High-Frequency Trading and Market Stability*

Dion Bongaerts† and Mark Van Achter‡

First version: March 2013
This version: November 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 liquidity providers in the fact that they combine speed and information processing. We compare a setting with HFTs to settings with liquidity providers 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, endogenous entry barriers and rents, or to 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.

JEL Codes: D53, G01, G10, G18

Keywords: Market Freeze, Liquidity Dry-Up, Systemic Risk, Latency, Informed Trading, Allocative Efficiency.

*We would like to thank Jean-Edouard Colliard, Sugato Chakravarty, Hans Degryse, Jérôme Dugast, Frank de Jong, Thierry Foucault, Nicolae Gârleanu, Terry Hendershott, Johan Hombert, Pete Kyle, Katya Malinova, Albert Menkveld, Sophie Moinas, Christine Parlour, Ioanid Roşu, conference participants at the 2015 Conference on the Industrial Organization of Securities Markets (Frankfurt), the 2015 CIFR Symposium Celebrating the 30 Years Since Kyle Met Glosten & Milgrom (Sydney), 2014 FIRS Conference (Quebec), the 2014 EFA Conference (Lugano) and seminar participants at HEC Paris and Erasmus University Rotterdam for helpful comments and suggestions. This paper was awarded with the 2015 De la Vega prize granted by the Federation of European Securities Exchanges. 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.[2] 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.

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).[3] 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.[4] Hence, policy makers and regulators have become increasingly concerned that HFT-based liq-

[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 (2012)).

[2] They exploit e.g. short-lived information on order book dynamics, trade dynamics, past stock returns, cross-stock correlations, cross-asset correlations and cross-exchange information delays. See Brogaard (2011a) for a further discussion of the different types of short-term information used by HFTs.

[3] 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.

[4] 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 Brogaard, Moyaert and Riordan (2014), 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.

uidity provision could come at the expense of an evaporation of liquidity when it is most needed (see e.g. CFTC-SEC (2010) and Niederauer (2012)).[5]

Our paper addresses exactly this concern. More specificially, we analyze whether or not HFTs (i) destabilize financial markets, (ii) improve liquidity provision, 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.[6] 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 (as e.g. documented in Korajczyk and Murphy (2015)). Nowadays, the simultaneous investment in both technological advances (which is the setup closest to real-life HFTs) generates synergy benefits for the HFTs which are unprecedented. Historically, such benefits have been much smaller or even non-existent.[7]

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 invest in *speed technology only* which allows to monitor the market faster. We demonstrate that if this technology is most cost-efficient[8], fast liquidity providers take over the whole market, while nobody adopts the new technology if it is too expensive. Overall, this outcome is allocatively efficient given the available technology. 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[9], which will then end up with the non-users. As such, LFTs bear disproportionally large adverse selection losses when

[5]As a further example of the increased regulatory scrutiny regarding HFTs, the European Commission has included the analysis of HFTs in its review of MiFID. In order to prevent systemic risk created by HFTs, it considers the possibility to subject them to regulatory oversight and capital requirements.

[6]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 are limit orders. Jovanovic and Menkveld (2011) find that the HFT they are focusing on is on the passive side in 70% to 80% of the transaction it conducts on Chi-X and Euronext.

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

[8]With "most cost-efficient" we mean that the ratio of speed over technology cost is more favorable for advanced traders than for LFTs. Throughout the paper, "efficiency" is considered at speed per cost level for individual liquidity providers, and in an allocative sense for the entire economy.

[9]As such, they are able to mitigate their exposure to the risk of being picked off (Copeland and Galai (1983)). This effect is also documented for liquidity-providing HFTs in Aït-Sahalia and Saglam (2013), Hoffman (2014), Jovanovic and Menkveld (2011), and Menkveld and Zoican (2015).

providing liquidity to "toxic order flow".[10] If this technology is sufficiently attractive (i.e., due to a small additional participation cost, accurate predictions on informed order flow, intense uninformed trading or low informed trading losses), we find that some traders will indeed adopt it. Interestingly, though, not all LFTs will do so in equilibrium. The reason is that if too many liquidity-providing traders become informationally advanced, LFTs will completely leave the market. As a result, there will be no liquidity suppliers to absorb (alleged) incoming informed order flow and costly market freezes arise. Therefore, the endogenous adoption rate of such technology is limited. On the bright side, this prevents freezes from occurring in equilibrium. However, it also allows the informationally advanced traders to pocket rents, while being less efficient at providing liquidity than LFTs. Moreover, because of the arising endogenous entry barrier, there is limited participation and insufficient competition. Liquidity suffers from both the allocative inefficiency as well as the constrained competition. As a third step, we explore a setting in which traders can opt to invest in technology that *combines speed and information superiority*, which reflects the bundled package HFTs dispose of. The overall effect of this setting depends on whether speed technology is cost-efficient or not. If speed technology is cost-inefficient, synergy benefits between speed and information technology actually increase the adoption likelihood as profits from informational superiority may cross-subsidize the high speed costs. In this case, market freezes do not occur and liquidity degrades for the same endogenous barriers to entry limiting competition as indicated in the second step. However, if speed technology is cost-efficient, the gains from speed superiority may create an allowance for the costs resulting from market freezes. All liquidity is then provided by HFTs, which are the most cost-efficient providers, but still falls short of that in the fully efficient setting. The reason is that internalizing the freeze losses reduces participation and therefore liquidity.

The resulting main insights can be summarized as follows. First, allowing LFTs to invest in speed technology only yields allocatively 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. The severeness of these problems is limited as market freezes are prohibitively expensive and are therefore avoided. Yet, endogenous barriers to entry arise. These limit competition and allow informationally advanced traders to pocket rents while offering inferior liquidity provision services. Third, if LFTs are allowed to purchase speed and information processing technology simultaneously (i.e., become HFTs), the overall impact hinges on the cost-efficiency of the speed technology. If speed technology is cost-inefficient, cross-subsidization from informational gains can nonetheless lead to its adoption. Market freezes in this case do not materialize and HFTs are even able to

[10] See also Biais, Declerck and Moinas (2014), Easley, Lopèz de Prado and O'Hara (2012), Han, Khapko and Kyle (2014) and Malinova, Park and Riordan (2013).

pocket rents. If, in turn, speed technology is cost-efficient enough, adoption rates can grow so large that costly market freezes can occur in equilibrium. These freezes limit participation, and therefore lower liquidity.

Korajczyk and Murphy (2015) provide empirical evidence that in normal times, HFTs take on the bulk of liquidity provision. Yet, in stressful periods, HFTs reduce their liquidity provision significantly, while the liquidity provision of designated market makers (i.e., LFTs) remains mostly unchanged.[11] Our model fully corroborates with their results, and provides a theoretical rationale for their distressing storyline. More specifically, 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 informationally efficient price.[12] However, a different storyline unfolds when suspicions of informed trading are high. In such situations, HFTs will shun the market as documented in Korajczyk and Murphy (2015), 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 crowded 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. 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.[13]

Our model also yields insights on how financial markets should be optimally organized and regulated to alleviate the potential market stability concerns HFTs bring. In particular, we assess the effectiveness of several proposed (or implemented) regulatory measures to manage HFT activity: (i) a financial transaction tax, (ii) minimum latency requirements, (iii) the introduction of (contingent) make-take fees, and (iv) affirmative liquidity provisions. Those measures are shown to affect the equilibrium number of HFTs and LFTs (and as such, the aforementioned trade-off between liquidity and systemic risk) in different ways. Furthermore, in an extension we explore a fully dynamic setting of the model that leads to similar conclusions. This setting allows us to endogenize the information production process of information technology at the cost of added

[11] Korajczyk and Murphy (2015) analyze liquidity provision to large institutional trade packages. These are often split throughout the day to avoid detection by other market participants, but maybe mistakenly interpreted by HFTs as sequential informed trades.

[12] These findings indeed concur with the early 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).

[13] This finding also 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 (2015), Bongaerts, Kong and Van Achter (2015) and Budish, Cramton and Shim (2013).

complexity. More specifically, technologically-advanced traders are able to 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.

Our model contributes to different strands in the literature. First, it relates to the literature assessing the impact of HFT activity on financial markets. While numerous theoretical contributions emerged in recent years on this topic[14], only a very limited number of recent papers document the channels through which the emergence of HFTs induces increased systemic risk on financial markets. In particular, Aït-Sahalia and Saglam (2014) study a single HFT optimizing his trading strategy accounting for real-time order flow information. They show that this strategic HFT will provide less liquidity when volatility increases, inducing markets to become fragile in volatile times. Cespa and Vives (2015) analyze how the introduction of fast dealers (which are continuously present in the market) may cause liquidity shocks to propagate over time. Dugast and Foucault (2014) explore a speculator's trade-off between trading fast or slow upon a signal. They show that cheaper fast trading technologies induce mini-flash crashes to occur at higher frequency. Weller (2013) highlights a mutually-beneficient specialization equilibrium with HFTs supplying inexpensive fast liquidity and LFTs providing risk-bearing services at higher cost. Yet, his model still yields liquidity dry-ups under extreme market circumstances. Our model allows liquidity suppliers to choose between an HFT and an LFT role. Introducing this endogenous participation choice yields novel insights on the development, allocative efficiency and fragility of the current liquidity provision industry.

