Eibelshäuser, Steffen; Smetak, Fabian

Working Paper

Frequent batch auctions and informed trading

SAFE Working Paper, No. 344

Provided in Cooperation with:
Leibniz Institute for Financial Research SAFE

Suggested Citation: Eibelshäuser, Steffen; Smetak, Fabian (2022) : Frequent batch auctions and informed trading, SAFE Working Paper, No. 344, Leibniz Institute for Financial Research SAFE, Frankfurt a. M.

This Version is available at:
https://hdl.handle.net/10419/251785

Steffen Eibelshäuser | Fabian Smetak

Frequent Batch Auctions and Informed Trading

SAFE Working Paper No. 344 | February 2022

Leibniz Institute for Financial Research SAFE
Sustainable Architecture for Finance in Europe
info@safe-frankfurt.de | www.safe-frankfurt.de

Frequent Batch Auctions and Informed Trading*

Steffen Eibelshäuser† Fabian Smetak‡

March 11, 2022

Abstract

We study liquidity provision by competitive high-frequency trading firms (HFTs) in a dynamic trading model with private information. Liquidity providers face adverse selection risk from trading with privately informed investors and from trading with other HFTs that engage in latency arbitrage upon public information. The impact of the two different sources of risk depends on the details of the market design. We determine equilibrium transaction costs in continuous limit order book (CLOB) markets and under frequent batch auctions (FBA). In the absence of informed trading, FBA dominates CLOB just as in Budish et al. (2015). Surprisingly, this result does no longer hold with privately informed investors. We show that FBA allows liquidity providers to charge markups and earn profits – even under risk neutrality and perfect competition. A slight variation of the FBA design removes the inefficiency by allowing traders to submit orders conditional on auction excess demand.

Keywords: market design, market microstructure, liquidity provision, high-frequency trading, continuous limit order book, frequent batch auctions, sniping, latency arbitrage

JEL Classification: G10, D47

*We thank Matthias Blonski, Peter Gomber, Andreas Hackethal, Jan Pieter Krahnen and Loriana Pelizzon as well as seminar participants at SAFE and E-Finance Lab in Frankfurt for helpful comments and discussions.

†Goethe University Frankfurt, eibelshaeuser@econ.uni-frankfurt.de

‡Goethe University Frankfurt, Leibniz Institute for Financial Research SAFE, smetak@safe-frankfurt.de

Contents

1 Introduction and Related Literature

In modern financial markets, highly correlated securities are traded on multiple financial exchanges. This fragmentation has been fostered by regulation both in the US (Reg NMS, 2005) and the EU (MiFID, 2004; MiFID II & MiFIR, 2014) in order to stimulate competition among trading venues. Prices remain aligned across exchanges by two mechanisms: Either by liquidity providers updating their limit prices or by high-frequency trading firms (HFTs) performing *latency arbitrage*. Latency arbitrage can be referred to as the practice of exploiting a time disparity in order execution (e.g. due to differences in trading technologies). When some market participants have speed advantages over liquidity providers, they can profitably trade on mispricings in prevailing bid-ask quotes, thereby bringing back prices to their "fundamental value". Latency arbitrage, however, imposes an additional adverse selection risk on liquidity providers, as shown by Budish et al. (2015) and Foucault et al. (2003, 2013, 2017). When prices on some exchanges respond to information more quickly, prices on other exchanges become stale. While liquidity providers try to adjust their prices, other HFTs try to "snipe" the old quotes before the adjustment, thereby earning arbitrage rents. In equilibrium, liquidity providers take this adverse selection risk into account and quote wider spreads, leading to increased transaction costs for investors. The risks of liquidity provision – and thus bid-ask spreads – are highly sensitive to the market microstructure, i.e. to the rules by which traders interact on the market.[1]

The currently predominant design of financial markets is the continuous limit order book (CLOB) setup, also known as continuous double auction. Under CLOB, financial exchanges process orders serially, in the order of receipt. Therefore, only the fastest traders get to transact or update their quotes, inducing a potentially wasteful arms race for trading speed (Budish et al., 2015). Both economists and practitioners have proposed alternative market designs intended to prevent latency arbitrage. One such alternative is the frequent batch auction (FBA) market design, raised by policy makers (Farmer and Skouras, 2012) and academics (Budish et al., 2015). The idea of FBA is to depart from continuous trading and to introduce periodic market clearing with uniform price double auctions. Budish et al. (2015) propose a trading model populated by noise traders and trading firms, the latter sorting into liquidity providers and stale-quote snipers. The authors show that FBA can mitigate arbitrage

[1] Readers interested in a broader literature review on market microstructure may consult O'Hara (1995, 2015), Madhavan (2000), Harris (2003), or Biais, Glosten, and Spatt (2005).

considerably. Once all firms have access to the same speed technology, stale-quote sniping is eliminated entirely and competitive liquidity providers quote efficient prices at zero bid-ask spreads.

Evidence for the positive welfare effects of batch auctions in financial markets is mounting. Wah and Wellman (2016) and Wah et al. (2016) present simulations of agent-based trading models where FBA leads to lower transaction costs than CLOB. Menkveld and Zoican (2017) argue that slower trading leads to fewer trades among HFTs which decreases adverse selection and thus spreads. Baldauf and Mollner (2020) point out an important trade-off between liquidity provision and price discovery; they find that FBA implements outcomes on the "efficient frontier" of that trade-off while CLOB does not. Aldrich and López Vargas (2020) conduct a series of laboratory experiments confirming that FBA leads to lower transaction costs than CLOB. On the empirical side, Riccò and Wang (2020) analyze the transition from FBA to CLOB on the Taiwan Stock Exchange in March 2020 and find that FBA had lower spreads (and lower trading volume). Ibikunle and Zhang (2021) use data on UK-listed stocks and find that an increase in latency arbitrage is associated with a rise in FBA volume; the authors interpret this as sub-second periodic auctions providing a "safe haven" for slower traders. Besides lower spreads, auctions might have additional advantages. There is an older literature on not-so-frequent batch auctions showing that these can improve market quality, particularly information aggregation and price efficiency (Madhavan, 1992; Economides and Schwartz, 1995; Kandel et al., 2012; Pagano and Schwartz, 2003; Pagano et al., 2013). In a more recent paper, Jagannathan (2020) shows that batch auctions have the potential to dampen shocks and crashes.

Despite their potential advantages, frequent batch auctions do not seem to have caught on.[2] Budish et al. (2020) argue that financial exchanges themselves profit from the speed race by selling speed technology such as fast data feeds and co-location services and therefore have little incentive to change their market design. Other academics point out that FBA might come with some disadvantages as well. Du and Zhu (2017) present a periodic auction model with volatile private information, inventory costs and fully strategic traders. In equilibrium traders engage in "demand reduction",

[2]According to Securities and Authority (2019), periodic auctions in European trading venues only account for around 1.5% of total trading volume as of January 2019. European venues with periodic auction designs include Cboe Periodic Auctions (accounting for roughly 70% of periodic auction volume in Europe), ITG POSIT Auction, Nasdaq Auction on Demand, Turquoise Uncross, Turquoise Lit Auctions, UBS MTF, GS Sigma-X MTF and MS MTF.

leading to an important trade-off: Fast trading (many auctions per time) allows quick asset reallocation due to news, but quoting in each auction is inefficient. The authors estimate the optimal trading frequency to be in the range of a couple of auctions per second. Fricke and Gerig (2018) show that risk-averse traders suffer from increased execution uncertainty in FBA. Using U.S. data they estimate the optimal batch frequency to be about half a second. Similarly, Bellia et al. (2020) argue that the lack of immediate execution in FBA might reduce participation of HFTs which in turn may impede the quality of price discovery. Finally, Haas et al. (2020) illustrate yet another trade-off linked to FBA. In their model, batch auctions decrease the number of sniping opportunities, thus improving liquidity, but less frequent trading also increases the likelihood of more informed traders per auction, thus hampering liquidity. The practitioner side also appears to be skeptical of FBA. For example, Dorre (2020) harshly criticizes FBA in a series of blog posts, albeit without exact scientific reasoning. Jagannathan (2020) conjectures that there are some "aspect[s] of reality academic research may be missing".

The main objective of this paper is to compare the CLOB and FBA market designs on the basis of the bid-ask spreads and markup inefficiencies that they imply. While existing theoretical literature has mainly focused on adverse selection risks from *either* high-frequency arbitrageurs *or* from privately informed investors, we will use a dynamic trading model that allows for the presence of both of these risks simultaneously. As will become evident, this does not constitute a minor variation but can have game-changing effects on the efficiency of the individual market designs. We measure the (in)efficiency of a particular market design by the excessive markups that investors expect to pay per unit of time to liquidity providers – we will refer to this measure as the "expected markup flow". The core argument will be that liquidity provision under FBA is inefficient and that the severity of this inefficiency can exceed the one from latency arbitrage under CLOB. To arrive at this conclusion, we follow the seminal work of Budish et al. (2015) and use a variation of their original model with one main adjustment: We allow some traders to have private information about the fundamental value of the asset. When trading can be motivated by private information, trades have price impact and the order flow determines the expected fundamental value. In a sealed-bid batch auction, however, the order flow is unknown ex ante and so is the true fundamental value during the auction. We will show that fully strategic bidders take this into account and that quoting becomes inefficient: In equilibrium, market

makers can charge additional markups and make strictly positive expected profits – even under risk neutrality and perfect price competition. These markups translate into increased transaction costs and hence into welfare losses for investors. The inefficiency resembles bid shading (also known as demand reduction), a well-known phenomenon in multi-unit, uniform price double auctions, and the following result applies: *Every equilibrium in multi-unit, uniform price auctions is inefficient due to bid shading* (Ausubel et al., 2014, Theorem 1, p. 1380). From that perspective, our paper is similar to Jovanovic and Menkveld (2021) who also consider an auction setup with an unknown number of bidders. However, the authors focus on randomization in mixed-strategy equilibria, similar to the classic model of sales (Varian, 1980), while we focus on inefficiencies of *pure-strategy equilibria*. In our model, there are two key determinants of trading costs: Firstly, the relative frequencies of privately informed trading versus publicly observable news events and, secondly, the absolute frequency of trading relative to the length ot the batch interval. When overall trading activity is high and privately informed trading is relatively frequent, expected markup flow in equilibrium is lower in CLOB than under FBA, and vice versa.

We provide three sets of results: Firstly, we prove existence and uniqueness of pure-strategy equilibria in CLOB and under FBA (uniqueness of equilibrium quotes up to permutation of trading firms). We derive closed-form representations of equilibrium bid-ask spreads and markups of liquidity providers under each market design. Secondly, we highlight the distinct inefficiencies of FBA and CLOB and show that the former can strictly exceed the latter. More precisely, we derive a boundary in the parameter space between FBA and CLOB such that each market design is welfare-dominant (in terms of expected markup flow) on one side of the boundary. Thirdly, we show that a slight variation of the FBA design – in which traders may submit quotes conditional on excess demand within an auction – is welfare-optimal in the sense that it leads to zero markups and lowest possible transaction costs for investors. Overall, we intend to contribute to the literature by formally shedding light on yet undiscovered inefficiencies and by offering a potential solution. We believe that these results are not only appealing from a theoretical perspective but can also provide additional insights for policy makers and market design authorities.

