K., Park, A. and R. Riordan, 2013, Do Retail Traders Suffer from High Frequency Traders?, Working Paper.

Martinez, V., and Roşu, I., 2011, High-Frequency Traders, News and Volatility, Working Paper.

Maskin, E., and Tirole, J. (1988). A theory of dynamic oligopoly. II. Price competition, kinked demandcurves and Edgeworth cycles, Econometrica 56, 571–599.

Menkveld, A., 2011, Electronic Trading and Market Structure, UK Government Foresight Driver Review 16.

Menkveld, A., 2012, High Frequency Trading and the New-Market Makers, Working Paper.

Menkveld A. and Z. Yueshen, 2011, The Anatomy of a Flash Crash: When Search Engines Replace Broker-Dealers, Working Paper.

Nanex, 2013, How to Destroy $45 Billion in 45 Milliseconds, http://www.nanex.net/aqck2/4197.html.

Niederauer, D., 2012, Market Structure: Ensuring Orderly, Efficient, Innovative and Competitive Markets for Issuers and Investors: Congressional Hearing Before the Subcommittee on Capital Markets and Government Sponsored Enterprises of the Committee on Financial Services US House of Representatives, 112th Congress. Congressional Testimony, Panel I. http://financialservices.house.gov/uploadedfiles/112-137.pdf.

Pagnotta, E., 2010, Information and Liquidity Trading at Optimal Frequencies, Working Paper.

Pagnotta, E. and T. Philippon, 2012, Competing on Speed, Working Paper.

Parlour, C., 1998, Price Dynamics in Limit Order Markets, Review of Financial Studies 11, pp. 789-816.

Roşu, I., 2009, A Dynamic Model of the Limit Order Book, Review of Financial Studies 22, pp. 4601-4641.

Russolillo, S., 2013, Google Suffers “Mini Flash Crash” Then Recovers, Wall Street Journal, April 22. http://blogs.wsj.com/moneybeat/2013/04/22/google-suffers-mini-flash-crash-then-recovers/.

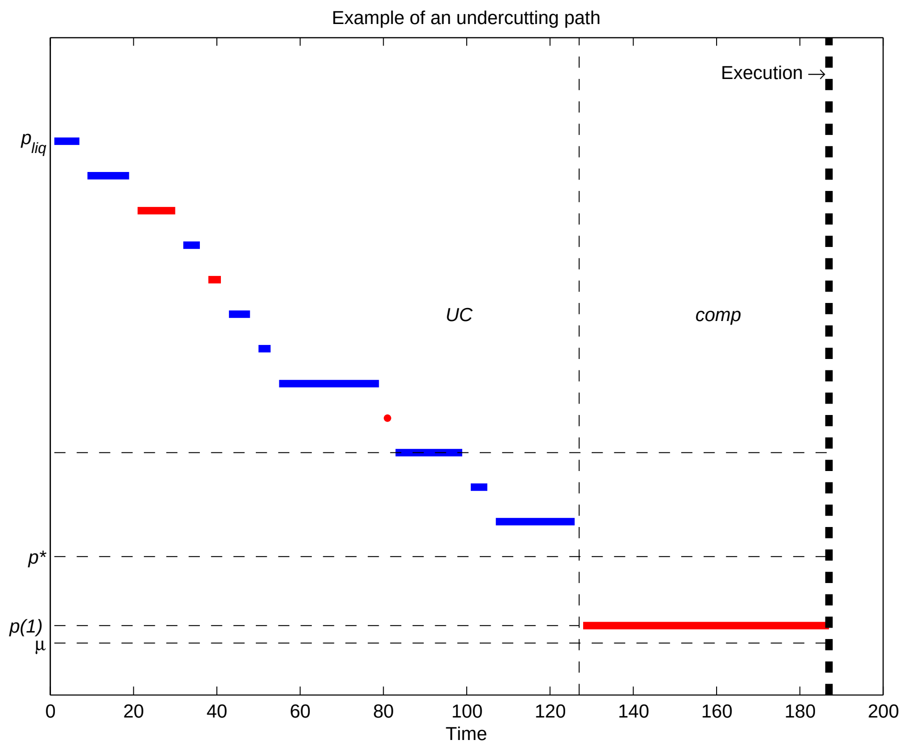
Sornette, D. and S. von der Becke, 2011, Crashes and high frequency trading, UK Government Foresight Driver Review 7.