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 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 or slow. 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.

[14]See e.g. Baldauf and Mollner (2015), Baruch and Glosten (2013), Bernales and Daoud (2013), Biais, Foucault and Moinas (2015), Biais, Hombert and Weill (2010), Bongaerts, Kong and Van Achter (2015), Budish, Cramton and Shim (2015), Colliard (2013), Foucault, Hombert and Roşu (2015), Garriott (2015), Han, Khapko and Kyle (2014), Hoffmann (2014), Jovanovic and Menkveld (2015), Li (2014), Martinez and Roşu (2011), Menkveld and Zoican (2015), Pagnotta (2010), Pagnotta and Philippon (2015), and Roşu (2015).

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 presents an analysis of some regulatory measures, while Section 7 provides extensions and robustness checks of the model. 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 D.

2 Setup

The model consists of two stages. In the first stage (participation stage) players decide on participation or not. In the second stage (trading stage), the participating players compete for incoming order flow. We start by presenting the market in the trading stage.

The market is characterized by a limit order book for a security with payoff $\widetilde{V}$. Given the available *public* information on this asset, its fundamental value 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 smaller δ implies a finer grid. On the grid, as a notational convention, 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 $p(-i) = \langle \mu \rangle^- - i \cdot \delta$, and $i \in \mathbb{N}$. We assume that $\mu - \langle \mu \rangle^- = \langle \mu \rangle^+ - \mu = \frac{\delta}{2}$ (i.e., 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.[15] 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.

Over time, which is continuous and 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 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 arriving liquidity-demanding trader 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, and $\widetilde{T}$ is exponentially distributed with intensity ν_{liq}. In turn, if $\zeta = inf$, with intensity ν_{inf} an informed trader

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

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.[16] 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 in the participation stage (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.[17] 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.[18] The derivation of the number of ATs and LFTs is closely related to the average liquidity level in the book (that is, the average effective spread) that realizes in the trading stage and which is labeled S. Moreover, as will be further detailed below, during the trading game the four trader types differ in two other respects: (i) the magnitude of their 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. In turn, 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.

[16] 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.

[17] 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.

[18] Note that when $m + n < 1$, some traders simply choose not to participate. 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.

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.

In the trading stage, liquidity providers arrive randomly over time 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 submit limit orders (as documented in e.g., Baron et al. (2014), Brogaard et al. (2014), Hagströmer and Nordén (2013), and Hendershott and Riordan (2013)), 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.[19] 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. These signals are assumed to be identical for all ATs. Signals $s = liq$ and $s = inf$ are correct with probabilities ϕ_1 and ϕ_2, 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. This assumption imposes the parameter restriction $\phi_1 = 1 - \frac{\bar{\pi}(1-\phi_2)}{1-\bar{\pi}}$. We assume that $\min(\phi_1, \phi_2) \in (0.5, 1]$, such that the signals are guaranteed to be informative.

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.[20] In particular, every time the market freezes, the mass of advanced traders incurs a cost mC_F, to be split equally among all constituents. Hence, upon the occurrence of a freeze, ATs face an additional cost density of C_F. We normalize freeze costs for LFTs to zero.

3 Equilibrium

The aim of this section is to provide a formal definition of the equilibrium. First, AT and LFT limit order placement strategies are characterized. Such a strategy is a

[19] This assumption is in line with Aït-Sahalia and Saglam (2013), Hoffman (2014), and Jovanovic and Menkveld (2011). It is empirically validated in Brogaard et al. (2014) and Malinova, Park and Riordan (2013).

[20] 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 (see e.g. Kirilenko et al. (2011), and Biais and Foucault (2014)).

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 of type k 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, 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. 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.

3.2 Trading Equilibrium Definition

Upon arrival to the book, each trader $k \in \{AT, LFT\}$, optimizes her reaction function R_k to the current best quote $\widehat{a}$:

$$R^*_k = \arg\max_{R_k \in Q} \Pi_k(R_k, \widehat{a}) \tag{2}$$

where all traders optimally 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.[21]

The solutions of these optimization problems yield the optimal order 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

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

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

The derivation of the equilibrium number of ATs and LFTs is closely related to the average liquidity level in the book (that is, the average effective spread S). As more participation leads to more intense competition for providing liquidity, S is typically declining monotonically in $\gamma m + n$. If all expected revenues are exactly offset by investments in the most liquidity-enhancing technology, we would obtain an "allocatively efficient" spread level S^{effic}. However, as will be further detailed in Section 5, we may

[21] 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.

for instance have that endogenous barriers to entry allow for rents. While liquidity may then be provided by the most cost-efficient suppliers, participation will be lower and competition less intensive, ultimately inducing higher effective spreads. Furthermore, superior information processing technology may become so widely adopted that a substantial fraction of all informed trades can be avoided by liquidity providers altogether. As this effectively lowers operating costs, liquidity can be provided more cheaply and effective spreads can be lower. There is however an offsetting effect in 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.

Brogaard (2011a) provides a decomposition of the profitability of HFTs which is argued to be highly dependent on their superior information processing capacity. Rents may emerge from market making activities, collecting liquidity rebates, successfully performing statistical pattern detection, upholding the law of one price and potentially manipulating markets. These rents, however, are likely short-term oligopoly gains stemming from (i) the decrease in the competition for liquidity provision by crowding out adversely-selected LFTs (see Biais, Martimort, Rochet (2000)), and (ii) the limited entry of competitive HFTs. In turn, Baron, Brogaard and Kirilenko (2014) compute the average trading profits for HFTs predominantly using limit orders and argue that they do not systematically earn profits in line with our zero-profit condition. More generally, our setting corresponds to a longer-term equilibrium state with free entry reflecting the assertion that as the HFT industry matures the initial oligopoly gains will gradually dissolve. It underpins our aim to analyze the impact of HFTs on market quality and stability in a setting in which the technological advances are widely available to all market participants, and in which any externalities related to oligopoly rents are absent.

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.[22] The main features and trade-offs put forward

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

in this paper will largely extend to the unrestricted version of the informed case.

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 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.

Consider a time τ (assumed earlier than the time of arrival $\widetilde{T}$ 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:*

$$\Phi(R^*_{AT}(\widehat{a})) = \Phi(R^*_{LFT}(\widehat{a})) = \frac{\nu_{liq}}{\nu_{liq} + \lambda(\gamma m + n)} \equiv \Phi. \tag{5}$$

For a limit order undercutting to $p(1)$, we have:

$$\Phi(R^*_{AT}(\widehat{a})) = \Phi(R^*_{LFT}(\widehat{a})) = 1. \tag{6}$$

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.[23] 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 - Uninformed Trading Case). With time and price priority enforced, any market participant $k \in \{LFT, AT\}$ follows the following strategy when observing quote $\widehat{a}$ upon arrival:*

$$R_k = \begin{cases} p_{liq} & \text{if } \widehat{a} - \delta \geq p_{liq} \\ \widehat{a} - \delta & \text{if } p_{liq} > \widehat{a} - \delta \geq \tilde{p}^* \\ p(1) & \text{if } \widehat{a} - \delta < \tilde{p}^* \end{cases}, \tag{7}$$

where

$$\tilde{p}^* = \left\langle \mu + \frac{\delta}{2\Phi} \right\rangle^+ = p(1) + \left\lfloor \left\lfloor \frac{1-\Phi}{2\Phi} \right\rfloor \right\rfloor \cdot \delta \tag{8}$$

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 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

[23] 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.

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).

The early empirical literature has found that ATs in general improve market liquidity (see e.g. Brogaard, Hendershott and Riordan (2013), Hasbrouck and Saar (2012), Hendershott, Jones and Menkveld (2011), and Malinova, Park and Riordan (2013)). Lemma 1 and Proposition 1 provide insights 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 $\gamma \geq 1$. Moreover, these results confirm the Brogaard (2011b) finding that ATs are often able to position their limit orders ahead of the queue of limit orders (i.e., they post quotes at least equal to the best quotes 50% of the time and stand alone at the best quotes 19% of the time). Furthermore, they are in line with the result in Brogaard and Garriott (2015) that the liquidity improvement after HFT entry to the market is driven by the competition among HFTs.

4.2 Informed Trading Case

In this subsection, we work out the model including information asymmetry. Within the uninformed trading case, the market would be dominated by either ATs or LFTs, depending on the cost of speed (see Proposition 3). 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.[24] However, this information processing 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 $\nu_{inf} = \infty$.[25] One could think about this assumption as

[24]See also Biais, Declerck and Moinas (2014), Easley, Lopèz de Prado and O'Hara (2012), Han, Khapko and Kyle (2014) and Malinova, Park and Riordan (2013).