The remainder of this paper is organized as follows. Section 2 illustrates some of the main results with simple examples. Section 3 introduces the trading model and

discusses its key assumptions. Section 4 presents the equilibrium concepts used for the different market designs and derives closed-form solutions for equilibrium inefficiencies under CLOB and FBA. Section 5 compares equilibrium trading costs and inefficiencies and establishes conditions for CLOB to welfare-dominate FBA. Section 6 presents a slight modification of FBA with conditional quoting schedules such that full efficiency is restored. Finally, Section 7 concludes. We provide proof ideas of important results throughout the text while detailed proofs are deferred to Appendix A. The Online Appendix contains supplementary materials and discussions.

2 Introductory Examples

Besides direct costs such as custody fees or broker commissions, bid-ask spreads constitute the lion's share of transaction costs. Should market microstructures have a significant bearing on those costs, clever market design can help to mitigate these inefficiencies. Albeit bid-ask spreads might appear negligible, even small decreases can translate into considerable welfare improvements given the enormous volume transacted on financial exchanges across the world.[3] We focus in particular on CLOB and FBA in this regard. The two distinct characteristics of the CLOB market design are *continuous trading* and *serial order processing*: Traders can send messages (i.e. post, cancel or adjust orders) at any time and the financial exchange processes incoming orders serially, one after the other, in order of receipt. Therefore, CLOB rewards speed advantages as only the fastest traders can transact or update their quotes. In the FBA design, by contrast, the trading day is divided into discrete intervals and uniform price, sealed-bid double auctions are conducted frequently over the course of the trading day. During an auction interval, traders can submit or modify orders at any time and the financial exchange collects – but does not immediately process – all incoming messages. At the end of the auction interval, the exchange *batches* all outstanding orders and treats them *as if* having arrived at the same time. It then computes aggregate demand and supply schedules and a uniform market clearing price at which each transaction is executed.[4] After market clearing quantities and the price

[3] It is worth to mention that transaction costs in general – and bid-ask spreads in particular – have significantly declined throughout the last decades (Jones, 2002, 2013). Nevertheless, this development does *not* imply that today's bid-ask spreads are optimal. We do not neglect previous improvements, but rather ask how transaction costs can be decreased further.

[4] Supply and demand schedules can intersect in different ways (horizontally, vertically) which leads to different equilibrium prices. More detailed discussions can be found in Budish et al. (2015).

have been computed and transactions have been executed, the order history of the current batch interval is made public by the exchange. The following simple examples summarize the (far more general) results of this paper in a nutshell. They shed some light on how – and why – the respective market designs might be vulnerable to adverse selection risks from informed investors and HFTs.

2.1 CLOB and FBA with HFTs

Suppose there is a single asset with current fundamental value $v \equiv 0$. The asset value may change over time due to publicly observable news events. Assume that a positive event increases v by $J = 1$. There are two kinds of market participants: Infinitely many investors and N high-frequency trading firms (HFTs), one of which acts as a market maker (MM) who quotes bid and ask prices at which she is willing to buy and sell the asset. Investors arrive stochastically at the exchange and have exogenous motivations to trade. HFTs and the MM do not have intrinsic trading motives. They observe – and can act upon – news events simultaneously and posses equally fast trading technologies.

Consider what happens in the CLOB market design once a positive news event occurs and the fundamental value jumps to $v = 1$. All trading firms observe the event simultaneously and instantaneously send their orders to the exchange. The MM's old quotes have become stale so she wants to adjust them upwards by $J = 1$. The other $N - 1$ HFTs try to snipe the old quotes by submitting market buy orders. Since all firms are equally fast, their orders arrive at the exchange simultaneously and the MM's old quotes are sniped with probability $\frac{N-1}{N}$. To account for the resulting losses, the MM quotes a positive bid-ask spread and earns profits from trading with investors. Equilibrium quotes are such that all firms make equal expected profits and no one has incentives to deviate. Hence, even in the absence of private information or technological advantages, CLOB allows for arbitrage. These rents come at the expense of investors (e.g. mutual funds, retail investors, etc.) who are faced with higher trading costs in the form of wider bid-ask spreads.[5]

What would the same situation look like in the FBA market design? Just as under

[5] Budish et al. (2015) highlight another inefficiency: A socially wasteful arms race for speed among HFTs which eventually also harms investors in the form of higher trading costs.

CLOB, trading firms could instantaneously send their orders to the exchange upon a positive news event (HFTs: market buy orders, MM: quote adjustment). In contrast to CLOB, the exchange *batches* all outstanding orders at the end of an auction and does not process them serially. Therefore, the canceled quotes of the MM do not even enter the supply schedule when the market clearing price is computed and only the adjusted quotes are considered. FBA thereby eliminates sniping risk entirely and the MM quotes zero bid-ask spreads, implying zero costs for investors.[6]

2.2 CLOB and FBA with HFTs and Informed Investors

Now suppose that some traders may have private information about changes in the fundamental value of the asset. In this case, marketable orders (i.e. trades) have price impact. They carry informational content and a buy (sell) order can be regarded as a signal that the fundamental value of the asset has increased (decreased). Market makers are systematically disadvantaged when trading with informed investors. They therefore quote positive bid-ask spreads in order to compensate these losses with profits earned from trading with uninformed investors. This informational disadvantage for market makers and the corresponding increase in bid-ask spreads is the same in both market designs. The risk from stale quote sniping, however, is still present only under CLOB and has to be accounted for by quoting higher bid-ask spreads. It seems straightforward that trading costs in the FBA design will again be strictly lower than under CLOB. Any markups - other than accounting for privately informed traders - should be competed away by HFTs who can also act as market makers and undercut profitable quotes. As it turns out, this does *not* have to be the case: Market makers in the FBA design *can* charge additional markups without being undercut. To illustrate this point, suppose that a given trading firm ("the seller") holds two shares of the asset with current fundamental value $v = 0$ and wishes to sell both shares through a uniform price auction. Further suppose that due to the presence of informed traders, buy orders have price impact normalized to $\Delta = 1$. Therefore, if one buy order arrives in the auction interval, the expected fundamental value of the asset is 1. If two buy orders arrive, the expected fundamental value is 2, and so on. To keep the setup simple, suppose that the number of buy orders in the auction is either 1 or 2 with 50% probability each. Under FBA, the trading price is determined as the mid-price

[6]Profits earned from quoting positive bid-ask spreads would be immediately competed away by other trading firms who can also act as market makers and would undercut these quotes.

where supply and demand schedules intersect. To make zero expected profit (and for the market to match buyers and sellers efficiently), it seems intuitive for the seller to quote the first share at an ask price of $a_1 = 1$ and the second share at $a_2 = 2$. In case of one (two) arriving buy order(s), the resulting trading price would be $p = 1$ ($p = 2$) with zero expected profit for the seller. But is this really an equilibrium? The answer, somewhat surprisingly, is *no*. The seller can strictly improve by increasing the limit price for the first share to $\hat{a}_1 = 1.5$ – even in case of risk neutrality and *perfect* market maker competition. To see this, suppose that other firms would step in if quoted ask prices are too high and if undercutting them is profitable. How high can ask quotes be set so that still no other firm has an incentive to underbid the seller's quotes? For the second share, there is no room for improvement. For any ask price $\hat{a}_2 > 2$, another trader would undercut the quote by $\varepsilon > 0$, thus rendering $\hat{a}_2$ the third best ask price which would never execute. If two buy orders arrive, the trader can sell at the market clearing price $a_2 - \varepsilon > 2$ and make positive profits. Therefore, in equilibrium, the second ask must be $a_2 = 2$. However, the seller can increase her first quote to $\hat{a}_1 = 1.5$ (thereby make positive expected profits), and still no other trader can undercut her profitably. Suppose some other trader would undercut to $a' = 1.5 - \varepsilon$. If one buy order arrives, that trader gets to trade at the clearing price $p = a'$ and makes expected profits of

$$\begin{aligned} \mathbb{E}[\pi' | 1 \text{ buy order}] &= p(a', \hat{a}_1, a_2, 1 \text{ buy order}) - \mathbb{E}[v | 1 \text{ buy order}] \\ &= 1.5 - \varepsilon - 1 = 0.5 - \varepsilon. \end{aligned}$$

If two buy orders arrive, however, *the seller's* quote $\hat{a}_1 = 1.5$ is the second best ask and hence determines the uniform trading price $p = \hat{a}_1$. The other trader incurs losses of

$$\begin{aligned} \mathbb{E}[\pi' | 2 \text{ buy orders}] &= p(a', \hat{a}_1, a_2, 2 \text{ buy orders}) - \mathbb{E}[v | 2 \text{ buy orders}] \\ &= 1.5 - 2 = -0.5. \end{aligned}$$

Since each event occurs with 50% probability, the overall expected profit of the other trader is -0.5ε. He would hence decide not to underbid $\hat{a}_1$ in the first place. The resulting equilibrium ask price of $\hat{a}_1 = 1.5$ can be decomposed into $\hat{a}_1 = \Delta + x_1 = 1 + x_1$, where 1 is the expected fundamental value in case one buy order arrives and $x_1 = 0.5$ is the additional markup that the seller can charge in the FBA market design.

If one buy order arrives, the seller's profit is equal to x_1 and if two buy orders arrive, her profit is zero. Overall, this yields strictly positive expected profits of 0.25.

The MM makes positive expected profits in equilibrium and nobody can profitably prevent her from doing so. This induces positive trading costs above what is charged to account for risks from informed investors. The inefficiency arises in uniform price auctions when (i) marketable orders have price impact and (ii) multiple units can be traded in an auction. The phenomenon resembles *bid shading* (or *demand reduction*) which occurs when orders for some units have a potential impact on the price of other units. This externality is priced in and equilibrium prices are excessive. Private information thus has differential implications for CLOB and FBA. Unlike in the first example where private information was absent, it is no longer obvious whether FBA is still the dominant choice in terms of efficient liquidity provision and trading costs for investors.[7]

2.3 Potential Solution for FBA with Informed Investors

The quoting inefficiency in FBA with informed traders arises because quotes for some units can impact the price of other units. This externality disappears if price competition is independent across units. One way to achieve this is to allow bidding via price-quantity schedules, i.e. to allow quoting *conditional on demand in an auction*. This restores independent price competition for each unit, leading to the classic Bertrand outcome with competitive prices for buyers. Continuing the previous example, suppose the seller could submit ask quotes conditional on the number of buy orders arriving as follows:

$$a_{\text{one buy order}} = 1\Delta = 1, \qquad a_{\text{two buy orders}} = 2\Delta = 2.$$

[7]It should be mentioned that there does exist a unique, symmetric *mixed-strategy* equilibrium with zero expected profits for the seller under FBA. An example thereof is provided in the ONLINE APPENDIX. The simple trading model hence indicates that FBA can induce inefficient stable prices or efficient random prices. Either scenario might not be desirable. Since our focus is on strategic behavior of HFTs and market makers as inspired by Budish et al. (2015), we continue to focus on pure-strategy equilibria. Recent research on price dispersion (Jovanovic and Menkveld, 2021) or flickering quotes and fleeting orders (Baruch and Glosten, 2013) makes specific use of stochastic quoting behavior. Mixed-strategy equilibria can hence be particularly fruitful with seemingly random patterns in (high-frequency) data.

In other words, if one buy order arrives, the seller is willing to sell one unit at a price of 1, whereas if two buy orders arrive, the seller is willing to sell two units at a price of 2 each. This is efficient quoting and in fact the only equilibrium outcome – the clearing price will always be equal to the expected value. Contrary to the previous section, the seller can no longer profitably raise any quote. If the seller increased any of her quotes, other firms would step in and undercut the quote – without any downside. Perfect price competition is restored and buyers can transact at the best possible prices.