Tabb Group, 2012, Global Equity Trends: State of the Industry Q1 2012.

UK Government’s Foresight panel, 2012, The Future of Computer Trading in Financial Markets, Working Paper.

Vlastelica, R., 2013, Symantec Shares Plunge, Traders See Mini “Flash Crash”, Reuters, April 30. http://www.reuters.com/article/2013/04/30/symantec-tradehalt-idUSL2N0DH1WK20130430.

Figure 1: **Example of an undercutting path in the uninformed setting**



The figure shows an example undercutting path when there is no asymmetric information. The x-axis shows time elapsed since the first quote has been posted, while the y-axis displays price ticks. Blue exposures are HFT exposures while red exposures are LFT exposures.

A APPENDIX - Proofs

Proof of Lemma 1. First we prove ***P1***. Suppose $\hat{a} \geq p(2)$. Posting $a \geq \hat{a}$ will lead to no execution and therefore zero payoff. Posting $a \in [p(1), \hat{a}) \cap Q$ guarantees a positive payoff as $p(1) > \mu$ and $\bar{\pi} = 0$.

Next, we prove ***P1***. Suppose $\hat{a} = p(1)$. Posting $a < p(1)$ on Q cannot be optimal, as any quote on Q falling short of $p(1)$ must be lower than μ and would therefore lead to a loss. Any quote $a \geq p(1)$ joins the queue and has zero execution probability. Hence the payoff of any such a quote is zero. Hence, $R_k^*(p(1)) = p(1)$ is (weakly) optimal.

Now we derive the expressions for execution probabilities. Assume the posted quote $a = R_k^*(\hat{a}) > p(1)$. In equilibrium, any liquidity provider arriving to the market will undercut due to ***P1***. Hence the execution probability is given by the probability that the liquidity demander arrives before another liquidity supplier. The arrival rate of liquidity suppliers is given by $\lambda(\gamma m + n)$ and is independent of k, because liquidity providers are atomistic. The arrival rate of liquidity demanders is given by ν_{liq}. Applying standard rules for the calculations with exponential distributions yields (4).

Now assume $a = R_k^*(\hat{a}) = p(1)$, where $\hat{a} > p(1)$. As it is never optimal for any liquidity supplier to undercut, execution is guaranteed and hence, $\Phi = 1$.

Q.e.d. ■

Proof of Proposition 1. From Lemma 1, it follows that undercutting is the only action undertaken in equilibrium by both ATs and LFTs. Undercutting to any quote larger than the incoming market order trader's reservation price p_{liq} is sub-optimal as it will always generate a zero payoff. Hence, upon observing a quote strictly larger than p_{liq} upon arrival, the optimal strategy is to undercut to p_{liq}. As the limit order execution probability Φ is independent of the number of ticks with which the undercutting takes place, undercutting to a quote lower than p_{liq} is always sub-optimal.

If p_{liq} or a lower quote are observed upon arrival, undercutting by one tick or undercutting to $p(1)$ are the only actions that will be undertaken in equilibrium by both ATs and LFTs. Undercutting by more than one tick to a price which is strictly larger than $p(1)$ is always sub-optimal as the limit order execution probability Φ is independent of the number of ticks with which the undercutting takes place.

Next, let us determine the threshold price at which arriving traders prefer to undercut to $p(1)$ instead of undercutting by one tick.

First, trader k faces the following trade-off. If she quotes the competitive price, she secures execution in the running iteration and obtains with certainty a profit equal to $p(1) - \mu = \frac{\delta}{2}$. If instead she undercuts by only one tick to a price $p > p(1)$, her expected payoff equals $\Phi\,(p - \mu)$ as she will be undercut by the subsequently-arriving liquidity provider. It follows that undercutting by only one tick is the best response if

$\Phi(p-\mu) \geq \frac{\delta}{2}$, implying that the exact threshold price where this inequality reverses is at $\tilde{p}^* = \mu + \frac{\delta}{2\Phi}$.