[25]The model can be extended to allow for more patient informed liquidity demanders, at the expense

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, 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 Information Processing Matters: Equilibria with Smart ATs and HFTs

In this section, we derive results for the cases in which ATs have access to information technology. That is, cases in which we have ATs that are only smart and cases when we have ATs that are both smart and fast (i.e. 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:

$$(p_{liq} - \mu)\hat{P}(\zeta = liq|\psi_{AT})\Phi(\zeta = liq) \geq (\mu_{inf} - p_{liq})\hat{P}(\zeta = inf|\psi_{AT})\Phi(\zeta_l = inf) \qquad (9)$$

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:

of reduced tractability and increased notational complexity. The main results will be largely unaffected as explained in Subsection 7.2.

$$\hat{P}(\zeta = inf|\psi_{AT}) = \begin{cases} \phi_2 \text{ if } s = inf, \\ 1 - \phi_1 \text{ if } s = liq. \end{cases} \tag{10}$$

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

$$\Phi(\zeta = inf) = 1, \tag{11}$$

$$\Phi(\zeta = liq) = \Phi. \tag{12}$$

Substituting these expressions and (5) into (9) and rewriting indicates that an AT will never post a quote to an empty book at all if:

$$(\gamma m + n) > \frac{\nu_{liq}(p_{liq} - \mu)\phi_1}{\lambda(\mu_{inf} - p_{liq})(1 - \phi_1)}. \tag{13}$$

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

$$(\gamma m + n) \leq \frac{\nu_{liq}(p_{liq} - \mu)(1 - \phi_2)}{\lambda(\mu_{inf} - p_{liq})\phi_2}. \tag{14}$$

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. The information set ψ_{LFT} refers to whether the order book an LFT arrives to is empty or not. 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:

$$(p_{liq} - \mu)\hat{P}(\zeta = liq|\psi_{LFT})\Phi(\zeta = liq) \geq (\mu_{inf} - p_{liq})\hat{P}(\zeta = inf|\psi_{LFT})\Phi(\zeta = inf). \tag{15}$$

Naturally, this inequality is more likely to be violated when the posterior probability of informed trading ($\hat{P}(\zeta = inf|\psi_{LFT})$) is larger, informed trading losses are larger, uninformed trading gains are lower and uninformed trading execution probabilities are lower. The posterior probability of informed trading is given by a complex expression. Some basic properties of this expression are given by the proposition below.

Lemma 2 *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 the fraction of ATs (m)and decreasing in the fraction of LFTs (n).*

Proof. See Appendix. ■

5 Profitability, Participation and Market Failure

5.1 Uninformed trading case

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 the competitive price $p(1)$. 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 respectively given by:*

$$E\left(\sum_{\widehat{a}} \Pi_{AT}(R^*_{LFT}(\widehat{a}))\right) = (1 - f_{LFT})m^{-1}(E(\Pi^{UC} + \Pi^{comp})), \tag{16}$$

$$E\left(\sum_{\widehat{a}} \Pi_{LFT}(R^*_{LFT}(\widehat{a}))\right) = f_{LFT}n^{-1}(E(\Pi^{UC} + \Pi^{comp})), \tag{17}$$

where

$$E(\Pi^{UC}) = \sum_{i=0}^{Z} \frac{\nu_{liq}\bar{\lambda}^i}{(\nu_{liq}+\bar{\lambda})^{i+1}}(p_{liq} - i \cdot \delta - \mu), \tag{18}$$

$$E(\Pi^{comp}) = (1 - P_{UC})(p(1) - \mu). \tag{19}$$

$$P_{UC} = \sum_{i=0}^{Z} \frac{\nu_{liq}\bar{\lambda}^i}{(\nu_{liq}+\bar{\lambda})^{i+1}}, \tag{20}$$

$$f_{LFT} = \frac{n}{n+\gamma m}. \tag{21}$$

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 are 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 (i.e., $(1 - P^{UC})$) times the guaranteed profit of half a tick. With the expressions in Proposition 2, we can derive 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 3 *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 3 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 cost-efficient producer of liquidity provision services. As liquidity provision at those expected revenues is not profitable for the least cost-efficient liquidity provider, the most cost-efficient liquidity providers must dominate the market.

Hence, if $\frac{C_A}{\gamma} \leq C_L$ 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. Hence, expected spreads must be at their allocatively efficient level S^{effic}, given the available technology. In case ATs are able to produce liquidity provision services at lower costs, they completely take over the market and do so at lower spreads.[26]

5.2 Informed Trading Case

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 consider settings with fast ATs and informed trading.

5.2.1 Only Speed Matters: Equilibria with Fast ATs

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 (see Proposition 1). The only difference is in the participation stage, where participation is more costly. Therefore, the equilibrium strategies are the same as the equilibrium strategies arising from the uninformed case with the following modifications to participation cost densities:

$$\tilde{C}_L = \frac{C_L + \bar{\pi}\frac{1}{n+\gamma m}(\mu_{inf} - p_{liq})}{1-\bar{\pi}}, \tag{22}$$

$$\tilde{C}_A = \frac{C_A + \bar{\pi}\frac{\gamma}{n+\gamma m}(\mu_{inf} - p_{liq})}{1-\bar{\pi}}. \tag{23}$$

[26] While existence and uniqueness of this equilibrium can be shown relatively easily, deriving closed-form solutions for the masses of ATs and LFTs as well as the average effective spread is difficult. However, closed form expressions can be obtained in closed form when we impose that $p_{liq} = \mu + 1.5\delta$. We present these results in Appendix A.

In line with the previous section, we find that the availability of speed technology in itself is good. If it is cost-inefficient, it will not be adopted and vice versa. Hence, we get efficient allocation, given the available technology. 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 4 *If LFTs can only choose to adopt speed technology, the availability of this technology never reduces liquidity. If it is cost-efficient enough, it takes over the whole market and market liquidity improves. The adopted technology is always most cost-efficient.*

Proof. See Appendix. ■

5.2.2 Information matters: Equilibria with smart ATs and HFTs

The availability of information processing technology may trigger information asymmetry problems. The extent to which those arise depends on model parameters and hence, we can get different types of equilibria. To analyze those, we need to define expected profit functions for LFTs and ATs with information technology conditional on their optimal quote posting strategies. For convenience, we set $b = \gamma m$ for the rest of this section. Due to Lemma 3 this transformation is without loss of generality.

Now, let us first consider a case in which information technology exists but changes little. We define:

$$g(b+n|\tilde{p}_{liq} = p_{liq}) = \frac{E(\Pi^{UC} + \Pi^{comp})}{n+b}, \tag{24}$$

where $\tilde{p}_{liq}$ is the starting point of the undercutting process. One can easily verify that $g'(\cdot) < 0$. Absent any information technology, $g(\cdot)$ would be the marginal (and average) expected gain from trading for LFTs and ATs.

If Equation (14) is satisfied in equilibrium, ATs always quote at an empty book as in the speed-only case. In this case, information technology is irrelevant and the most cost-efficient liquidity provider dominates the market.

Proposition 5 *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:*

$$g^{-1}\left(\min\left(\frac{\tilde{C}_A}{\gamma}, \tilde{C}_L\right)\middle| p_{liq}\right) \leq \frac{\nu_{liq}(p_{liq}-\mu)(1-\phi_2)}{\lambda(\mu_{inf}-p_{liq})\phi_2}. \tag{25}$$

Under this condition, liquidity provision is allocatively efficient.

Proof. See Appendix. ■

The idea behind Proposition 5 is as follows. If participation costs are high, competition at the opening stage is low and execution probability Φ is relatively high. With Φ high, posting a quote in an empty book is more likely to be sufficiently profitable to not care about potential adverse selection. Naturally, the hurdle to always participate is reached more easily if uninformed trades are very profitable, losses to trading with informed parties are small and information technology has low accuracy.

Let us now consider what happens if Equation (25) is violated (that is, when ATs optimally condition on their signals). To this end, we need different marginal profit functions for ATs and LFTs and we need to differentiate between scenarios in which LFTs participate and where they do not.

We define

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

Whenever Equation (25) is violated and Equation (15) is satisfied, $g_A(b+n)$ is the marginal profit from trading for ATs. In case Equation (25) and Equation (15) are both violated, the marginal profit from trading for ATs is given by:

$$h_A(b,n) = \\ (1-\bar{\pi})\left(\phi_1(1-\Phi)g(b+n|\tilde{p}_{liq}=p_{liq}-\delta)+\frac{\phi_1(p_{liq}-\mu)\Phi-(1-\phi_1)(\mu_{inf}-p_{liq})}{b}\right). \quad (27)$$

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

Let us also define

$$g_L(b,n)=g_A(n+b)+\bar{\pi}\frac{(1-\phi_2)(p_{liq}-\mu)\Phi-\phi_2(\mu_{inf}-p_{liq})}{n}. \quad (28)$$

Whenever Equation (15) is satisfied and Equation (25) is violated, $g_L(b,n)$ is the marginal profit from trading for LFTs. Whenever Equation (15) and Equation (25) are violated, marginal profit from trading for LFTs is given by:

$$h_L(b+n)=(1-\bar{\pi})\phi_1(1-\Phi)g(b+n|\tilde{p}_{liq}=p_{liq}-\delta). \quad (29)$$

One can verify that $h'_A(\cdot)<0$ on its domain. The marginal cost density for LFTs is given by C_L. For ATs, the marginal cost density equals $\frac{C_A}{\gamma}$ if Equation (15) is satisfied and equals $\frac{C_A}{\gamma}+\bar{\pi}C_F$ if Equation (15) is violated.