3 General Setup of the Dynamic Trading Model

This section introduces the general building blocks of the trading model we use. Most modeling assumptions are closely related to those in Budish et al. (2015) in order to allow for an easier comparison of results. The key difference is that we augment the financial market model by privately informed investors so that marketable orders have price impact. Further details on the market microstructure are intentionally left unspecified in this section since equilibrium analyses are conducted separately for CLOB and FBA in the subsequent section.

3.1 Model Primitives and Basic Setup

Time: Time t runs continuously on $[0, T]$, where $T > 0$. We abstract from any latency and assume that submitted orders arrive at the exchange *instantaneously*.

Asset: There is a single risky asset with time-varying fundamental value V_t which evolves according to a Poisson jump process. For $t \in (0, T)$, V_t remains unobservable to all agents. At each jump time, the fundamental value jumps up or down by a constant jump size $J > 0$ with equal probability. Jump directions are stochastically independent.[8] While some of the jumps are observed publicly, others are only observed privately. Further details will be explained below.

Prices: Prices can be any real number, i.e. the tick size is zero.

[8]The specific jump size distribution allows for closed-form solutions, but it should be generalizable to arbitrary, symmetric distributions without affecting the qualitative results of this paper.

Market: The asset is traded in an anonymous limit order market. Attention is confined to limit orders and market orders. All orders must be for one single unit of the asset.[9]

Traders: The model is populated by three types of traders, all of whom are risk-neutral and do not discount future payoffs. They differ in their motivations to trade.

(i) There are infinitely many **informed traders**. According to an exogenous Poisson process with rate $\lambda_i \geq 0$, one single informed trader arrives at the market, privately observes a jump of size $\pm J$, and issues a market buy or market sell order for one unit of the asset accordingly. Upon order execution, the corresponding trader exits the market forever.

(ii) There are infinitely many **uninformed traders**, also referred to as **noise traders**, who trade for exogenous motivations such as rebalancing their portfolios or for liquidity reasons. According to an exogenous Poisson process with rate $\lambda_n > 0$, one single uninformed trader arrives at the market, randomly issues a market buy or a market sell order for one unit of the asset, and exits the model forever. Uninformed traders can be modeled as to have a private valuation component $\tilde{U}$ and to value the asset at $V_t + \tilde{U}$. Private valuation components $\tilde{U}$ are equal to $\pm U$ (where $U \geq J$)[10] with equal probability and are independent and identically distributed across investors. Due to the symmetric distribution of $\tilde{U}$, uninformed investors buy and sell with equal probability.

(iii) There are $N \geq 2$ perfectly competitive **trading firms** who may submit limit and market orders at any time. Trading firms can freely choose to act as either *market makers* (MMs) or *high-frequency traders* (HFTs) and this choice is determined by whatever role yields higher expected profits. All firms are assumed to be infinitely fast, i.e. they can react *instantaneously* to any market event.

Market Events: The model features three types of stochastic market events. Their arrival is governed by three independent Poisson processes. At *public news events*, the

[9] The single-unit restriction keeps the model tractable without departing too far from today's trading reality. For most liquid assets, individual order sizes are fairly small and mostly single-digit. One reason for small order sizes is provided by Kyle (1985) and Back and Baruch (2007) who have shown that large block orders should optimally be split into several smaller orders to avoid price impact and to minimize effective transaction costs.

[10] $U \geq J$ is a technical restriction ensuring that noise traders are willing to transact at bid-ask spreads.

fundamental value jumps by $\pm J$. They occur with Poisson intensity $\lambda_j \geq 0$ and are observed simultaneously by all agents. *Informed trade* occurs with Poisson intensity λ_i and is characterized by privately observable jumps of size $\pm J$ in the fundamental value. They are only observable to one informed investor who submits a market order for one unit of the asset in the corresponding direction. Finally, *noise trade* occurs with Poisson intensity λ_n. A noise trader arriving at the market randomly submits a market buy or sell order for one unit of the asset with equal probability.

Information and Updating: Any trader arriving at the market at time t can observe the past history $\mathcal{H}_t$ of all prices, quotes, trades, and publicly observable jumps in the fundamental value of the asset. $\mathcal{H}_t$ is also referred to as the *public information available at time t*. All parameters of the model are common knowledge. Furthermore, the initial fundamental value v_0 and the order flow are public information and summarized by $\mathcal{H}_t$ at any point in time. Orders are anonymous, however, so informed and uninformed trades cannot be distinguished and agents act in a game of imperfect information. Market participants perform standard Bayesian updating to form an unbiased expectation about the valuation of the asset.

Strategies: Informed and uninformed investors arrive at the market according to an exogenous stochastic process, mechanically trade one unit at the best available price and leave the market afterwards. Their behavior is not strategic. Strategic interaction is limited to competition among trading firms. They are permanently active in the market and may submit, modify and cancel orders at any time. Further details are provided in Section 4 where equilibrium analyses are presented for each market design.

3.2 Discussion of the Model Setup

The model setup is closely related to Budish et al. (2015) in order to simplify a comparison of results. The key distinction of our model is that we allow for traders with private information as in Glosten and Milgrom (1985).[11] The possibility of informed trading causes an adverse selection problem for liquidity providers which is

[11] Augmenting the model of Budish et al. (2015) by privately informed investors has also been done by Budish et al. (2020), however, with different modeling assumptions and focus. While their paper considers stock exchange competition between multiple trading venues, our paper focuses on competitive liquidity provision on a single exchange.

accounted for by increased bid-ask spreads.[12] Furthermore, informed trading causes a rational Bayesian price impact corresponding to the expected informational content of marketable orders. While privately informed trading has a long tradition in the literature of financial markets (Treynor, 1981; Copeland and Galai, 1983; Glosten and Milgrom, 1985; Kyle, 1985) it has been abstracted from in much of the recent literature (with the notable exception of Baldauf and Mollner (2020), for example). Our model provides a synthesis of Budish et al. (2015) and Glosten and Milgrom (1985). Without informed trading ($\lambda_i = 0$), our model reduces to the original Budish-Cramton-Shim model with constant jump size J. Without public news events ($\lambda_j = 0$), our model corresponds to a continuous-time version of the Glosten-Milgrom model. Further discussion is deferred to the Online Appendix.

Since one of our key objectives is to shed some light on yet undiscovered inefficiencies of FBA as compared to CLOB, assumptions that leave room for ambiguity have been made in the most conservative manner, i.e. in a way that is most favorable for the FBA design. This concerns in particular the instantaneous reaction time and the common speed technology. Assuming zero latency and equally fast trading technologies eliminates sniping risk under FBA completely.[13] Budish et al. (2015) allow for different trading technologies (fast and slow) and show that even in this case, frequent batch auctions can alleviate sniping risk considerably.[14] In summary, if we can find a situation where CLOB dominates FBA in terms of trading costs and inefficiencies under the assumptions that we impose, then this will be especially true when relaxing some of these assumptions.

[12] Easley et al. (1996) have shown empirically that the adverse selection risk associated with informed trading is a major source of bid-ask spread.

[13] When all trading firms have access to a common no-latency speed technology, order cancellations and adjustments will always arrive at the same time as sniping attempts within an auction interval. Since all orders are batched at the end of an auction, only the MM's adjusted quotes are relevant and her stale quotes do not enter her supply schedule. This eliminates latency arbitrage.

[14] More precisely, they show that the proportion of an auction interval of length τ which allows for latency arbitrage is $\frac{\delta}{\tau}$ where $\delta := \delta_{\text{slow}} - \delta_{\text{fast}}$ is the difference in latency of the speed technologies. By choosing the length of the auction interval long enough, sniping risk can be reduced considerably. See Budish et al. (2015) Section VII.B and Figure VII for a more detailed exposition.

4 Equilibrium Analysis

This section provides equilibrium analyses for the CLOB and the FBA market design. First, we introduce necessary equilibrium notions. We then present optimal quoting strategies of trading firms and compute the resulting equilibrium bid-ask spreads and the distinct inefficiencies as represented by the expected markup flow.

4.1 Equilibrium Concepts

A natural solution concept for stochastic games in continuous time with imperfect information is a *pure-strategy, stationary Markov Perfect Equilibrium* (MPE) (Maskin and Tirole, 2001). While this equilibrium concept *does* apply to the CLOB design, the discrete-time implementation of sealed-bid, frequent batch auctions together with private information leads to non-existence problems under FBA. In their recent paper, Budish et al. (2020) also augment a similar trading model by private information and encounter analogous equilibrium existence issues. The authors suggest an alternative solution concept, *Order Book Equilibrium* (OBE), which strictly weakens MPE but still tries to capture the spirit of competitive liquidity provision à la Glosten and Milgrom (1985). We will follow the authors and use the OBE notion under FBA.

Stationary Markov Perfect Equilibrium (in CLOB)

As the order book can be monitored continuously under CLOB, trading firms can *condition their strategies* on its current state. Strategies hence constitute a complete contingent plan of action: Whenever some trading firm deviates from equilibrium play (which will be reflected in the order book configuration), other firms can *instantaneously* react and play their best responses against the deviation. If such reactions render the deviation unprofitable, total profits earned from deviating are essentially zero (and the deviation attempt should not be initiated in the first place). Continuous-time trading and reactions that occur *at the same instant* make stationary Markov Perfect Equilibria well-defined under CLOB. An MPE is a subgame perfect equilibrium in which players may only use time-independent Markov strategies, i.e. strategies that depend on the current state only (Fudenberg and Tirole, 1991; Maskin and Tirole, 2001). In the present model, the payoff relevant state at time t is given by $\mathcal{S}_t = \{P_t, J_t, (b_t, a_t)\}$ where $P_t = \mathbb{E}[V_t|\mathcal{H}_t]$ denotes the estimate of the fundamental value given public information, $J_t \in \{0, -J, J\}$ indicates whether there is a jump at

time t, and (b_t, a_t) summarizes the current bid and ask quotes in the order book. The dynamic MPE constitutes a static Nash equilibrium at each point in time.

Order Book Equilibrium (in FBA)

The introductory example with private information in Section 2 showed that a market maker under FBA can charge additional markups and thus make positive profits. Other firms cannot undercut these quotes as they would incur losses. But *given that no one undercuts her quotes*, the market maker has incentives to increase markups even further. Therefore, stationary MPEs do not exist under FBA and a weaker solution concept is required. Note, however, that once the market maker deviates by increasing the markups, other firms will have an incentive to undercut these higher quotes which renders the deviation unprofitable.[15] In order to enhance the comparability of our results to Budish et al. (2015) and Budish et al. (2020), we follow the latter authors and use their alternative solution concept of an *Order book Equilibrium* (OBE) in the FBA setup. We briefly outline the main intuition in the following – a more detailed treatment can be found in the last-mentioned authors' paper or theory appendix and in our Online Appendix. Whereas an MPE requires that no player has a profitable deviation at any point in time, OBE allows for profitable deviations to exist as long as they are rendered unprofitable by specific reactions of some other player.[16] An OBE is a set of orders of the high-frequency trading firms that captures the idea of a "rest point" of the order book (Budish et al., 2020): First, there do not exist *strictly profitable safe price improvements*. These are deviations that improve prices of existing quotes. More specifically, new quotes constitute a price improvement over older ones when it is weakly cheaper trading against liquidity-providing limit orders in the new set of quotes and there exists some limit order for which it is strictly cheaper. A strictly profitable price improvement is *safe* if it remains strictly profitable, even if some other trading firm profitably withdraws liquidity (i.e. takes out limit order(s) from the book) in response to firm i's deviation.[17] Second, there do not exist *robust*

[15] Such reactions can only happen at the end of an auction as this is when supply and demand schedules are released and the deviation becomes visible. Any possible equilibrium under FBA hence cannot capture the idea of instantaneous reactions *within a given auction*.