As a final step, we still need to account for the fact that $\tilde{p}^*$ may not be positioned on the price grid. To do so, denote the greatest integer strictly lower than x by $\lfloor\lfloor x \rfloor\rfloor$. Then,

$$\tilde{p}^* = \left\langle \mu + \frac{\delta}{2\Phi} \right\rangle^+ = p(1) + \left\lfloor\left\lfloor \frac{1-\Phi}{2\Phi} \right\rfloor\right\rfloor \delta, \tag{29}$$

where $\tilde{p}^*$ is the smallest price on the grid such that the inequality is satisfied.

Q.e.d. ■

Proof of Proposition 2. We will now work out the unconditional expected profits in each of the two parts along the equilibrium path.

Let us start with region UC. To facilitate exposition, let us define the random variables b as the number of ticks away from p_{liq} on which execution takes place, q_t the number of ticks the best standing quote is away from p_{liq} and t_b the time at which execution takes place. The market-wide expected aggregate profit earned in region UC is given by

$$E(\Pi^{UC}) = \sum_{i=0}^{Z} P(b=i)(p_{liq} - i\delta - \mu).$$

The probability of execution i ticks away from p_{liq} can be derived as follows. We have that

$$P(b=i) = \int_{t=0}^{\infty} P(q_t = i) P(t_b > t) \nu_{liq} dt. \tag{30}$$

The probability $P(q_t = i)$ is given by a Poisson distribution with parameter $\bar{\lambda}t$, while $P(t_b > t) = \exp(-\nu_{liq} t)$. Substituting these distribution functions into (30), we get

$$\begin{aligned} P(b=i) &= \int_{t=0}^{\infty} \frac{1}{i!}(\bar{\lambda}t)^i \exp(-\bar{\lambda}t)\exp(-\nu_{liq}t)\lambda_{liq} dt, &(31)\\ &= \int_{t=0}^{\infty} \frac{\nu_{liq}\bar{\lambda}^i}{(\nu_{liq}+\bar{\lambda})^{i+1}} \left[(\nu_{liq}+\bar{\lambda})^{i+1}\frac{1}{i!}t^i \exp(-(\nu_{liq}+\bar{\lambda})t) \right] dt. &(32) \end{aligned}$$

The part in square brackets can be recognized as the pdf of a Gamma distribution with parameters $(i+1, \nu_{liq}+\bar{\lambda})$, while all other terms are multiplicative, do not depend on t and can therefore be put in front of the integration. By definition, a pdf integrates to 1 over its support, such that we have

$$P(b=i) = \frac{\nu_{liq}\bar{\lambda}^i}{(\nu_{liq}+\bar{\lambda}^{i+1}}. \tag{33}$$

Let us now continue with the *comp* region. Let us define the probability of execution in the UC region

$$P_{UC} = \sum_{i=0}^{Z} P(b = i). \tag{34}$$

If execution takes place outside the UC region, it must take place in the *comp* region where execution is guaranteed to the first one posting a quote $p(1)$. Hence,

$$E(\Pi^{comp}) = (1 - P_{UC})(p(1) - \mu) \tag{35}$$

trivially follows.

Now we still need to show how expected aggregate profits accrue to LFTs and ATs. This depends on the expected exposures of both groups. As expected quote life is independent of trader type, the expected exposure of a group depends on how often it can be expected to post an undercutting quote relative to the other group. Hence, the fraction of time that the market is exposed to LFT quotes is given by

$$f_{LFT} = \frac{n}{n + \gamma m}. \tag{36}$$

Q.e.d. ■

Proof of Proposition 3. Define $b = \frac{m}{\gamma}$ and substitute into (10) to (15). Applying the chain rule for differentiation to get the derivatives of the expected revenue densities (10) and (11) with respect to b and n respectively gives:

$$\frac{\partial E\left(\sum_{\widehat{a}} \Pi_{LFT}(R^*_{LFT}(\widehat{a}))\right)}{\partial n} = \frac{\partial E\left(\sum_{\widehat{a}} \Pi_{AT}(R^*_{LFT}(\widehat{a}))\right)}{\partial b} = \frac{-(E(\Pi^{UC} + \Pi^{comp}))}{(n+b)^2} + \frac{\partial(E(\Pi^{UC} + \Pi^{comp}))}{\partial(n+b)} \frac{1}{n+b}. \tag{37}$$