We now present the different possible equilibria with information technology graphically. To analyze the participation in equilibrium, we need to differentiate different

scenarios. Consider Figures 2 to 6 in the (n, b) space. In each of this figures, the red solid curve depicts the participation threshold for LFTs in an empty book. For combinations (n, b) above this line, adverse selection is too severe to participate. The other curves are indifference curves for functions $g_A = C_A$ (green asterisks), $g_L = C_L$ (purple stars), $h_L = C_L$ (cyan circles), and $h_A = C_A + \bar{\pi} C_F$ (blue diamonds). Above each of those curves, additional entry for the respective player type is not optimal while below it is. If we get an equilibrium above the red solid curve, this is a freeze equilibrium. Any equilibrium at or below the red curve does not involve freezes. In an equilibrium, additional entry is suboptimal by definition. Because LFTs can choose whether or not to participate in an empty book, LFT net profit should be 0 or participation should be 0.

If the LFT indifference curves always lie above the AT indifference curves, LFTs are more efficient ($C_L << \frac{C_A}{\gamma}$) and ATs do not participate (as is the case in, for example, Figure 2). This solution achieves maximal (allocative) efficiency and maximal participation. As a consequence, the average effective spread is at its efficient value.

For ATs there is a large discontinuity in their net profit functions due to freeze losses. If the AT indifference curve left the red solid curve lies below the indifference curve of the LFT, but right the red solid curve lies above the indifference curve of the LFT (such as in Figure 3), an equilibrium will materialize in the point where g_L hits the red solid curve. In this equilibrium, average AT profits are strictly positive, yet AT entry does not take place due to the discontinuity in marginal profit. This gives rise to an inefficiency in that there is insufficient entry (vs. the situation without information technology). This happens when C_L is close to $\frac{C_A}{\gamma}$ and $\bar{\pi} C_F$ is rather large. Except for when C_L exactly equals $\frac{C_A}{\gamma}$, this allocation is inefficient as there is some participation of the less efficient type. Moreover, in such equilibria, ATs pocket rents due to the endogenous entry barrier and insufficient participation. If $C_L < \frac{C_A}{\gamma}$, (zero-sum) informed trading gains cross-subsidize inefficient speed technology; if $C_L > \frac{C_A}{\gamma}$, the endogenous barrier to entry prevents ATs from pushing out inefficient LFTs. The allocative inefficiency and limited participation push the average effective spreads above the efficient level.

If AT indifference curves are always above the LFT indifference curves, such as in Figure 4 (this happens, for example, when $\frac{C_A}{\gamma} < C_L$ and $\bar{\pi} C_F$ is small), ATs dominate the market and we get an equilibrium with freezes at the spot where the dark blue squared curve crosses the y-axis. In this equilibrium, the gains from superior speed technology and lower adverse selection essentially create a tolerance for bearing freeze costs. If $\frac{C_A}{\gamma} < C_L$, this solution achieves allocative efficiency as only the most efficient types participate, but is suboptimal from a participation perspective. Because of anticipated freeze costs, entry from ATs is lower than it would be absent information technology. As a result, the average effective spread exceeds the efficient spread.

The equilibria described above are boundary solutions. There are also two types of

equilibria that are interior solutions; in such solutions we always have co-existence of LFTs and HFTs. In the first type of equilibrium $\frac{C_A}{\gamma} > C_L$, but the difference is not too large; this scenario is depicted in Figure 5. As a result, the purple starred indifference curve for LFTs starts out above the green asterisk indifference curve for ATs when n is large, but levels off due to adverse selection. As a result, the two curves cross each other. In this setting, there are three possible types of equilibria here. Either we have the oligopoly equilibrium at the crossing of the purple starred and the solid red curve, or an LFT only equilibrium at the maximum of n (or where the purple starred curve hits the x-axis), and finally, an instable equilibrium where the purple starred and the green asterisk curve cross each other (as at this point, marginal benefits for LFTs and ATs equal zero). The LFT only equilibrium is efficient from both an allocation and participation perspective and hence achieves the efficient average effective spread. The oligopoly equilibrium is inefficient from both perspectives as before. The instable interior equilibrium is efficient from a participation perspective, but inefficient from an allocative perspective as there is some strictly positive mass of inefficient ATs.

The other interior equilibrium is stable and materializes for example when $\frac{C_A}{\gamma} < C_L$ and $\bar{\pi}C_F$ is not too large. Figure 6 depicts this situation. In this case, starting from only LFTs, ATs keep entering, even beyond the point where the remaining LFTs refuse to participate in the opening stage. However, LFTs do participate once the undercutting starts. Because LFTs do not participate in the opening stage, the dark blue squared curve is not as steep as the light blue circled curve. At some point, they cross each other, at which point it is sub-optimal for LFTs to enter and also sub-optimal for ATs to enter. This type of equilibrium is inefficient from a participation as well as an allocative perspective. Unless C_L exactly equals $\frac{C_A}{\gamma}$, there is always participation of a less efficient type. Moreover, the anticipated freeze costs lower participation below what it should have been (as with any equilibrium with freezes). As a result, the average effective spread falls short of the efficient spread.

While it is hard to determine which scenario is currently in play, the available empirical evidence does provide some indications. Brogaard (2011b) shows HFTs are low cost competitors for LFTs which suggests that speed technology indeed is cost-efficient in practice and thus rules out a scenario as in Figure 2. More recent empirical evidence appears in line with the results in one of the other scenarios. In particular, Breckenfelder (2013) documents LFTs are crowded out of the market for liquidity provision due to the competition from cost-efficient HFTs. His results indicate that the quality of liquidity that HFTs potentially provide is of particular concern. Biais, Declerck and Moinas (2014) show the LFT crowding out to be related to asymmetric information. Their results indicate that as HFTs are less exposed to adverse selection, they execute a larger proportion of trades via passive limit orders than LFTs. Relatedly, Korajczyk and Murphy (2015) show that in stressful periods (corresponding to high suspicions of informed

trading in our model), HFTs reduce their liquidity provision significantly. In contrast, the (fairly limited) liquidity provision of LFTs remains mostly unchanged. Tong (2015) also examines whether HFTs provide a reliable source of liquidity when liquidity is most demanded by institutional investors (i.e., in case of a large buy-sell imbalance). She documents the expected short-livedness of HFT liquidity, and more importantly highlights HFT liquidity to be most expensive to institutional investors in case they exhibit a large trade imbalances. This finding is also confirmed by van Kervel and Menkveld (2015).

6 Effectiveness of HFT Regulatory Measures

Several measures have been suggested or recently implemented by regulators or trading venues to get more grip on HFTs. These include (i) transaction taxes, (ii) latency restrictions, (iii) affirmative liquidity provision, and (iv) make-take fees.[27] The framework introduced here helps to analyze the effectiveness of each of those proposals given the stated goals. We refrain from a stand on which measures maximize aggregate welfare because the objective function is somewhat unclear here. In particular, higher HFT participation could improve liquidity in normal times, but lead to market freezes in distressed times. The preference for the one vs the other depends on the relative benefit of liquidity in good times vs. freezes in bad times.

First, let us have a look at *transaction taxes*. Imposing an exogenous unavoidable transaction tax would lower expected revenues. Using a similar argument as in Section 5.2.1 this is equivalent to having a larger participation cost. Obviously, if transaction taxes are only levied on HFTs (as occurs in France, where an HFT tax was adopted in August 2012) the cost of being an HFT goes up and being fast may not be cost-efficient anymore. A substantial transaction tax may result in the realization of scenario as in Figure 2, in which HFTs are completely pushed out of the market and one would think that cost-efficient LFT liquidity provision results. However, as AT participation costs are artificially increased, this is not allocatively efficient given the technology available.

Alternatively, a scenario as in Figure 4 could be replaced by a scenario as in Figure 3, implying that instead of freezes and under-investment in speed technology, we now obtain a cost-inefficient adoption of speed technology. Moreover, introducing a transaction tax then even has the perverse effect that HFTs can pocket rents and leads to under-participation. Furthermore, 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

[27]For a detailed overview and assessment of the proposed policy measures to deal with HFTs, we refer to Section 6 of the final project report on the future of computer trading in financial markets performed by the UK Government Office for Science (2012).

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 effective spreads increase. Hence, we prevent market freezes, but potentially at the cost of lower liquidity overall. Even if transaction taxes are uniformly applied the same effects on liquidity and market stability may pertain, as HFTs will suffer relatively more if speed technology is cost-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).

Second, policymakers have suggested to impose *latency restrictions* on HFTs. Depending on the exact form these latency restrictions take, HFTs then become more like smart ATs in our model; effectively γ decreases while C_A stays the same. By Lemma 3 this is equivalent to a cost increase for HFTs. Again, we could move from a scenario as in Figure 4 to a scenario as in Figure 3 or even 2 and the same effects resort as in the transaction taxes case.