[16] Profitable deviations can hence only be sustained for short time periods, i.e. an auction interval.

[17] Note that a price improvement can never withdraw liquidity itself as it would otherwise no longer be weakly cheaper to trade against *all* liquidity-providing limit orders in the new set of quotes. Further note that liquidity withdrawals can never render another quote or deviation unprofitable in our FBA setup and any profitable price improvement is hence automatically *safe*: In case of excess demand, for instance, fewer quotes in the order book due to a liquidity withdrawal can only mean

deviations. These are strictly profitable deviations (other than price improvements) of any firm *that remain strictly profitable* even if another firm *profitably reacts* to this deviation by safe profitable price improvements or liquidity withdrawals (Budish et al., 2020). We follow the latter authors and term the following solution concept an *Order Book Equilibrium* (OBE).

Definition 1 (Order Book Equilibrium)**.**
Given the state $\mathcal{S}_t = \{P_t, J_t, (b_t, a_t)\}$ *at time* t*, an* ***Order Book Equilibrium*** *at time* t *of the strategic interaction between the high-frequency trading firms is a set of orders such that the following conditions hold for all firms:*

1. *There do not exist* strictly profitable safe price improvements*.*
2. *There do not exist* strictly profitable robust deviations*.*

OBE captures the idea that likely anticipated reactions by rival trading firms to an existing profitable deviation can discipline a given firm not to pursue the deviation in the first place (Budish et al., 2020). They can hence uphold equilibrium quote levels and no trading firm has an incentive to add or withdraw liquidity-providing quotes from the order book. Firstly, there do not exist any *strictly profitable price improvements*: It cannot be possible in an OBE that some firm can still profitably *improve* quotes. Secondly, there do not exist strictly profitable *robust deviations*: It cannot be possible in an OBE to *worsen* quotes without triggering underbidding incentives of other firms that would render the initial deviation unprofitable.

In market designs that feature continuous order processing (e.g. CLOB), the stronger concept of a MPE coincides with that of an OBE: Trading firms can condition their actions on the order book which is observed *at every instant*. Hence, in a MPE, there do not even exist *any* profitable deviations given the strategies of the other firms. Summing up, we will solve for pure-strategy, stationary Markov Perfect Equilibria under CLOB and for pure-strategy, Order Book Equilibria under FBA.[18]

(weakly) higher trading prices as supply and demand would intersect at (weakly) higher ask quotes. This would – if anything – increase an incumbent market maker's profits.

[18]Reverting to the example from Section 2.2, the suggested equilibrium in which the market maker charges markups under FBA does not constitute an MPE but it fulfills all properties of an OBE.

4.2 The Continuous Limit Order Book Market Design

We derive the equilibrium in the CLOB market design in the following. In other words, we present optimal strategies of trading firms and compute the resulting equilibrium bid-ask spreads of the market maker and the resulting expected markup flows. Equilibrium actions of trading firms given informational events are straightforward: At *public news events*, the liquidity provider sends a message to adjust her quotes by $\pm J$ while arbitrageurs simultaneously submit market orders to snipe the stale quotes.[19] At *other trading times* (i.e. upon arrival of either informed or uninformed investors), the liquidity provider renews her quotes by the rational price impact Δ of any single order according to Bayesian updating. There are hence two sources of adverse selection risk for the market maker in this model. First, she potentially trades against informed traders who possess superior private information. Second, upon publicly observable jumps in the fundamental value, she trades against other HFTs who pick up her quotes that have become stale. The liquidity provider will incur expected losses in both cases. To compensate for these losses, she will quote a strictly positive bid-ask spread which yields expected profits when trading with uninformed investors. By allowing for instantaneous modifications of limit orders, trading firms providing liquidity stand in *perfect* competition. If one firm posts a limit order with a limit price that promises strictly positive expected profits, other trading firms will immediately undercut this order until prices are low enough to yield zero expected profits. The following theorem establishes existence and a characterization of a stationary Markov Perfect Equilibrium in pure strategies in the CLOB market design.

Theorem 1 (CLOB Equilibrium)**.**
There exists a pure-strategy, stationary Markov Perfect Equilibrium of the continuous limit order book market model. The equilibrium is unique up to permutation of trading firms.

Bid and ask prices in equilibrium satisfy

$$b_t = \mathbb{E}[V_t|\mathcal{H}_t, sell] \;=\; P_t - \frac{s_{CLOB}}{2} \qquad \text{and} \qquad a_t \;= \mathbb{E}[V_t|\mathcal{H}_t, buy] \;=\; P_t + \frac{s_{CLOB}}{2},$$

where $P_t = \mathbb{E}[V_t|\mathcal{H}_t]$ denotes the asset's expected value at time t conditional on information $\mathcal{H}_t$ available up to time t, $\mathbb{E}[V_t|\mathcal{H}_t, sell]$ and $\mathbb{E}[V_t|\mathcal{H}_t, buy]$ denote the expected

[19] When assuming infinitely many trading firms (i.e. $N \to \infty$), the market maker's stale quotes will be sniped with probability one upon publicly observable news events.

fundamental value of the asset, conditional on a market sell order and market buy order arriving, respectively, and s_{CLOB} *denotes the spread given by*

$$s_{CLOB} \;=\; 2J\,\frac{\lambda_i+\lambda_j}{\lambda_i+\lambda_j+\lambda_n}. \tag{4.1}$$

After a non-arbitrage buy and sell trade, the asset's conditional expected value jumps to $P_t = P_{t^-} + \Delta$ *and* $P_t = P_{t^-} - \Delta$*, respectively, where the price impact* Δ *is given by*

$$\Delta \;=\; J\,\frac{\lambda_i}{\lambda_i+\lambda_n}. \tag{4.2}$$

Trading costs for investors are equal to $\frac{s_{CLOB}}{2}$ for every trade *and can be decomposed into*

$$C_{CLOB} \;=\frac{s_{CLOB}}{2} \;=\; \Delta \;+\; x_{CLOB} \;=\; \Delta \;+\; J\,\frac{\lambda_j\lambda_n}{(\lambda_i+\lambda_j+\lambda_n)(\lambda_i+\lambda_n)},$$

where x_{CLOB} *denotes the markup component that is due to the risk of stale-quote arbitrage.*

The "expected markup flow" for investors captures expected markups per unit of time. *It is given by*

$$\mathbb{E}[x^f_{CLOB}] \;=\; x_{CLOB}\,(\lambda_i+\lambda_n) \;=\; J\frac{\lambda_j\lambda_n}{(\lambda_i+\lambda_j+\lambda_n)}. \tag{4.3}$$

Proof*:* See Appendix A.1.

Proof idea:
Optimality and uniqueness (up to permutation) follow from profit maximization and perfect competition among trading firms. The equilibrium bid-ask spread results from an equal-profit condition for all trading firms. Since informed and uninformed trade cannot be distinguished, the rational price impact Δ derives from a manipulation of $\mathbb{E}[V_t|\mathcal{H}_t, \text{buy}, i\vee n]$.

The liquidity provider quotes bid and ask prices symmetrically around the mid-price. Positive spreads follow from the twofold asymmetric information problem that the market maker faces, resulting from (i) adverse selection risk from informed traders (Δ) and (ii) sniping risk from arbitrageurs (x_{CLOB}). Both of these risks are subsumed

in the half-spread $\frac{s_{\text{CLOB}}}{2}$ which represents the total trading costs for every trade in this model.[20] More informative in the lights of a comparison of the distinct inefficiencies between the continuous CLOB design and the discrete FBA setup are what we call *"expected markup flows"*.[21] These are the markups which investors expect to pay to the market maker *per unit of time* – in addition to the rational price impact of orders – and serve as our foremost *measure of inefficiency* of a particular market design. The price impact Δ from informed trading captures the rational component of trading costs and reflects an "efficient" protection of market makers against adverse selection risk. It will always be the same, irrespective of the specific market microstructure. The distinct inefficiency of the CLOB design is hence given by $\mathbb{E}[x^f_{\text{CLOB}}]$ and is attributable to the conjunction of continuous-time trading and serial order processing. It will be the key determinant when examining conditions under which either market design is welfare-dominant. Figure 1 below illustrates an exemplary path of equilibrium prices and quotes under CLOB.

4.3 The Frequent Batch Auction Market Design

Budish et al. (2015) propose frequent batch auctions as an alternative market design to CLOB. Under FBA, the trading day is divided into discrete, sealed-bid auction intervals, each of length τ. At the end of an interval, the exchange batches all outstanding orders and computes aggregate demand and supply schedules and a uniform market clearing price. The order history of the current interval is made public by the exchange thereafter.

Before equilibrium strategies and resulting quotes in the FBA setup are discussed, some important differences to the CLOB design should be recapitulated. First, discrete-time uniform price auctions *eliminate* latency arbitrage. Second, continuous-time trading and independent Poisson arrival rates made it sufficient for liquidity providers under CLOB to quote prices at unit depth on either market side. The FBA market design, by contrast, leads to a *deeper order book*. Trading volume of

[20] We abstract from other explicit or implicit trading costs, such as inventory costs or order commissions.

[21] Under CLOB, we could also consider "expected trading cost flows", i.e. $\mathbb{E}[C^f_{\text{CLOB}}] = C_{\text{CLOB}}(\lambda_i + \lambda_n)$, capturing total trading costs per unit of time. However, it includes the rational price impact which is not a distinct market design inefficiency. We therefore consider $\mathbb{E}[x^f_{\text{CLOB}}]$ to be our key inefficiency measure.

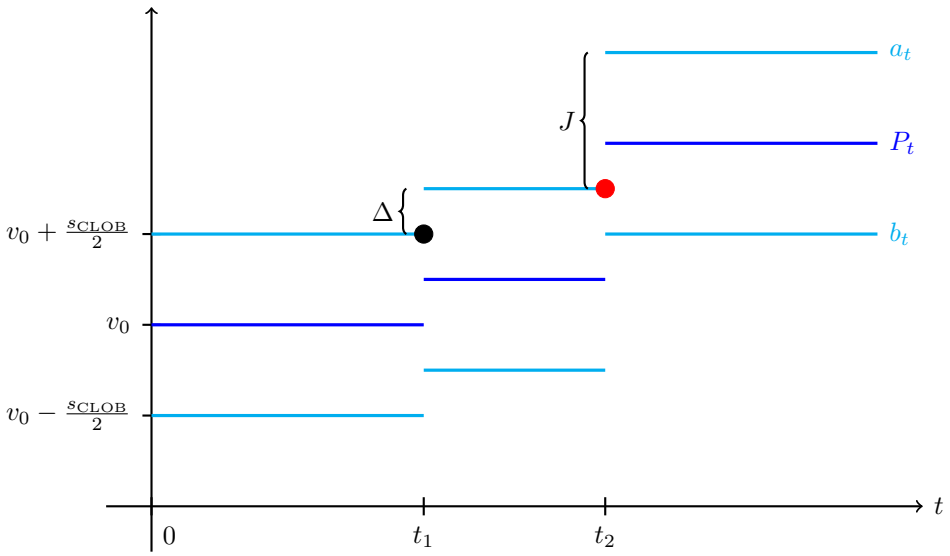

Figure 1: At time $t = 0$, bid $b_0 = v_0 - \frac{s_{\text{CLOB}}}{2}$ and ask $a_0 = v_0 + \frac{s_{\text{CLOB}}}{2}$ lie symmetrically around the asset's initial fundamental value v_0. From then on, the expected fundamental value $P_t = \mathbb{E}[V_t|\mathcal{H}_t]$ evolves as a jump process. At time t_1, a buy trade occurs and induces price impact of Δ. At time t_2, a public news event (fundamental value jumps by $+J$) leads to an arbitrage trade at the old ask price based on information before the event.