Hence, marginal expected revenue densities are equal. On the other hand, marginal expected cost densities are given by C_L and $\frac{C_A}{\gamma}$, respectively. Hence, given $n + b$, expected revenue minus expected costs for ATs always exceeds that for LFTs if $C_L > \frac{C_A}{\gamma}$. Moreover, the partial derivatives of (10) and (11) with respect to n and b are all four strictly negative. As entry is free, it will take place as long as marginal revenue exceeds expected costs. Hence we must have for each player type in equilibrium either marginal costs equals marginal profits or participation is zero. As a result we have that $n = 0, m > 0$ if $C_L > \frac{C_A}{\gamma}$ and $n > 0, m = 0$ if $C_L < \frac{C_A}{\gamma}$.

Q.e.d. ■

Proof of Proposition 4. Let us define the event B that a specific LFT arrives to an empty order book, let the event S denote suspicion from the ATs and NS no suspicion from the ATs. Then Bayes rule gives

$$\hat{P}(\zeta = inf|\psi_{LFT}) = P(\zeta = inf|B) = \frac{P(B|\zeta = inf)}{P(B)}, \tag{38}$$

$$P(B) = P(B|\zeta = inf) + P(B|\zeta = liq), \tag{39}$$

$$P(B|\zeta = inf) = P(B|\zeta = inf, S)P(S|\zeta = inf) + P(B|\zeta = inf, NS)P(NS|\zeta = inf), \tag{40}$$

$$P(B|\zeta = liq) = P(B|\zeta = liq, S)P(S|\zeta = liq) + P(B|\zeta = liq, NS)P(NS|\zeta = liq), \tag{41}$$

$$P(S|\zeta = inf) = \frac{P(\zeta = inf|S)P(S)}{P(\zeta = inf)}, \tag{42}$$

$$P(S|\zeta = liq) = \frac{P(\zeta = liq|S)P(S)}{P(\zeta = liq)}. \tag{43}$$

Moreover, we have that

$$P(B|\zeta = inf, S) = P(B|\zeta = liq, S) = \frac{1}{n}, \quad P(B|\zeta = inf, NS) = P(B|\zeta = liq, NS) = \frac{1}{n + \gamma m}, \tag{44}$$

$$P(\zeta = inf) = P(S) = \bar{\pi}, \qquad P(\zeta = liq) = 1 - \bar{\pi}, \tag{45}$$

$$P(NS|\zeta = inf) = 1 - P(S|\zeta = inf), \qquad P(NS|\zeta = liq) = 1 - P(S|\zeta = liq), \tag{46}$$

$$P(\zeta = inf|S) = \phi_2. \tag{47}$$

Substituting in, we get

$$\hat{P}(\zeta = inf|\psi_{LFT}) = \frac{\phi_2 \frac{1}{n} + (1 - \phi_2)\frac{1}{n+\gamma m}}{\phi_2 \frac{1}{n} + (1 - \phi_2)\frac{1}{n+\gamma m} + \frac{1}{n}\bar{\pi}(1 - \phi_2) + \left(1 - \frac{\bar{\pi}(1-\phi_2)}{1-\bar{\pi}}\right)}. \tag{48}$$

The partial derivatives (where $\phi_2 > \bar{\pi}$) are given by:[27]

$$\frac{\partial \hat{P}(\zeta = inf|\psi_{LFT})}{\partial m} = \frac{n\gamma(1 - \bar{\pi})(\phi_2 - \bar{\pi})}{(2n(-1 + \bar{\pi}) + m\gamma(-\phi_2 + \bar{\pi}(-1 + 2\phi_2)))^2} > 0, \tag{49}$$

$$\frac{\partial \hat{P}(\zeta = inf|\psi_{LFT})}{\partial \gamma} = \frac{mn(1 - \bar{\pi})(\phi_2 - \bar{\pi})}{(2n(-1 + \bar{\pi}) + m\gamma(-\phi_2 + \bar{\pi}(-1 + 2\phi_2)))^2} > 0, \tag{50}$$

$$\frac{\partial \hat{P}(\zeta = inf|\psi_{LFT})}{\partial \phi_2} = \frac{m\gamma(1 - \bar{\pi})(n + m\gamma\bar{\pi})}{(2n(-1 + \bar{\pi}) + m\gamma(-\phi_2 + \bar{\pi}(-1 + 2\phi_2)))^2} > 0, \tag{51}$$