Third, we can have a look at *affirmative liquidity provision*. In its strictest sense, this means that a liquidity provider is forced to provide liquidity at competitive prices at all times the venue is open and regardless of market conditions.[28] Several leading European markets (including Germany, France, Italy, The Netherlands, Sweden and Norway) already attribute comparable designated market maker obligations to a subset of liquidity-providing LFTs.[29] In the US, contracts of this type are currently still prohibited by FINRA rule 5250. If successfully implemented on an individual basis (i.e. each HFT should always provide liquidity), we may move from a scenario as in Figure 4 or 3 to a scenario as in Figure 2. This is clearly an allocative efficiency improvement. Yet, the enforcement costs of such implementation may be high. As an alternative, regulators could punish all active HFTs upon a freeze materializing. This effectively gives a regulator control over C_F in our model. While much cheaper to implement, this implementation would only have the potential to move from a scenario as in Figure 4 to a scenario as in Figure 3.[30]

Finally, 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

[28]The current MiFID II draft mentions such a proposal in Article 17.

[29]Designated market makers receive periodic payments from the firm in exchange for their services. The set maximum spread obligations are typically binding as documented in Anand, Tanggaard, and Weaver (2009). See e.g. Bessembinder, Hao and Zheng (2015) for a thorough theoretical analysis of designated market maker obligations.

[30]Note that another 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.

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"¨This system would impose a take fee in normal times. The revenues can then be used to finance a make fee in times of market freezes in order to incentivize liquidity providers to quote in an empty book and resolve the freeze. This way, liquidity in normal times is reduced similar to the case with a transaction tax, but freezes are avoided (we may end up in the case that HFTs always post in an empty book irrespective of the signal).

7 Extensions and Robustness

In this section, we address robustness issues and present possible extensions to our model. The first extension relates to the fact that market participants in practice can have dual roles. We argue that an endogenous choice between market and limit orders could incentivize LFTs to leave the market even more easily. Next, we explore how the model results would alter when allowing for patient informed liquidity demanders. Thereafter we discuss how in particular our definition of 'efficient' changes when freeze costs are low. Finally, in Subsection 7.4, we generalize the model towards a dynamic setting. This dynamic model allows the generation of signals s to be endogenously derived rather than exogenously assumed, while preserving the results from earlier sections.

7.1 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. Freezes can arise in our setting because LFTs start to exit the market.

In reality, LFTs may have an exogenous trading need, for example due to fund outflows. Such LFTs trade off execution uncertainty with transaction costs. The transaction cost component in this trade-off consists of expected revenues of providing liquidity as well as the expected opportunity costs of using market orders. Our model only covers the effect of expected revenues of providing liquidity. With the entry of more cost-efficient HFTs, both expected revenues as well as opportunity costs decline, shifting the trade-off more and more towards market orders. Therefore, the dual role of LFTs would only make it more likely for LFTs to stop providing liquidity. As a result, equilibria with AT rents or with AT dominance and freezes will materialize more often.

7.2 Patient Informed Liquidity Demanders

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 the 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.[31] 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 due to adverse selection concerns. Yet, posting the first quote of the sequence would be less risky.

7.3 Information Gains Exceed Freeze Costs

So far, we assumed that expected freeze costs always at least offset any informational gain. As a result, liquidity in freeze equilibria would always fall short of first best. If informational gains were to exceed freeze costs, participation could be even higher than in the speed-only case. As a result, the equilibrium with strictly positive probability of Freezes (Equilibrium as in Figure 4) would become the efficient scenario (as avoiding adverse selection effectively lowers the cost of liquidity provision). The probability of such equilibria materializing would also increase as $h_A(\cdot)$ would be higher than before (due to lower C_F).

7.4 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 setting, then information production can be endogenized at the cost of additional complexity. 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:

$$\begin{bmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{bmatrix}, \tag{30}$$

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.[32]

Next, we assume that informed order flow is less patient than uninformed order flow

[31] 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.

[32] 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)).

(i.e., $\nu_{inf} > \nu_{liq}$) and that informationally advanced traders can perfectly observe the historical evolution of the order book. In this dynamic setting, the difference in patience between informed and uninformed liquidity demanders allows for inference about trading types in previous periods by informationally advanced traders. 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 we assume $\nu_{inf} = \infty$ as in our main model, informationally advanced traders can perfectly infer the state of nature of the previous stage game. In that case, we obtain a perfect Bayesian equilibrium, and the signal accuracy parameters are given by:

$$\phi_1 = \alpha, \ \phi_2 = \beta. \tag{31}$$

The dynamic version of the model is essentially a repeated version of the baseline model. As a result, the results from earlier sections carry over. The dynamic model has two important features. First, it shows that information signals can be endogenously generated in the model. Second, because the signals are a result of superior technology for processing public information, signals are common across ATs. This validates the assumption in the baseline model of identical signals for all ATs.

For the fully dynamic setting some conditions need to be satisfied for LFTs to be unable to learn, and for the learning of informationally advanced traders 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 C.

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 limit order book model in which HFTs compete for incoming order flow with low-frequency traders (LFTs), such as traditional market makers or institutional investors. In line with practice, HFTs are modeled to be superior over LFTs in two dimensions. 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).

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 crowded out of the maket for liquidity provision. As such, our model is the first to formally capture and analyze this potential systemic risk HFT activity brings to financial markets.[33] Temporary market freezes could arise with increasing frequency in equilibrium as HFTs gain a larger market share and get access to more cost-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, affirmative liquidity provisions and make-take fees).

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. This is also in line with recent empirical evidence presented in Biais, Declerck and Moinas (2014) and Korajczyk and Murphy (2015). 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 (see Kirilenko et al. (2011) and Biais and Foucault (2014)). Moreover, they have no affirmative obligation to make markets over time and tend to retract in bad times as evidenced in the flash crash (CFTC-SEC (2010)).[34] Furthermore,

[33]See e.g., Biais and Woolley (2011) and Biais and Foucault (2014) for intuitive assessments of the risk of crowding out LFTs.

[34]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.

More generally, Korajczyk and Murphy (2015) provide empirical evidence that in stressful times HFTs reduce liquidity provision significantly, while liquidity provision by designated market makers (i.e., LFTs) remains mostly unchanged.

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 dealing with the aging of society) and in the financing of the economy.[35]

[35]See also Biais and Woolley (2011) and Biais and Foucault (2014) for a further outlook.

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, 2014, High Frequency Traders: Taking Advantage of Speed, NBER Working Paper.

Anand, A., Tanggaard, C. and D. Weaver, 2009, Paying for Market Quality, Journal of Financial and Quantitative Analysis 44, pp. 1427-1457.

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

Baldauf, M. and J. Mollner, 2015, High-Frequency Trading and Market Performance, Working paper.

Baron, M., Brogaard, J. and A. Kirilenko, 2014, The Trading Profits of High-Frequency Traders, Working Paper.

Baruch, S. and L. Glosten, 2013, Fleeting Orders, Working Paper.

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

Bessembinder, H., Hao, J. and K. Zheng, Market Making Contracts, Firm Value, and the IPO Decision, forthcoming *Journal of Finance*.

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

Biais, B., Declerck, F. and S. Moinas, 2014, Fast Trading and Prop Trading, Working Paper.

Biais, B. and T. Foucault, 2014, High-Frequency Trading and Market Quality, Bankers, Markets, and Investors 128, pp. 5-19.

Biais, B., Foucault,T. and S. Moinas, 2015, Equilibrium Fast Trading, forthcoming *Journal of Financial Economics*.

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, pp. 1655–1689.

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

Biais, B., Martimort, D. and J.-C. Rochet, 2000, Competing Mechanisms in a Common Value Environment, Econometrica 78, pp. 799–837.

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

Bongaerts, D., Kong, L. and M. Van Achter, 2015, Trading Speed Competition: Can the Arms Race Go Too Far?, Working Paper.

Breckenfelder, H.-J., 2013, Competition between High-Frequency Traders, and Market Quality, Working Paper.

Brogaard, J., 2011a, High Frequency Trading, Information, and Profits, UK Government Foresight Driver Review 10.

Brogaard, J., 2011b, The Activity of High Frequency Traders, Working Paper.

Brogaard, J. and C. Garriott, High-Frequency Trading Competition, Working Paper.

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

Brogaard, J., Hagströmer, B., Nordén, L. and R. Riordan, 2014, Trading Fast and Slow: Colocation and Liquidity, Working Paper.

Brogaard, J., Moyaert, T. and R. Riordan, 2014, High Frequency Trading and Market Stability, Working Paper.

Budish, E., Cramton, P. and J. Shim, 2015, 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).

Colliard, J.-E., 2013, Catching Falling Knives - Speculating on Market Overreaction, ECB Working Paper.

Copeland, T. and D. Galai, 1983, Information Effects on the Bid-Ask Spread, Journal of Finance 38, pp. 1457-1469.

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, 2014, False News, Informational Efficiency, and Price Reversals, 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.

ESMA, 2014, High-Frequency Trading Activity in EU Equity Markets, Economic Report.