HTFs is determined by the *excess demand* (or *order imbalance*) Z_τ, defined as the total number of buy orders, D_τ, minus the total number of sell orders, S_τ, in any auction interval. Since upward and downward jumps in the fundamental value occur with equal probability, the distribution of Z_τ will be symmetric around zero. As do Budish et al. (2015), we assume that excess demand is bounded, i.e. $|Z_\tau| \leq Q + 1$, where Q is a large integer.[22] In each auction interval, liquidity providers must offer at least $Q + 1$ units on each side of the market. Due to the presence of informed traders, these supply schedules are *staggered*. The q^{th} best ask quote, $a_q^{(\tau)}$, has to account for a rational price impact of $q \cdot \Delta$. Third, trading under CLOB is organized by a liquidity provider who quotes bid and ask prices and hence enables trade for all other market participants. Under FBA, buy and sell orders of market participants are matched with each other and quotes of the market maker are only needed to clear *excess* demand or supply. From a more technical perspective, the aforementioned differences make the derivation of equilibria under FBA substantially more involved. This has two main reasons: First, we want to allow for an arbitrarily deep order book

[22]This boundedness assumption is innocuous for a sufficiently large Q. We will analyze equilibrium quotes for an *infinitely deep* order book ($Q \to \infty$) below and show that these quotes converge quickly.

whose configuration is not further restricted. More precisely, we wish to allow for the possibility that a given trading firm can own none, one, several, or even all quotes in the order book. Second, although the notion of an Order Book Equilibrium narrows down the possible reactions to a given deviation, the strategy space of trading firms is still infinitely large. They can add or shift quotes to any desired price level and the profitability of doing so *does* depend on the precise actions they take and on the configuration of the order book (i.e. on which other quotes they already own). In order to account for this fact in the proofs of the following theorems, we let $\omega_q \geq 0$ denote the number of active quotes which a given firm of interest owns in the order book at or below the q^{th} best quote, and $\omega = (\omega_1, ..., \omega_{Q+1})$. Note that we naturally have $0 \leq \omega_1 \leq \cdots \leq \omega_{Q+1} \leq Q+1$. If excess demand in an auction interval is $Z_\tau = q$, all quotes up to and including the q^{th} quote will execute at the same uniform clearing price, i.e. the q^{th} best ask quote $a_q^{(\tau)}$. The revenue of a given trading firm is hence given by $\omega_q \, a_q^{(\tau)}$ while expected profits depend on the possibility to charge markups under FBA. We will show in the following that liquidity providers *can* charge additional markups and earn positive expected profits – even in a *perfectly competitive* market. Remarkably, such an inefficient quoting outcome turns out to be the unique OBE and there does *not* exist any efficient equilibrium with zero markups as the following theorem shows.

Theorem 2 (No Zero Markups in FBA)**.**
The FBA quoting schedule (up to permutation of trading firms for each quote) with zero markups for any unit, i.e. bid and ask schedules of the form

$$\begin{aligned} b_q^{(\tau)} &= P_\tau - q\Delta \qquad q = 1, ..., Q+1 \\ a_q^{(\tau)} &= P_\tau + q\Delta \qquad q = 1, ..., Q+1 \end{aligned} \tag{4.4}$$

is ***not*** *an Order Book Equilibrium. It is also neither a pure-strategy Nash equilibrium in the finitely-repeated game, nor in the discounted infinitely-repeated game.*

Proof*:* See Appendix A.2.

Proof idea:
OBE: Consider, for instance, the liquidity provider who submits the lowest ask quote $a_1^{(\tau)}$ and show that she has a strictly profitable deviation that cannot be rendered unprofitable with reactions covered by the OBE notion.

Nash-Equilibrium: Given the zero-markup schedule, there always exists a strictly profitable unilateral deviation by some liquidity provider.

Theorem 2 shows that any pure-strategy equilibrium under FBA (should it exist) cannot be efficient. The following theorem establishes existence of the Order Book Equilibrium under FBA. It provides a closed-form representation of the unique markups that liquidity providers can charge and characterizes the inefficiencies that arise.

Theorem 3 (FBA Equilibrium)**.**
There exists an Order Book Equilibrium in pure strategies of the FBA market model. If excess demand in an auction is truncated to $Q+1$, i.e. $|Z_\tau| \leq Q+1$, supply and demand schedules in equilibrium satisfy

$$b_q^{(\tau)} = P_\tau - q\Delta - x_q, \qquad q = 1, \dots, Q+1$$
$$a_q^{(\tau)} = P_\tau + q\Delta + x_q, \qquad q = 1, \dots, Q+1$$

where $\Delta = J\frac{\lambda_i}{\lambda_i+\lambda_n}$ is the usual expected price impact of an order and x_q represents the additional markup that the liquidity provider can charge in the q^{th} best quote. These markups are given by $x_{Q+1} = 0$ and

$$x_q = \Delta \sum_{k=0}^{Q-q} \prod_{s=k}^{Q-q} \alpha_{Q-s}, \qquad q = 1, ..., Q, \tag{4.5}$$

$$\text{where} \qquad \alpha_q := \frac{p_{q+1}}{p_q + p_{q+1}} \qquad \text{and} \qquad p_q := \mathbb{P}(Z_\tau = q \mid |Z_\tau| \leq Q+1).$$

Equilibrium markups are unique up to permutation of trading firms for each submitted quote. The order imbalance measure Z_τ follows a symmetric truncated Skellam distribution with parameter $\frac{1}{2}\tau(\lambda_i + \lambda_n)$. The total expected markup flow that investors expect to pay to liquidity providers per unit of time is given by

$$\mathbb{E}[x^f_{FBA}] = \frac{2}{\tau}\sum_{k=1}^{Q} k\, p_k\, x_k \tag{4.6}$$

Proof*:* See Appendix A.3.

Proof idea:

Markups: Arise from a zero-profit condition for underbidding attempts of other firms.

OBE Existence and Uniqueness: Show by induction that the markup x_q, $q = 1, ..., Q+1$, as given by (4.5) does neither allow for strictly profitable safe price improvements nor for strictly profitable robust deviations while any markup $\tilde{x}_q$ different from (4.5) does so.

Probability distribution of excess demand: $Z_\tau := D_\tau - S_\tau$ is the difference between two independent Poisson random variables and thus follows a truncated Skellam distribution.

Theorem 3 quantifies the inefficiencies that result in the FBA market design, once augmented by privately informed investors. The equilibrium in the FBA market design is remarkable in that *perfectly competitive* liquidity providers can charge additional markups over and above what is necessary to account for informed trade and no other firm can prevent them from doing so. We will show that all markups $x_1, ..., x_Q$ are strictly positive, and so is the resulting expected markup flow $\mathbb{E}[x^f_{\text{FBA}}]$. Therefore, equilibrium prices under FBA are excessive. Figure 2 illustrates equilibrium quotes under FBA with an order book depth of $Q + 1 = 4$ and highlights the inefficiency.

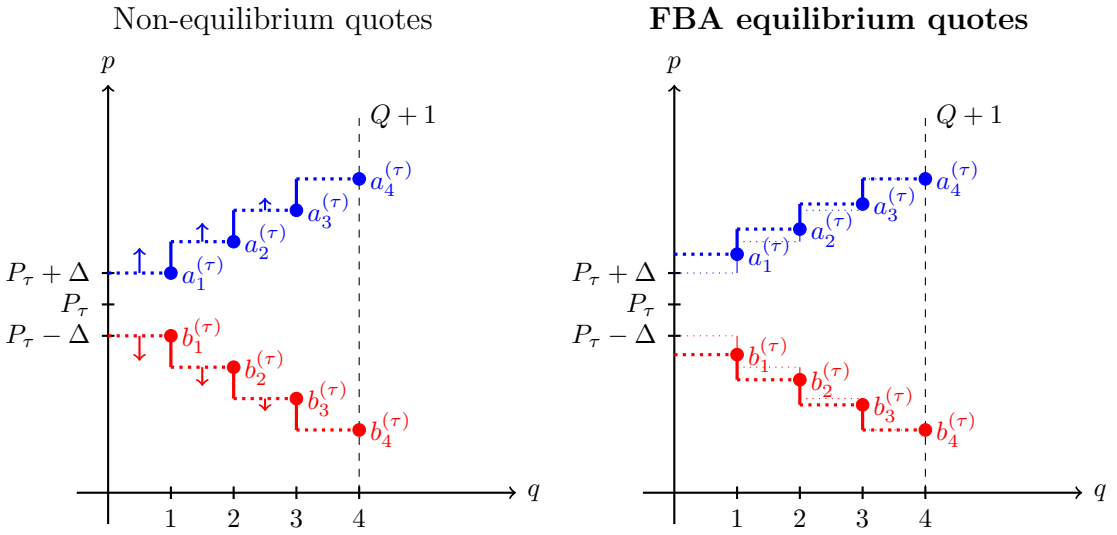

Figure 2: *Left*: The zero-markup schedule (only accounting for the price impact Δ) does *not* reflect equilibrium quotes under FBA. The arrows indicate that ask (bid) quotes can be increased (decreased) – even in case of perfect competition and under risk neutrality. *Right*: FBA equilibrium quotes. The liquidity provider has an incentives to shade all bids except for the quotes of the last unit.

Intuitively, equilibrium markups under FBA with informed trade can be upheld – despite perfect competition among trading firms – because quoted units are not "independent". Suppose some previously inactive tradding firm underbids a given ask quote, say $a_q^{(\tau)} = q\,\Delta + x_q$, by some small amount $\varepsilon > 0$. Then this undercutting attempt influences the trading price if excess demand turns out to be $Z_\tau \geq q$. More precisely, the underbidding firm would benefit if $Z_\tau = q$ since she now offers the q^{th} best quote and earns profits in the amount of $x_q - \varepsilon$. But she would incur losses if $Z_\tau > q$. This is because the old previously q^{th} best quote at markup x_q becomes the new $(q+1)^{th}$ best quote, and similarly, all quotes above the new undercutting quote shift positions by one. As a consequence, each quote above the new undercutting quote accounts too little for the price impact of informed trade and hence results in expected losses.[23] This trade-off allows market makers to charge strictly positive markups x_q for $q = 1, ..., Q$ that cannot be undercut by other firms. For the last unit in the order book, liquidity providers cannot charge a positive markup, i.e. $x_{Q+1} = 0$. This is because excess demand is assumed to be bounded by $Q + 1$ and this unit can hence be undercut at no risk.

The resulting equilibrium in the FBA market design is *asymmetric*: Liquidity providers earn positive expected profits while other trading firms earn zero profits. As equilibrium markups sizes vary between different quotes, liquidity providers will *not* make equal profits in equilibrium – but none has an incentive to deviate. This emphasizes the fact that equilibrium outcomes are determinable *up to permutation of trading firms for each quote*. The closed-form representation (4.5) shows that equilibrium markups depend on the probability distribution of the order imbalance measure $Z_\tau = D_\tau - S_\tau$. Buy and sell orders are independent and equally likely as they are both generated from informed and uninformed investors arriving at the market with known Poisson intensities λ_i and λ_n, respectively. Therefore, Z_τ is the difference between two statistically independent Poisson-distributed random variables and hence follows a truncated Skellam distribution.[24] The closed-form solution of the markups and the

[23] For instance, the old previously q^{th} best quote $a_q^{(\tau)} = q\,\Delta + x_q$ accounts for a rational price impact of $q\,\Delta$. However, due to the undercutting attempt, it now determines the uniform trading price if $Z_\tau = q+1$ in which case the rational price impact to be accounted for should be $(q+1)\,\Delta$. The undercutting firm hence incurs losses in the amount of $\Delta - x_q$. Analogous reasoning applies for any $Z_\tau > q+1$.