$$\frac{\partial \hat{P}(\zeta = inf|\psi_{LFT})}{\partial n} = \frac{-m\gamma(1 - \bar{\pi})(\phi_2 - \bar{\pi})}{(2n(-1 + \bar{\pi}) + m\gamma(-\phi_2 + \bar{\pi}(-1 + 2\phi_2)))^2} < 0, \tag{52}$$

$$\frac{\partial \hat{P}(\zeta = inf|\psi_{LFT})}{\partial \bar{\pi}} = \frac{-m\gamma(1 - \phi_2)(n + m\gamma\phi_2)}{(2n(-1 + \bar{\pi}) + m\gamma(-\phi_2 + \bar{\pi}(-1 + 2\phi_2)))^2} < 0. \tag{53}$$

[27] Calculations performed by Mathematica

If ATs do not employ a differential strategy upon observing an informed trade (i.e. ATs always or never submit a first quote), LFTs cannot learn anything about the state of the world from observing an empty book and we have that $P(\zeta = inf|B) = \bar{\pi}$.

Q.e.d. ■

Proof of Proposition 5. Combining Proposition 3 with the observation leading to (16) in section 4.2.1, we have that $m > 0, n = 0$ if $\tilde{C}_L > \frac{\tilde{C}_A}{\gamma}$ and $m = 0, n > 0$ if $\tilde{C}_L < \frac{\tilde{C}_A}{\gamma}$. (16) is obtained by applying the same function $f(x) = \frac{x}{1-\bar{\pi}} + \frac{\bar{\pi}}{(1-\bar{\pi})(n+\gamma m)}(\mu_{inf} - p_{liq})$ to both C_L and C_A. Because $f(x)$ is linear with strictly positive coefficient on the linear term, $f(x)$ is strictly increasing. Hence, rank ordering of input is preserved. Therefore, $\tilde{C}_L > \frac{\tilde{C}_A}{\gamma}$ iff $C_L > \frac{C_A}{\gamma}$ and $\tilde{C}_L < \frac{\tilde{C}_A}{\gamma}$ iff $C_L < \frac{C_A}{\gamma}$.

Q.e.d. ■

Proof of Proposition 6. Let us assume that (21) is satisfied in equilibrium. In this setting, ATs always quote in an empty book and hence we resort to the case with speed only. Due to Proposition 5, liquidity is provided exclusively by the player type with lowest adjusted cost, i.e. ATs when $\min\left(\frac{\tilde{C}_A}{\gamma}, \tilde{C}_L\right) = \frac{\tilde{C}_A}{\gamma}$ and LFTs otherwise. (24) then ensures that (21) is satisfied in equilibrium.

Q.e.d. ■

Proof of Proposition 7. Because $g_A(\cdot)$ is strictly decreasing, it equals $\frac{C_A}{\gamma}$ at only one point. If at this point marginal expected profit for LFTs is strictly positive, LFTs will take over the whole market due to the fact that $g_A(\cdot)$ and $g_L(\cdot)$ are strictly decreasing in b and n (LFTs can enter, causing ATs to leave, which in turn attracts more LFTs, etc.).

If entry of ATs is profitable whenever LFT marginal profit equals zero, ATs will enter at the expense of LFTs. However, because of Proposition 4, (22) will bind. At this point, AT marginal profit must be strictly positive, while LFT marginal profit equals zero. Moreover, increasing b would be infeasible, due to the discontinuity in AT marginal profit. Hence, we must be in $(\bar{b}, \bar{n})$.

Under the same conditions, but with $h_A(\bar{b}, \bar{n}) > 0$, entry of ATs is also optimal in $(\bar{b}, \bar{n})$. As marginal profit for ATs and LFTs is strictly negative in b and n, we must have that liquidity is exclusively provided by ATs. Because LFTs do not participate and ATs stay away from an empty book when $s = inf$, freezes take place. The incidence rate of $s = inf$ is $\bar{\pi}$ due to the unconditionally unbiased nature of the signal. If $\bar{\pi} c_M > \phi_2(\mu_{inf} - p_{liq})$ as assumed, freezes cost more than is saved by avoiding toxic order-flow. This is inefficient.