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

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, 2015, News Trading and Speed, Journal of Finance, forthcoming..

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

Foucault, T., Kadan, O. and E. Kandel, 2013, Liquidity Cycles and Make/Take Fees in Electronic Markets, Journal of Finance 68, pp. 299-341.

Garriott, C., 2015, A Dynamic Model of Competition Among High-Speed Market Makers, Working Paper.

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.

Hagströmer, B. and L. Nordén, 2013, The Diversity of High-Frequency Traders, 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.

Han, J., Khapko, M. and A. Kyle, 2014, Liquidity with High-Frequency Market Making, Working Paper.

Hasbrouck, J., 2015, High-Frequency Quoting: Short-Term Volatility in Bids and Offers, Working Paper.

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.

Hendershott, T. and R. Riordan, 2013, Algorithmic Trading and the Market for Liquidity, Journal of Financial and Quantitative Analysis 48, pp. 1001–1024.

Hoffmann, P., 2014, A Dynamic Limit Order Market with Fast and Slow Traders, Journal of Financial Economics 113, pp. 156-169.

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, 2015, Middlemen in Limit-Order Markets, Working Paper.

van Kervel, V. and A. Menkveld, 2015, High-Frequency Trading around Large Institutional Orders, 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.

Korajczyk, R. and D. Murphy, 2015, High Frequency Market Making to Large Institutional Trades, Working Paper.

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

Li, W., 2014, High Frequency Trading with Speed Hierarchies, Working Paper.

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

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

Maskin, E., and J. Tirole, 1988, A Theory of Dynamic Oligopoly. II. Price Competition, Kinked Demand Curves 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.

Menkveld, A. and M. Zoican, 2015, Need for Speed? Exchange Latency and Liquidity, 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, 2015, 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.

Roşu, I., 2015, Fast and Slow Informed Trading, Working Paper.

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.

Tong, L., 2015, A Blessing or a Curse? The Impact of High Frequency Trading on Institutional Investors, Working Paper.

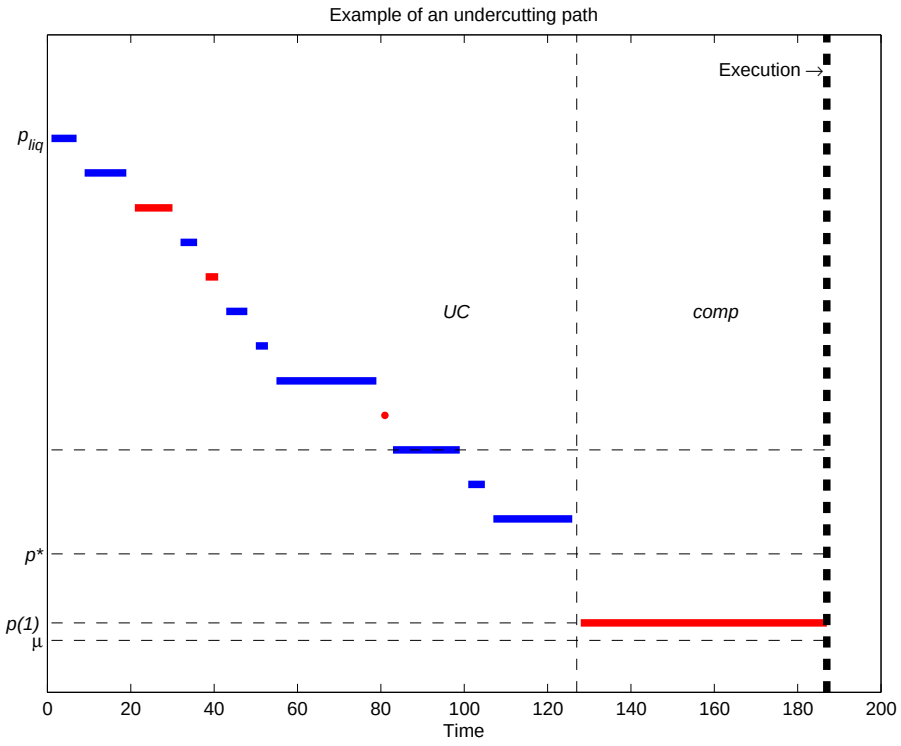
UK Government Office for Science, 2012, The Future of Computer Trading in Financial Markets: an International Perspective, Final Project Report.

Van Achter, M., 2010, A Dynamic Limit Order Market with Diversity in Trading Horizons, 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.

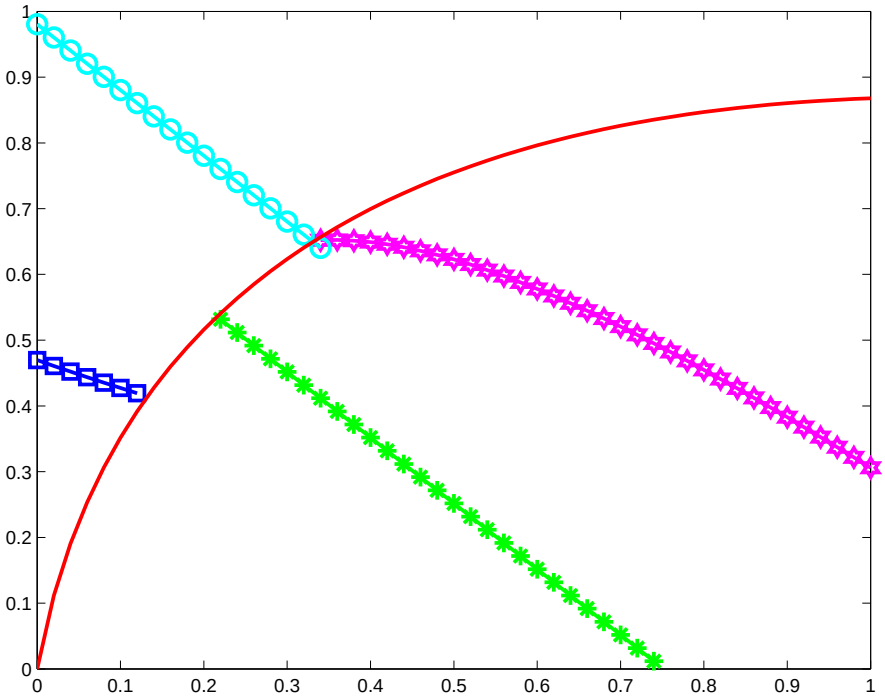
Weller, B., 2013, Fast and Slow Liquidity, Working Paper.

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



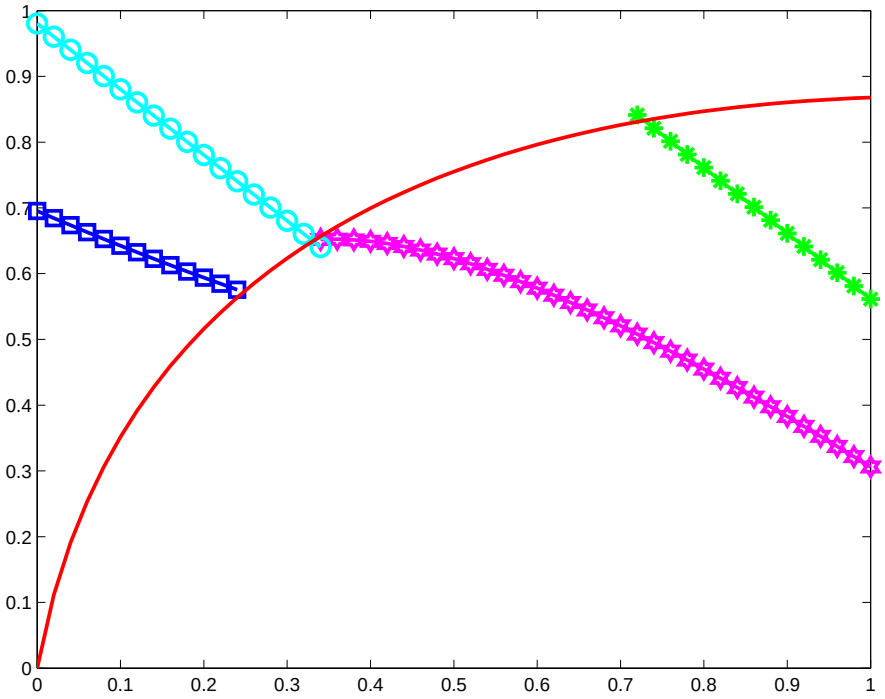
The figure shows an example undercutting path when there is no asymmetric information. The x-axis shows time elapsed since the first quote has been posted, while the y-axis displays price ticks. Blue exposures are HFT exposures while red exposures are LFT exposures.