[24] Further details of the order imbalance measure Z_τ and the Skellam distribution, including its probability mass function, are provided by Lemma A.4 in the Appendix. In general, the Skellam distribution is characterized by *two* parameters, namely the expected values (or arrival rates) of the

distribution of Z_τ allow us to derive important properties of the individual markups. These properties emphasize the characteristic of the quoting inefficiency under FBA and are summarized in the following proposition.

Proposition 1 (Properties of the FBA Markups)**.**
Suppose $\lambda_i > 0$ (existence of informed investors). Then markups in the FBA design, as given by (4.5), fulfill the following properties for $Q \in \mathbb{N}$ and $q = 1, ..., Q$:

(1)	$x_{Q+1} = 0$	(no markup on last unit)
(2)	$0 < x_q < \Delta$	(positivity and boundedness)
(3)	$x_q > x_{q+1}$	(monotonically decreasing)
(4)	$x_q(Q+1) > x_q(Q)$	(increasing in order book depth)
(5)	$\lim_{Q\to\infty} \Big(x_q(Q+1) - x_q(Q)\Big) = 0$	(convergence in order book depth)

Proof*:* See Appendix A.4.

These markup characteristics are appealing, both from a mathematical and from an economic standpoint: First, all markups - except for the last unit - are strictly positive in the presence of informed trade, with markups on "early units" being largest. Since Skellam probabilities for these units are also largest, this can translate into considerable quoting inefficiencies in the FBA market design.[25] Second, although the markups do depend on the depth of the limit order book, this dependence diminishes and disappears as $Q \to \infty$. In other words, all markups converge. Finally, markup sizes are well-behaved and markups do not "explode": They are all smaller than the price impact $\Delta = J\frac{\lambda_i}{\lambda_i+\lambda_n}$. Figure 3 below shows an exemplary simulation of the FBA markups for an order book depth of $Q + 1 = 150$.

Theorem 3 which established the FBA equilibrium still rests on the assumption that excess demand is bounded, i.e. $|Z_\tau| \leq Q + 1$ where Q is a large integer. The resulting equilibrium markups in (4.5) and hence the markup inefficiency in (4.6) *do* depend on this exogenous bound. The last property of Proposition 1 indicates that the boundedness assumption is innocuous. The following proposition substantiates this

two underlying Poisson processes (here, D_τ and S_τ). Since these arrival rates are identical in our symmetric case, we can characterize the Skellam distribution by a single parameter.

[25]For the symmetric Skellam distribution, it holds that $\mathbb{P}(|Z_\tau| = i) > \mathbb{P}(|Z_\tau| = j)$ for $i < j$ (see Lemma A.4 in the Appendix for details).

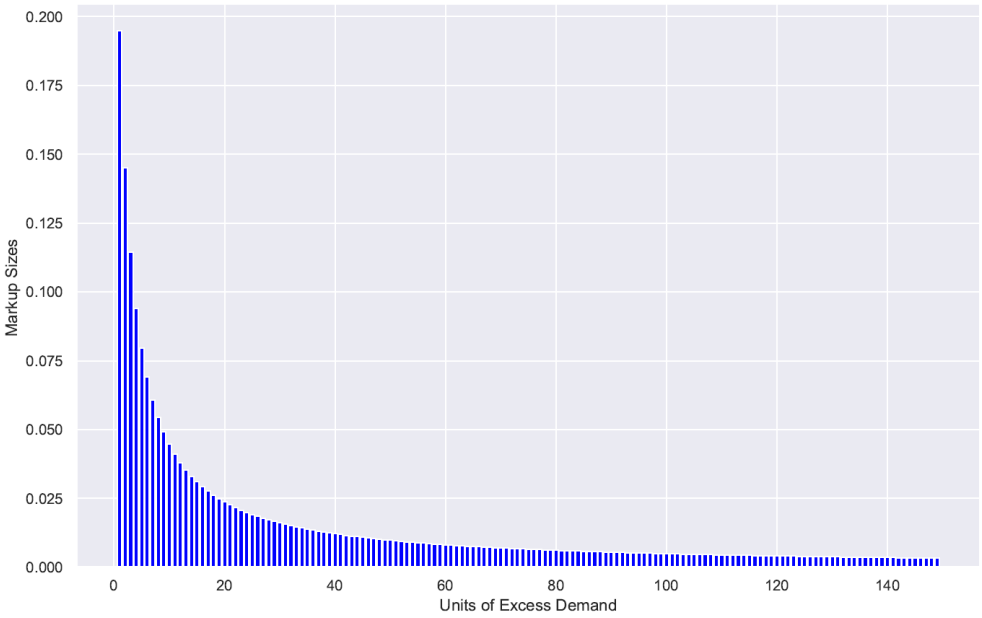

Figure 3: FBA markups for order book depth $Q+1 = 150$ ($\tau = J = \lambda_i = \lambda_n = \lambda_j = 1$).

point and analyzes equilibrium quotes for an infinitely deep order book.

Proposition 2 (FBA Quotes with an Infinitely Deep Order Book)**.**
FBA markups and hence equilibrium supply and demand schedules $a_1^{(\tau)}, ..., a_q^{(\tau)}, ...$ *and* $b_1^{(\tau)}, ..., b_q^{(\tau)}, ...$ *with*

$$\begin{aligned} b_q^{(\tau)} &= P_\tau - q\Delta - x_q, \qquad q = 1, 2, \dots \\ a_q^{(\tau)} &= P_\tau + q\Delta + x_q, \qquad q = 1, 2, \dots \end{aligned}$$

converge for fixed q *and* $Q \to \infty$*. Markups* x_q *are given by (4.5) for any* $q \in \mathbb{N}$ *and excess demand* Z_τ *follows a (now untruncated) symmetric Skellam distribution. The total expected markup flow remains finite, i.e.*

$$\mathbb{E}[x^f_{FBA}] \;=\; \frac{2}{\tau} \sum_{k=1}^{\infty} k\, \mathbb{P}(Z_\tau = k)\, x_k \;<\; \infty.$$

Proof*:* See Appendix A.5.

5 A Comparison of Efficiency and Trading Costs

Our equilibrium analyses have shown that CLOB and FBA suffer from different inefficiencies. Continuous trading and serial order processing give rise to stale quote arbitrage by HFTs. Discrete trading in auctions and processing orders in batches eliminates these arbitrage opportunities but allows liquidity providers to charge positive markups that cannot be competed away. In both cases, the inefficiencies induce higher trading costs that are mainly borne by investors. Budish et al. (2015) illustrate that in the absence of informed trading, FBA dominates the CLOB market design in terms of trading costs. Our equilibrium characterizations indicate that this result does no longer need to hold in the presence of privately informed investors. It is hence not clear a priori which market design is welfare dominant. In our setup, determining the more efficient design reduced to a comparison of the expected markup flows $\mathbb{E}[x^f_{\text{FBA}}]$ and $\mathbb{E}[x^f_{\text{CLOB}}]$ under FBA and CLOB, respectively. We will demonstrate that there exist model parameters such that CLOB is strictly preferable over FBA. The following lemma indicates that the Poisson arrival rates of the different market events will play an important role in this regard.

Lemma 1 (Expected Markup Flow)**.**
The following holds for the expected markup flows under CLOB and FBA, respectively:

$$(i) \quad \frac{\partial}{\partial \lambda_j}\Big(\mathbb{E}[x^f_{CLOB}]\Big) = J\frac{\lambda_n(\lambda_i+\lambda_n)}{(\lambda_i+\lambda_j+\lambda_n)^2} > 0 \quad \textit{and} \quad \frac{\partial}{\partial \lambda_j}\Big(\mathbb{E}[x^f_{FBA}]\Big) = 0.$$

$$(ii) \quad \mathbb{E}[x^f_{CLOB}]\Big|_{\lambda_j=0} = 0 \quad \textit{and} \quad \lim_{\lambda_j\to\infty} \mathbb{E}[x^f_{CLOB}] = J\,\lambda_n.$$

Proof*:* See Appendix A.6.

The first part of Lemma 1 shows that the inefficiencies under CLOB are increasing in the frequency of publicly observable jumps in the fundamental value. Since FBA eliminates stale quote arbitrage, its inefficiencies are independent of these jump events. The second part of Lemma 1 quantifies the markup inefficiencies under CLOB for extreme cases of public news events. These insights allow us to state the main theorem of this paper in the following. It highlights that none of the market designs is unambiguously superior and that a comparison of markup inefficiencies has to depend on model parameters, especially on the frequencies of different market events.

Theorem 4 (Inefficiency Comparison)**.**
Consider any given parameter constellation $(J, Q, \tau, \lambda_i, \lambda_n)$ *with* $\lambda_n > 0$*.*

(i) Without private information, expected markup flow under FBA is zero and strictly lower than under CLOB:

$$0 \;=\; \mathbb{E}[x^f_{FBA}] \;<\; \mathbb{E}[x^f_{CLOB}] \qquad \textit{for } \lambda_i = 0,\ \lambda_j > 0$$

(ii) Without public news events, expected markup flow under CLOB is zero and strictly lower than under FBA:

$$0 \;=\; \mathbb{E}[x^f_{CLOB}] \;<\; \mathbb{E}[x^f_{FBA}] \qquad \textit{for } \lambda_j = 0,\ \lambda_i > 0$$

Proof*:* See Appendix A.7.

The intuition behind Theorem 4 is straightforward: When $\lambda_i = 0$, there is no informed trade and orders do not have price impact, i.e. $\Delta = 0$. As the closed-form representation in (4.5) shows, all markups under FBA – and hence expected markup flow $\mathbb{E}[x^f_{\text{FBA}}]$ – will be zero. By contrast, the risk of stale quote sniping is still present under CLOB, so expected markup flow will be strictly positive. Conversely, when $\lambda_j = 0$, sniping risk is zero and the markup inefficiency under CLOB disappears. Under FBA, by contrast, liquidity providers can still charge additional markups in their quotes. This results in higher markup inefficiencies under FBA. Theorem 4 induces an important corollary which we state in the following. It provides a characterization of the possible situations that can arise when comparing the markup inefficiencies in both market designs and shows that there will always exist model parameters such that CLOB is more favorable than FBA in terms of expected markup flows.

Corollary 1 (Markup Flow Boundary)**.**
Consider any given parameter constellation $(J, Q, \tau, \lambda_i, \lambda_n)$ *with* $\lambda_n > 0$ *and* $\lambda_i > 0$*. Then, two cases can arise for expected markup flows:*

Case 1: *(CLOB unambiguously preferred over FBA)*

$\nexists\, \lambda_j$ *s.t.* $\mathbb{E}[x^f_{CLOB}] \geq \mathbb{E}[x^f_{FBA}]$*. This is the case whenever*

$$\mathbb{E}[x^f_{CLOB}] \xrightarrow{\lambda_j \to \infty} J\,\lambda_n \quad < \quad \frac{2}{\tau}\sum_{k=1}^{Q} k\,x_k\,p_k \;=\; \mathbb{E}[x^f_{FBA}], \tag{5.1}$$

and therefore $\mathbb{E}[x^f_{CLOB}] < \mathbb{E}[x^f_{FBA}] \ \forall \lambda_j \geq 0$.