Q.e.d. ■

B Internally consistent news announcements

In the dynamic extension of the model, we need to make sure that price movements are consistent with informed trading. In other words, it is important that prices move in the direction of the information in the market when the state of nature switches from inf to liq. However, we want to prevent LFTs from learning from price paths to keep tractability. To this end, we assume that public information can be released between iterations. In particular, we assume that information releases always occur if ζ_l switches from informed to uninformed, such that the efficient price μ can be updated to the value μ_{inf} from last period. Moreover, we assume that information from either side of the book is impounded in prices in a similar way such that there is no price drift up or down.[28] In order to have that information releases contain no information about ζ_l, certain conditions about the frequencies of public information releases need to be satisfied. Let us define the event A_l as a public information release (announcement) between iteration $l-1$ and l.

Assumption 1 *(Announcement uninformativeness) When the state of nature switches from inf to liq, public information is released (i.e. $P(A_l|\zeta_{l-1} = inf, \zeta_l = liq) = 1$). Moreover, information releases satisfy the following constraint*

$$\beta(1-\pi)P(A_l|\zeta_l = inf, \zeta_{l-1} = inf) + (1-\alpha)(1-\bar{\pi})(\frac{1}{\bar{\pi}} - 1)P(A_l|\zeta_l = inf, \zeta_{l-1} = liq) = (1-\beta)\bar{\pi} + \alpha(1-\bar{\pi})P(A_l|\zeta_{l-1} = liq, \zeta_l = liq) \quad (54)$$

Under assumption 1, we show below that public information releases are uninformative about the state of nature ζ_l. Note that the assumptions in this paragraph are not necessary to obtain our main results, but merely to show that the setup of our model is internally consistent.

In order to have information asymmetry that is consistent with future price movements, we have under assumption 1 that

$$P(A_l|\zeta_{l-1} = inf, \zeta_l = liq) = 1. \quad (55)$$

Moreover, we want the event A_l to be uninformative about the state of nature (to LFTs). This is the case when

[28]For tractability reasons, we refrain from also explicitly modeling the other side of the book.

$$P(\zeta_l = inf|A_l) = P(\zeta_l = inf) \ \rightarrow \tag{56}$$

$$\frac{P(A_l|\zeta_l = inf)P(\zeta_l = inf)}{P(A_l)} = P(\zeta_l = inf) \ \rightarrow \tag{57}$$

$$P(A_l|\zeta_l = inf) = P(A_l). \tag{58}$$

The only thing left to do now is to work out this constraint in terms of public news release probabilities for each type of transition. We can work out $P(A_l|\zeta_l = inf)$ first:

$$\begin{aligned} P(A_l|\zeta_l = inf) = P(A_l|\zeta_l = inf, \zeta_{l-1} = inf)P(\zeta_{l-1} = inf|\zeta_l = inf)+ \\ P(A_l|\zeta_l = inf, \zeta_{l-1} = liq)P(\zeta_{l-1} = liq|\zeta_l = inf). \end{aligned} \tag{59}$$

Applying Bayes rule twice, we have

$$P(\zeta_{l-1} = inf|\zeta_l = inf) = \frac{P(\zeta_l = inf|\zeta_{l-1} = inf)P(\zeta_{l-1} = inf)}{P(\zeta_l = inf)} = \frac{\beta\bar{\pi}}{\bar{\pi}} = \beta, \tag{60}$$

where $\bar{\pi} = \frac{1-\alpha}{2-\beta-\alpha}$, the long-term (unconditional) steady state probability of being in the informed state of nature. Similarly, we have

$$P(\zeta_{l-1} = liq|\zeta_l = inf) = \frac{(1-\alpha)(1-\bar{\pi})}{\bar{\pi}}. \tag{61}$$

Substituting these expressions into (59), we get

$$\begin{aligned} &P(A_l|\zeta_l = inf) = \\ &\quad P(A_l|\zeta_l = inf, \zeta_{l-1} = inf)\beta + P(A_l|\zeta_l = inf, \zeta_{l-1} = liq)(1-\alpha)(\frac{1}{\bar{\pi}} - 1). \end{aligned} \tag{62}$$