Figure 2: **To be filled in**



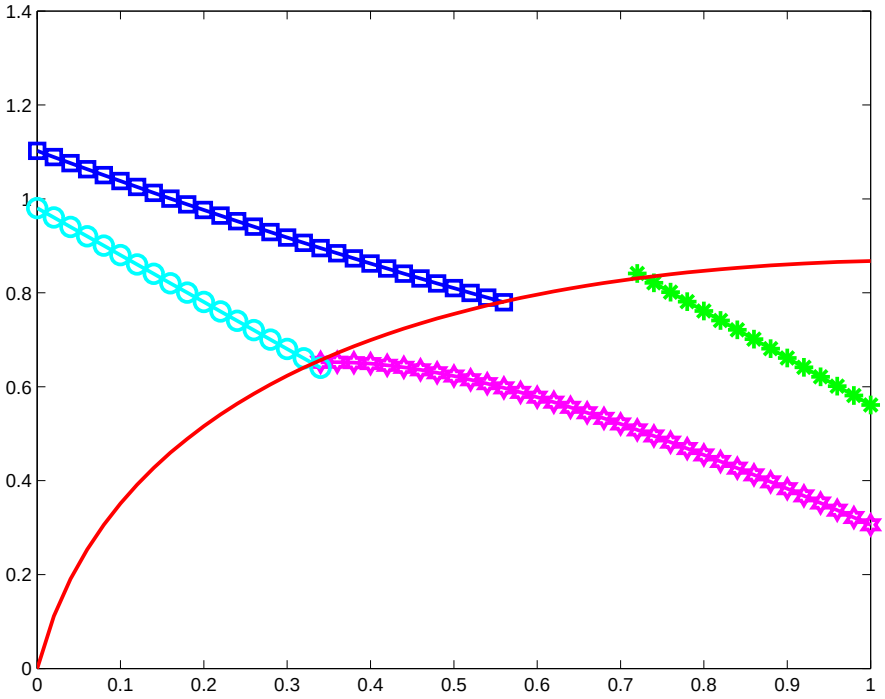
To be filled in.

Figure 3: **To be filled in**



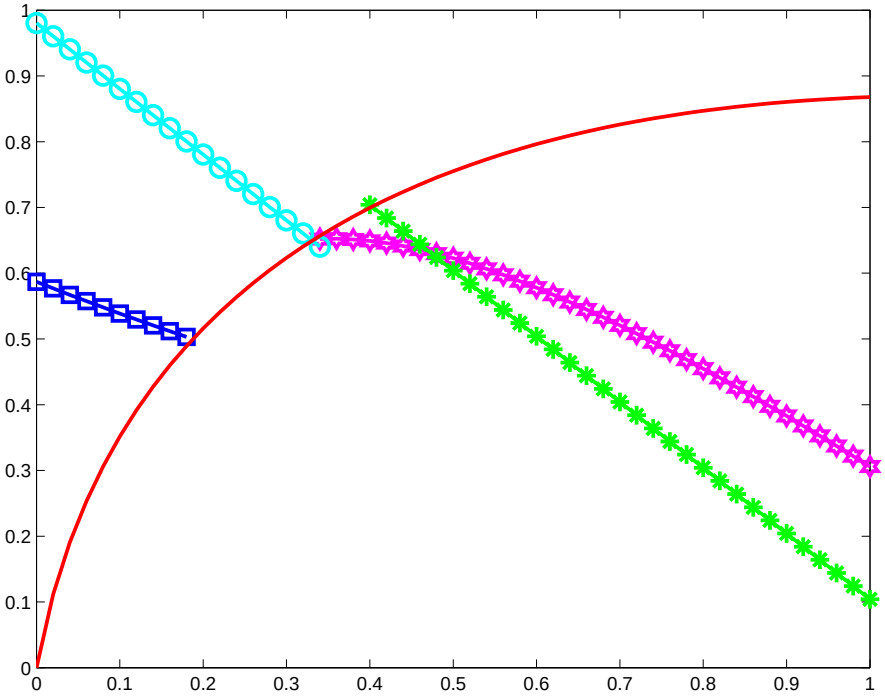
To be filled in.

Figure 4: **To be filled in**



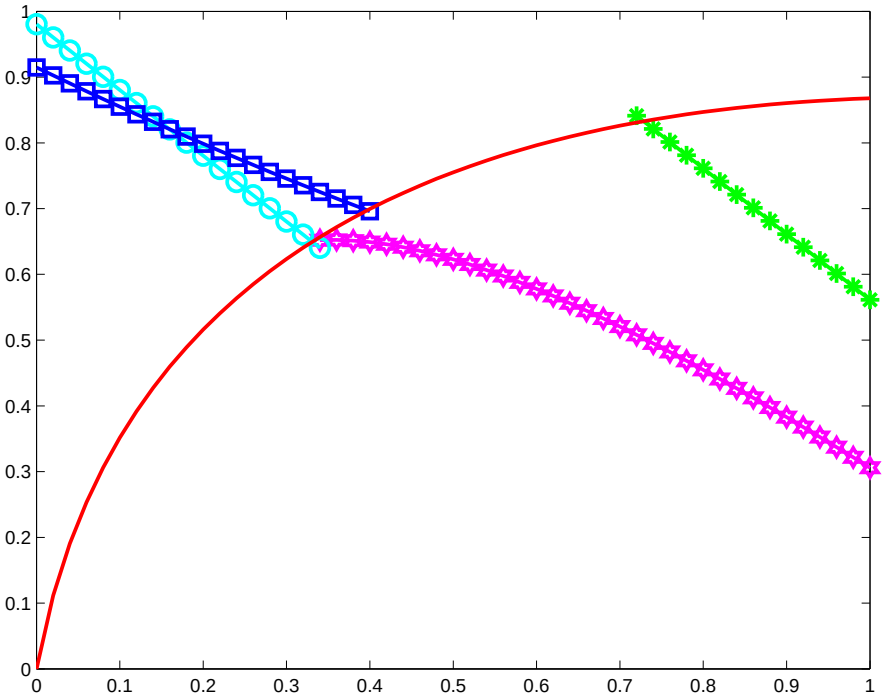
To be filled in.

Figure 5: **To be filled in**



To be filled in.

Figure 6: **To be filled in**



To be filled in.

A Simplification uninformed case

In order to improve tractability, we now introduce a simplified setting. This setting is a specific case of the more general setting. In particular, we impose that $p_{liq} = p(1)+\delta$. As a result, we have only undercutting and no jump of more than one tick to the competitive price. Under this assumption, several expressions simplify significantly. As before, we have that

$$\Phi = \frac{\nu_{liq}}{\nu_{liq} + \lambda(n+b)}. \tag{32}$$

We can now invert the function $g(\cdot)$ in closed form. We have for $g^{-1}(y)$ that

$$y = g(n+b) = \left(\frac{1}{2}\delta + \frac{\delta\nu_{liq}}{\nu_{liq} + \lambda(n+b)}\right)\frac{1}{n+b} \Rightarrow \tag{33}$$

$$(b+n)y = \frac{1}{2}\delta + \frac{\delta\nu_{liq}}{\nu_{liq} + \lambda(n+b)} \tag{34}$$

$$(b+n)y(\nu_{liq} + \lambda(n+b)) = \frac{1}{2}\delta(\nu_{liq} + \lambda(n+b)) + \delta\nu_{liq} \tag{35}$$

$$0 = \lambda y(n+b)^2 + (\nu_{liq}y - \frac{\lambda\delta}{2})(n+b) - 1\frac{1}{2}\nu_{liq}\delta \Rightarrow \tag{36}$$

$$n + b = g^{-1}(y) = \frac{-(\nu_{liq}y - \frac{\lambda\delta}{2}) \pm \sqrt{(\nu_{liq}y - \frac{\lambda\delta}{2})^2 + 6y\lambda\nu_{liq}\delta}}{2\lambda y}. \tag{37}$$

There is only one positive root and hence the inverse of $g(\cdot)$ is one-to-one and monotonic, as basic calculus predicts (since $g(\cdot)$ is continuous and monotonic on its domain). When we set $y = \min(C_L, \frac{C_A}{\gamma})$, this provides a closed form solution for the uninformed case. In this case, the average half spread is given in closed form by

$$p(1) - \mu + \frac{\nu_{liq}\delta}{\nu_{liq} + \lambda g^{-1}(\min(C_L, \frac{C_A}{\gamma}))}. \tag{38}$$

B 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,

$$\tilde{p}^* = \left\langle \mu + \frac{\delta}{2\Phi} \right\rangle^+ = p(1) + \left\lfloor \left\lfloor \frac{1 - \Phi}{2\Phi} \right\rfloor \right\rfloor \delta, \tag{39}$$

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

Q.e.d. ■

Proof of Lemma 2. 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

$$\hat{P}(\zeta = inf|\psi_{LFT}) = P(\zeta = inf|B) = \frac{P(B, \zeta = inf)}{P(B)}, \quad P(B, \zeta = inf) = P(B|\zeta = inf, S)P(\zeta = inf|S) \tag{40}$$