Case 2: *(Superiority depends on frequency of public news events)*

$\exists\, \lambda_j$ *s.t.* $\mathbb{E}[x^f_{CLOB}] \geq \mathbb{E}[x^f_{FBA}]$. *This is the case whenever*

$$\mathbb{E}[x^f_{CLOB}] \xrightarrow{\lambda_j \to \infty} J\,\lambda_n \quad \geq \quad \frac{2}{\tau}\sum_{k=1}^{Q} k\,x_k\,p_k \;=\; \mathbb{E}[x^f_{FBA}]. \tag{5.2}$$

In this case, there exist a unique $\lambda^*_j = \lambda_j(\lambda_i, \lambda_n, \tau, Q, J)$ *given by*

$$\lambda^*_j = \frac{\mathbb{E}[x^f_{FBA}]\,\left(\lambda_i + \lambda_n\right)}{J\lambda_n - \mathbb{E}[x^f_{FBA}]} \tag{5.3}$$

that induces equality between the expected markup flows under CLOB and FBA, i.e. $\mathbb{E}[x^f_{CLOB}] = \mathbb{E}[x^f_{FBA}]$. *For* $\lambda_j < \lambda^*_j$, *we have* $\mathbb{E}[x^f_{CLOB}] < \mathbb{E}[x^f_{FBA}]$, *and vice versa.*

Proof*:* See Appendix A.8.

Proof idea:

$\lambda_j \to \infty$ puts CLOB in the least favorable position as sniping risk is maximal. If expected markup flow is still smaller than under FBA, then CLOB is unambiguously preferred (Case 1). Otherwise, we can solve $\mathbb{E}[x^f_{\text{CLOB}}] = \mathbb{E}[x^f_{\text{FBA}}]$ for a unique λ^*_j (Case 2).

Corollary 1 can be interpreted as to attenuate the advantageous position from Budish et al. (2015) that FBA seems to have over CLOB. It shows that the interplay between the Poisson arrival rates of public news events (λ_j) and informed trade (λ_i) are key determinants for the relative inefficiencies in the different market designs. Two scenarios can arise for expected markup flows. In the first case, CLOB is unambiguously welfare-dominant since it induces lower markups than FBA for *any* λ_j, i.e. no matter how high the risk of stale-quote arbitrage is. In the second case, there exists a unique $\lambda^*_j > 0$, given by (5.3), which induces a *markup flow boundary* between CLOB and FBA. For any given level of uninformed trade, λ^*_j separates the $\lambda_j - \lambda_i-$parameter space into two parts and each market design is welfare-dominant in one of them.[26]

[26] One could also fix a given level of informed trade, λ_i, and consider the $\lambda_j - \lambda_n$-parameter space.

Corollary 1 allows us to make a precise prediction as to when each market design is preferable in terms of expected markups per unit of time. Most importantly, it shows that the dominant position of FBA without privately informed investors might no longer hold in the presence of informed trade. Figure 4 shows the markup flow boundary induced by λ_j^* for two different lengths τ of the auction interval.

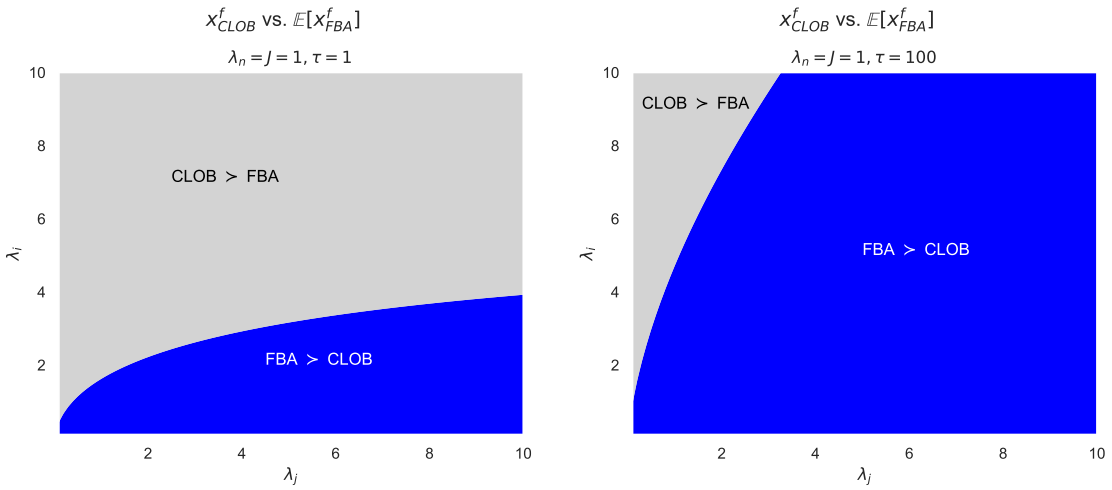

Figure 4: Markup flow boundary in the $\lambda_j - \lambda_i$−space induced by λ_j^* as given by (5.3). The length of the auction interval τ is varied (left: $\tau = 1$, right: $\tau = 100$).

Longer auction intervals make arrivals of multiple investors - and hence higher realizations of excess demand – more likely. While total expected markups within a longer interval increase as a consequence, the rate of this increase is less than linear in τ so that expected markups *per unit of time* decrease with τ. From a more intuitive perspective, longer intervals lead to more investor arrivals, on average. However, trading firms only clear *excess* demand, so many of these investors are matched with each other in an auction. Therefore, total markups paid to trading firms increase, but only by a factor that is less than τ. To increase efficiency in frequent batch auctions, longer intervals could be suggested but those might not be compatible with the fast-changing environment of today's financial markets in practice. A higher level of informed trade, λ_i, exacerbates the inefficiency under FBA and makes this market design less attractive. Figure 5 provides further intuition for the shape of the markup flow boundary. It depicts comparative statics for markup flows under CLOB and FBA of the auction interval length and the arrival rate of informed trade, respectively.

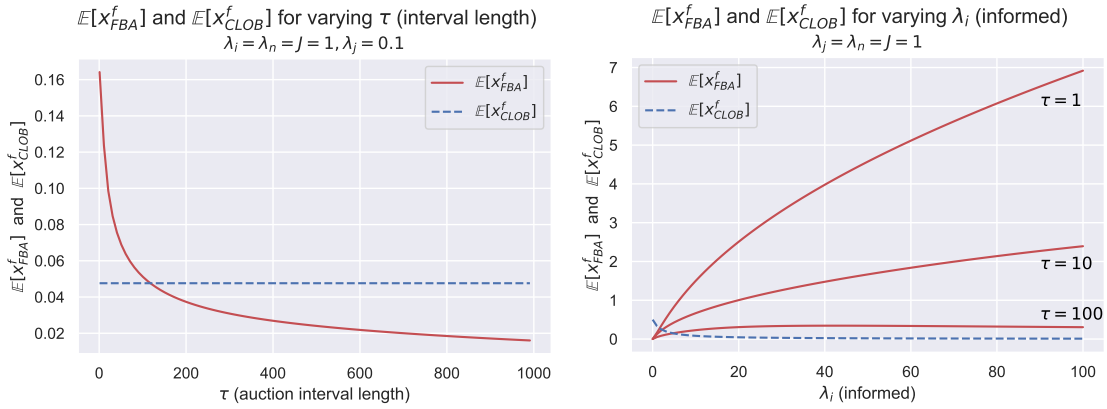

Figure 5: Comparative statics of the auction interval length τ (left) and the arrival rate of informed trade λ_i (right) on expected markup flows under CLOB and FBA.

Discussion of the Main Result

While FBA eliminates stale-quote arbitrage, it can lead to inefficient prices. When abstracting from informed trade ($\lambda_i = 0$), our model's implications are analogous to those in Budish et al. (2015) in that FBA is welfare-dominant. Once there are privately informed traders in the market, however, this result does no longer hold.[27] The equilibrium that we have derived relies on several important assumptions, most of which are common in the market microstructure literature. First, informed investors and noise traders can only use market orders and strategic interaction is limited to trading firms. Once informed investors could also use limit orders, further equilibria in the FBA market design would exist, one of which is efficient and leads to zero markups. Second, we have shown that there *exist* parameter constellations such that CLOB is strictly preferred over FBA in terms of markup inefficiencies. Our result critically relied on the frequencies of different market events. In practice, it is notoriously difficult to distinguish between informed and noise trade and to quantify publicly observable jumps in the fundamental value. Third, we have focused on informed trading and risks from stale-quote arbitrage but have abstracted from other costs in both market designs. Fourth, equilibrium outcomes were unique only up to permutation of trading firms and quotes. Therefore, it is undetermined how trading firms sort into HFTs and liquidity providers and how a certain equilibrium is reached. Finally, in order to obtain closed-form results, we have assumed a very simple process for the

[27] This is true although we have made modeling assumptions as conservatively (i.e. as unfavorable) as possible for trading costs under CLOB.

fundamental asset value. Nonetheless, we have shown that the seemingly dominant position of FBA over CLOB is itself contingent on several crucial assumptions - first and foremost, the abstraction from informed trade. Well-known results from auction theory mitigate the efficiency of frequent batch auctions and can lead to externalities that outweigh the costs from stale-quote arbitrage under CLOB.

6 A Possible FBA Adjustment

Although the underlying motive for introducing FBA (i.e. eliminating stale-quote arbitrage and a socially wasteful arms race for speed) is noble, we have shown that this market design does not come without unintended drawbacks. Nonetheless, the fundamental contributions of Budish et al. (2015) should not be devalued. Their seminal work has laid the foundation for a critical re-evaluation of existing microstructures and has shed light on important inefficiencies that can be reduced by clever market design alternatives.[28] Furthermore, the inefficiencies under FBA that we highlight can be eliminated with relatively straightforward adjustments as we show in this section. The main reason why liquidity providers can charge markups under FBA is that price competition for different quotes is not "independent". In an auction with excess demand $Z_\tau = k$, the clearing price is determined by the k^{th} best quote. Suppose some trading firm underbids the k^{th} ask quote a_k, this undercutting affects the trading price if excess demand turns out to be $Z_\tau \geq k$. The undercutting firm would benefit if $Z_\tau = k$ but would incur losses if $Z_\tau > k$. This trade-off allows market makers to charge strictly positive markups that cannot be undercut by other firms. In order to prevent the inefficiency, an adjustment should make price competition *independent* for each quote for every possible realization of excess demand.