Similarly, we can work out $P(A_l)$ as

$$\begin{aligned} P(A_l) = P(A_l|\zeta_{l-1} = inf, \zeta_l = inf)P(\zeta_{l-1} = inf, \zeta_l = inf)+ \\ P(A_l|\zeta_{l-1} = inf, \zeta_l = liq)P(\zeta_{l-1} = inf, \zeta_l = liq)+ \\ P(A_l|\zeta_{l-1} = liq, \zeta_l = inf)P(\zeta_{l-1} = liq, \zeta_l = inf)+ \\ P(A_l|\zeta_{l-1} = liq, \zeta_l = liq)P(\zeta_{l-1} = liq, \zeta_l = liq). \end{aligned} \tag{63}$$

Working out basic statistical identities, we have

$$P(\zeta_{l-1} = inf, \zeta_l = inf) = P(\zeta_l = inf|\zeta_{l-1} = inf)P(\zeta_{l-1} = inf) = \beta\bar{\pi}, \tag{64}$$

and similarly

$$P(\zeta_{l-1} = inf, \zeta_l = liq) = (1-\beta)\bar{\pi}, \tag{65}$$

$$P(\zeta_{l-1} = liq, \zeta_l = inf) = (1-\alpha)(1-\bar{\pi}), \tag{66}$$

$$P(\zeta_{l-1} = liq, \zeta_l = liq) = \alpha(1-\bar{\pi}). \tag{67}$$

Substituting everything into (58) and realizing that probabilities must be contained in the unit interval, any set of announcement probabilities satisfying the following set of constraints can be allowed:

$$\beta(1-\pi)P(A_l|\zeta_l = inf, \zeta_{l-1} = inf) + (1-\alpha)(1-\bar{\pi})(\frac{1}{\bar{\pi}} - 1)P(A_l|\zeta_l = inf, \zeta_{l-1} = liq) = \\ (1-\beta)\bar{\pi} + \alpha(1-\bar{\pi})P(A_l|\zeta_{l-1} = liq, \zeta_l = liq) \tag{68}$$

and

$$P(A_l|\zeta_l = inf, \zeta_{l-1} = inf) \in [0,1], \tag{69}$$

$$P(A_l|\zeta_l = inf, \zeta_{l-1} = liq) \in [0,1] \tag{70}$$

$$P(A_l|\zeta_{l-1} = liq, \zeta_l = liq) \in [0,1]. \tag{71}$$

C Notation Summary

Parameters		
Symbol	*Support*	*Description*
Q	$-$	price grid
δ	$(0, \infty]$	tick size
$p(i)$	Q	price level on the grid
μ	$(0, \infty]$	fundamental value conditional on public information only
p_{liq}	$(\mu, \infty]$	reservation price liquidity demanders
μ_{inf}	$(p_{liq}, \infty]$	true value of the asset in the informed state
$\hat{a}$	Q	standing best quote upon arrival
C_k	$(0, \infty]$	participation costs
c_M	$[0, \infty]$	freeze costs
λ	$[0, \infty]$	arrival intensity liquidity providers
γ	$[1, \infty]$	speed advantage of ATs
ν_{inf}, ν_{liq}	$[0, \infty]$	arrival intensities for informed and uninformed liquidity demanders respectively
ϕ_1, ϕ_2	$(0.5, 1]$	accuracy of signals $s = liq$ and $s = inf$ respectively
$\bar{\pi}$	$[0, 1]$	(unconditional) probability of $\zeta = inf$ state
α, β	$[0, 1]$	transition probabilities of staying in the liq and inf states respectively (dynamic extension only)
States of nature		
$\tilde{V}$	$\{\mu_{inf}, \mu\}$	Asset payoff
ζ	$\{inf, liq\}$	state of nature/liquidity demander type
s	$\{inf, liq\}$	signal about state of nature
ψ_k	$-$	information set
Indices		
k	$\{A, L\}$	liquidity provider type
i	$\{0, .., \infty\}$	ticks
t	$[0, \infty]$	time
l	$\{1, .., \infty\}$	iteration (i.e. stage game; dynamic extension only)
Decision variables		
m, n	$[0, 1)$	masses of ATs and LFTs respectively
a	Q	price quote to be submitted