Moreover, we have that

$$P(B|\zeta = inf, S) = P(B|\zeta = liq, S) = \frac{1}{n}, \quad P(B|\zeta = inf, NS) = P(B|\zeta = liq, NS) = \frac{1}{n + \gamma m}, \tag{41}$$

$$P(S) = \bar{\pi}, \qquad P(NS) = 1 - \bar{\pi}, \tag{42}$$

$$P(\zeta = inf|S) = \phi_2, \qquad P(\zeta = liq|S) = 1 - \phi_2 \tag{43}$$

$$P(\zeta = inf|NS) = 1 - \phi_1, \qquad P(\zeta = liq|NS) = \phi_1 \tag{44}$$

Substituting in, we get

$$\hat{P}(\zeta = inf|\psi_{LFT}) = \frac{\frac{1}{n}\phi_2\bar{\pi} + \frac{1}{n+\gamma m}(1-\phi_1)(1-\bar{\pi})}{\frac{1}{n}\phi_2\bar{\pi} + \frac{1}{n+\gamma m}(1-\phi_1)(1-\bar{\pi}) + \frac{1}{n}(1-\phi_2)\bar{\pi} + \frac{1}{n+\gamma m}\phi_1(1-\bar{\pi})} \tag{45}$$

$$= \frac{\frac{1}{n}\phi_2\bar{\pi} + \frac{1}{n+\gamma m}(1-\phi_1)(1-\bar{\pi})}{\frac{1}{n}\bar{\pi} + \frac{1}{n+\gamma m}(1-\bar{\pi})}. \tag{46}$$

The partial derivatives (where $\phi_2 > \bar{\pi}$) are given by:[36]

$$\frac{\partial \hat{P}(\zeta = inf|\psi_{LFT})}{\partial n} = \frac{-\gamma m(1-\bar{\pi})\bar{\pi}(\phi_1 + \phi_2 - 1)}{(n + \gamma m\bar{\pi})^2} < 0 \tag{47}$$

$$\frac{\partial \hat{P}(\zeta = inf|\psi_{LFT})}{\partial m} = \frac{n\gamma(1-\bar{\pi})\bar{\pi}(\phi_1 + \phi_2 - 1)}{(n + \gamma m\bar{\pi})^2} > 0. \tag{48}$$

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 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

[36] Calculations performed by Mathematica

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

$$P(b=i) = \int_{t=0}^{\infty} P(q_t = i) P(t_b > t) \nu_{liq} dt. \tag{49}$$

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 (49), we get

$$P(b=i) = \int_{t=0}^{\infty} \frac{1}{i!}(\bar{\lambda}t)^i \exp(-\bar{\lambda}t) \exp(-\nu_{liq}t) \lambda_{liq} dt, \tag{50}$$

$$= \int_{t=0}^{\infty} \frac{\nu_{liq}\bar{\lambda}^i}{(\nu_{liq} + \bar{\lambda})^{i+1}} \left[(\nu_{liq} + \bar{\lambda})^{i+1} \frac{1}{i!} t^i \exp(-(\nu_{liq} + \bar{\lambda})t) \right] dt. \tag{51}$$

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

$$P(b=i) = \frac{\nu_{liq}\bar{\lambda}^i}{(\nu_{liq} + \bar{\lambda}^{i+1}}. \tag{52}$$

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

$$P_{UC} = \sum_{i=0}^{Z} P(b=i). \tag{53}$$

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,

$$E(\Pi^{comp}) = (1 - P_{UC})(p(1) - \mu) \tag{54}$$

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

$$f_{LFT} = \frac{n}{n + \gamma m}. \tag{55}$$

Q.e.d. ■

Proof of Proposition 3. Define $b = \frac{m}{\gamma}$ and substitute into (16) to (21). Applying the chain rule for differentiation to get the derivatives of the expected revenue densities (16) and (17) with respect to b and n respectively gives:

$$\frac{\partial E\left(\sum_{\widehat{a}} \Pi_{LFT}(R^*_{LFT}(\widehat{a}))\right)}{\partial n} = \frac{\partial E\left(\sum_{\widehat{a}} \Pi_{AT}(R^*_{LFT}(\widehat{a}))\right)}{\partial b} = \frac{-(E(\Pi^{UC} + \Pi^{comp}))}{(n+b)^2} + \frac{\partial(E(\Pi^{UC} + \Pi^{comp}))}{\partial(n+b)} \frac{1}{n+b}. \quad (56)$$

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 (16) and (17) 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. As no party has information, the all players act as in Proposition 3, but with adjusted cost functions. Moreover, (22) and (23) are 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 $C_L < \frac{C_A}{\gamma}$. Combining Proposition 3 with (22) and (23), we have that $m > 0, n = 0$ if $\tilde{C}_L > \frac{\tilde{C}_A}{\gamma}$ and $m = 0, n > 0$ iff $\tilde{C}_L < \frac{\tilde{C}_A}{\gamma}$.

Q.e.d. ■

Proof of Proposition 5. Let us assume that (14) 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 4, 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. (25) then ensures that (14) is satisfied in equilibrium.

Q.e.d. ■

C 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.[37] 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*

$$\beta(1-\pi)P(A_l|\zeta_l = inf, \zeta_{l-1} = inf) + (1-\alpha)(1-\bar{\pi})(\frac{1}{\bar{\pi}} - 1)P(A_l|\zeta_l = inf, \zeta_{l-1} = liq) = (1-\beta)\bar{\pi} + \alpha(1-\bar{\pi})P(A_l|\zeta_{l-1} = liq, \zeta_l = liq) \quad (57)$$

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

$$P(A_l|\zeta_{l-1} = inf, \zeta_l = liq) = 1. \quad (58)$$

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

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

$$P(\zeta_l = inf|A_l) = P(\zeta_l = inf) \ \rightarrow \tag{59}$$

$$\frac{P(A_l|\zeta_l = inf)P(\zeta_l = inf)}{P(A_l)} = P(\zeta_l = inf) \ \rightarrow \tag{60}$$

$$P(A_l|\zeta_l = inf) = P(A_l). \tag{61}$$

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:

$$P(A_l|\zeta_l = inf) = P(A_l|\zeta_l = inf, \zeta_{l-1} = inf)P(\zeta_{l-1} = inf|\zeta_l = inf) + P(A_l|\zeta_l = inf, \zeta_{l-1} = liq)P(\zeta_{l-1} = liq|\zeta_l = inf). \tag{62}$$

Applying Bayes rule twice, we have

$$P(\zeta_{l-1} = inf|\zeta_l = inf) = \frac{P(\zeta_l = inf|\zeta_{l-1} = inf)P(\zeta_{l-1} = inf)}{P(\zeta_l = inf)} = \frac{\beta\bar{\pi}}{\bar{\pi}} = \beta, \tag{63}$$

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

$$P(\zeta_{l-1} = liq|\zeta_l = inf) = \frac{(1-\alpha)(1-\bar{\pi})}{\bar{\pi}}. \tag{64}$$

Substituting these expressions into (62), we get

$$P(A_l|\zeta_l = inf) = P(A_l|\zeta_l = inf, \zeta_{l-1} = inf)\beta + P(A_l|\zeta_l = inf, \zeta_{l-1} = liq)(1-\alpha)(\frac{1}{\bar{\pi}} - 1). \tag{65}$$

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

$$\begin{aligned} P(A_l) = {} & P(A_l|\zeta_{l-1} = inf, \zeta_l = inf)P(\zeta_{l-1} = inf, \zeta_l = inf) + \\ & P(A_l|\zeta_{l-1} = inf, \zeta_l = liq)P(\zeta_{l-1} = inf, \zeta_l = liq) + \\ & P(A_l|\zeta_{l-1} = liq, \zeta_l = inf)P(\zeta_{l-1} = liq, \zeta_l = inf) + \\ & P(A_l|\zeta_{l-1} = liq, \zeta_l = liq)P(\zeta_{l-1} = liq, \zeta_l = liq). \end{aligned} \tag{66}$$

Working out basic statistical identities, we have

$$P(\zeta_{l-1} = inf, \zeta_l = inf) = P(\zeta_l = inf|\zeta_{l-1} = inf)P(\zeta_{l-1} = inf) = \beta\bar{\pi}, \tag{67}$$

and similarly

$$P(\zeta_{l-1} = inf, \zeta_l = liq) = (1-\beta)\bar{\pi}, \tag{68}$$
$$P(\zeta_{l-1} = liq, \zeta_l = inf) = (1-\alpha)(1-\bar{\pi}), \tag{69}$$
$$P(\zeta_{l-1} = liq, \zeta_l = liq) = \alpha(1-\bar{\pi}). \tag{70}$$

Substituting everything into (61) 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:

$$\beta(1-\pi)P(A_l|\zeta_l = inf, \zeta_{l-1} = inf) + (1-\alpha)(1-\bar{\pi})(\frac{1}{\bar{\pi}} - 1)P(A_l|\zeta_l = inf, \zeta_{l-1} = liq) = (1-\beta)\bar{\pi} + \alpha(1-\bar{\pi})P(A_l|\zeta_{l-1} = liq, \zeta_l = liq) \tag{71}$$

and

$$P(A_l|\zeta_l = inf, \zeta_{l-1} = inf) \in [0,1], \tag{72}$$
$$P(A_l|\zeta_l = inf, \zeta_{l-1} = liq) \in [0,1] \tag{73}$$
$$P(A_l|\zeta_{l-1} = liq, \zeta_l = liq) \in [0,1]. \tag{74}$$

D 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_F	$[0, \infty]$	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