One possible approach to restore independent price competition across units is to allow market makers to submit quotes *conditional on excess demand* in an auction. Let most rules of the current FBA implementation remain intact. In particular, discrete-time auction intervals still eliminate stale-quote arbitrage, given common trading technologies. But suppose that, instead of submitting "unconditional" quotes, liquidity

[28] Another interesting approach, opposite to discretization, is due to Kyle and Lee (2017). They propose to make trading fully continuous by letting traders submit trade rates over time instead of quantities. Trade rates would allow to perfectly smooth larger trades over time, eliminating temporary price impact and thus limiting sniping opportunities for HFTs. While theoretically appealing, this suggestion seems more drastic and less practicable than variations of the FBA design.

providers can submit price-quantity schedules for every realization of excess demand. More precisely, each liquidity provider $i = 1, ..., N$ can submit an ask schedule (and an analogous bid schedule) of the form

$$\left(\left\{ (a^i_{1,j}, n^i_{1,j}) \right\}_{j=1}^{q^i_1} , \left\{ (a^i_{2,j}, n^i_{2,j}) \right\}_{j=1}^{q^i_2} , \ ... \ , \left\{ (a^i_{Q+1,j}, n^i_{Q+1,j}) \right\}_{j=1}^{q^i_{Q+1}} \right) \tag{6.1}$$

whereby $(a^i_{k,j}, n^i_{k,j})$ is the j's price-quantity quote (out of q^i_k in total) reflecting firm i's willingness to sell $n^i_{k,j}$ units (with $n^i_{k,j} \in \mathbb{N}_0$) at price $a^i_{k,j}$ if excess demand is $Z_\tau = k$. Since k is the maximum number of units that firm i could sell when $Z_\tau = k$, we impose $\sum_{j=1}^{q^i_k} n^i_{k,j} \leq k$. As $n^i_{k,j} = 0$ for $j = 1, ..., q^i_k$ is permitted, liquidity providers are not forced to submit quotes for *every* possible realization of excess demand.[29]

For each realization of excess demand $Z_\tau = k$ with $k = 1, ..., Q+1$, the uniform trading price is determined by the k^{th} best among all quotes that have been submitted conditional on $Z_\tau = k$. More formally, let M_k denote the multiset of all quotes submitted by the N trading firms, conditional on excess demand being $Z_\tau = k$.[30] In other words, M_k contains all quotes $a^1_{k,1}, ..., a^1_{k,q^1_k}, \ a^2_{k,1}, ..., a^2_{k,q^2_k}, \ ..., \ a^N_{k,1}, ..., a^N_{k,q^N_k}$ where each $a^i_{k,j}$ has multiplicity $n^i_{k,j}$ and hence $|M_q| = \sum_{i=1}^{N} \sum_{j=1}^{q^i_k} n^i_{k,j}$. Furthermore, let $m(M_k, k)$ denote the k^{th} smallest element of M_k. If excess demand is $Z_\tau = k$, then the trading price in the adjusted FBA design is given by $m(M_k, k)$. Note especially that if the k lowest quotes in M_k are all identical (as will be the case in equilibrium), the trading price is given by this common ask price. Finally, one has to specify how the k units are allocated to those firms who submit quotes below or equal to $m(M_k, k)$. *Price-priority* should be upheld while, in case of ties, we suggest a *pro-rata*-type rationing rule based on offered quantities. We denote by n^{low}_k (respectively n^{equ}_k) the total number of quotes in M_k with ask prices strictly below (respectively equal to) $m(M_k, k)$ and by $\mathcal{N}^{low}_k \subseteq \{1, ..., N\}$ (respectively $\mathcal{N}^{equ}_k \subseteq \{1, ..., N\}$) the subset of firms who submit these ask quotes. Several cases must be considered. First, if $n^{low}_k + n^{equ}_k = k$, each liquidity provider $i \in \mathcal{N}^{low}_k \cup \mathcal{N}^{equ}_k$ is allowed to sell her desired number of units. The same is true if $n^{low}_k + n^{equ}_k < k$, although some units of excess demand cannot be filled in this case.[31] Finally, if $n^{low}_k + n^{equ}_k > k$, the k

[29] In practice, simple rule-based algorithms should allow liquidity providers to submit their desired quoting schedules with relative ease.

[30] Informally speaking, a multiset is a modification of the concept of a set which allows for multiple instances of the same element.

[31] One could either carry this unfulfilled demand into the next auction interval, as Budish et al.

units must be rationed. Due to *price-priority*, firms within $\mathcal{N}_k^{low}$ can sell all their desired quantities with ask prices below $m(M_k, k)$ while the remaining units, $\widetilde{k} := k - \sum_{i \in \mathcal{N}_k^{low}} \sum_{j=1}^{q_k^i} n_{k,j}^i \mathbb{1}_{\{a_{k,j}^i < m(M_k,k)\}}$, must be rationed among the firms within $\mathcal{N}_k^{equ}$. With *pro-rata* rationing, the number of units allocated to each firm $i \in \mathcal{N}_k^{equ}$ is proportional to $n_{k,equ}^i$, the total number of quotes submitted by firm i with an ask price equal to $m(M_k, k)$.[32] Figure 6 illustrates how prices are determined and how units are allocated under the proposed FBA adjustment. Proposition 3 shows that markups can no longer be sustained in the adjusted FBA design and that the resulting unique pure-strategy equilibrium is efficient.

Proposition 3 (Equilibrium in the Adjusted FBA Setup)**.**
There exists a pure-strategy, stationary Markov Perfect Equilibrium in the adjusted FBA market model. Equilibrium bid and ask schedules are given by

$$\begin{aligned} b_q^{(\tau)} &= P_\tau - q\Delta \qquad q = 1, ..., Q+1 \\ a_q^{(\tau)} &= P_\tau + q\Delta \qquad q = 1, ..., Q+1, \end{aligned} \tag{6.2}$$

where $\Delta = J\frac{\lambda_i}{\lambda_i + \lambda_n}$ *denotes the price impact of an order. The equilibrium is efficient and induces zero expected markups. The equilibrium is unique up to the subsets of firms submitting the* k^{th} *best quote on each market side for each value of excess demand* $Z_\tau = k$ *with* $k = 1, ..., Q+1$.

Proof*:* The adjusted FBA setup gives rise to separate games of Bertrand competition for each quote so that classical zero-profit equilibrium results apply. ■

The proposed FBA adjustment creates $Q+1$ separate games of Bertrand competition. In each of these games, trading firms compete for price leadership for all possible values of excess demand. Undercutting solely affect the trading price for this level of excess demand and imposes no externality on other levels of excess demand. The classic zero-profit result in Bertrand competition applies and equilibrium prices are efficient.

(2014) suggest, or cancel the respective market orders (similar to fill-or-kill orders). In any case, $n_k^{low} + n_k^{equ} < k$ will never occur in equilibrium.

[32] More precisely, one possibility is to allow firm $i \in \mathcal{N}_k^{equ}$ to sell $\widetilde{n}_k^i := \left\lfloor \frac{n_{k,equ}^i}{\sum_{j \in \mathcal{N}_k^{equ}} n_{k,equ}^j} \widetilde{k} \right\rfloor$ units where $\lfloor \cdot \rfloor$ is the floor function. Should some units remain (which is the case when $\sum_{i \in \mathcal{N}_k^{equ}} \widetilde{n}_k^i < \widetilde{k}$), then those remaining units are allocated randomly among all firms within $\mathcal{N}_k^{equ}$ that have been assigned fewer units than desired by them (i.e. among those firms $i \in \mathcal{N}_k^{equ}$ for which $\widetilde{n}_k^i < n_{k,equ}^i$).

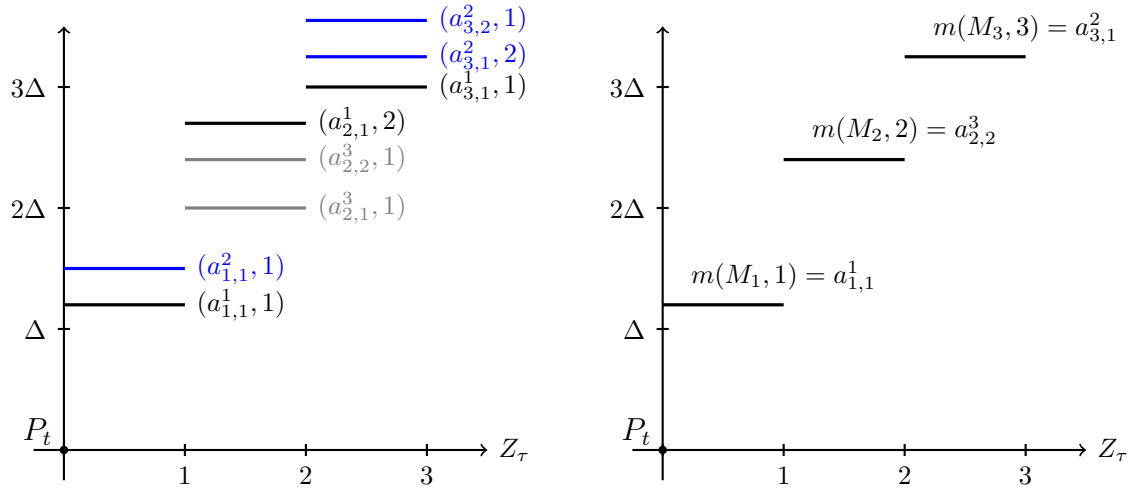

Figure 6: Price determination and quantity allocation in the proposed FBA adjustment. Note that this figure does *not* depict equilibrium quoting behavior. *Left*: Three trading firms, labeled 1 (black), 2 (blue), and 3 (gray) submit price-quantity-schedules as in (6.1), where $(a^i_{k,j}, n^i_{k,j})$ denotes the j's price-quantity quote of firm i, reflecting her willingness to sell $n^i_{k,j}$ units at price $a^i_{k,j}$ if excess demand is $Z_\tau = k$. *Right*: Resulting prices $m(M_k, k)$ are determined by the k^{th} best quote for excess demand of $Z_\tau = k$. If $Z_\tau = 1$ ($Z_\tau = 2$), firm 1 (firm 3) is the only seller of those units. If $Z_\tau = 3$, the quote $a^2_{3,1}$ of firm 2 determines the trading price. Firm 1 is allocated one unit since she was willing to sell below the trading price. The remaining two units are allocated to firm 2.

As trading firms submit quotes without markups, investors only face costs arising from the Bayesian price impact of orders due to the presence of informed traders. Consequently, our measure of inefficiency – expected markup flow – is zero for the adjusted FBA design.

7 Conclusion

The goal of this paper was to compare trading costs and (in)efficiencies under two different market designs: The continuous limit order book (CLOB) and frequent batch auctions (FBA). In their seminal and pioneering work, Budish et al. (2015) highlight that continuous trading and serial order processing can lead to distinct inefficiencies which can be avoided using frequent batch auctions. We have shown that such multi-unit auctions, in the presence of informed trading, induce other inefficiencies that closely resemble bid-shading or demand reduction, well-known concepts in auction theory. When some investors have superior private information and orders have price

impact, liquidity providers under FBA can charge markups and earn profits in equilibrium which cannot be competed away – even in a perfectly competitive environment and under risk-neutrality. Higher trading costs are mainly borne by investors who suffer from excessive equilibrium prices.

The main result of this paper shows that there exist circumstances such that the inefficiencies under CLOB are strictly lower than under FBA. Our inefficiency measure, expected markup flow, quantifies expected markups per unit of time that informed and uninformed investors have to pay under different market microstructures. Key determinants for markups are the frequency of publicly observable jumps in the fundamental value and the frequency of informed trading. We have shown that CLOB is either strictly preferable or there exists a critical level of public news events, λ_j^*, inducing a "markup-flow-boundary" between CLOB and FBA. In the latter case, CLOB is still the preferable market design in terms of markup inefficiencies for all public news frequencies smaller than λ_j^*, and FBA is preferable otherwise. Finally, we have proposed a straightforward adjustment to the FBA market design that fixes the aforementioned inefficiency and leads to a unique pure-strategy equilibrium with zero markups.

A Proofs

General remark: Proofs will be conducted for one side of the order book only; all reasoning symmetrically applies to the respective other side as well.

A.1 Proof of Theorem 1 (CLOB Equilibrium)

Bid-ask spread. In equilibrium, the liquidity provider and HFTs must earn identical expected profits since they are free to choose their role.

$$\underbrace{\overbrace{\lambda_n \frac{s_{\text{CLOB}}}{2}}^{\text{profit from uninformed}} \quad \overbrace{-\lambda_i (J - \frac{s_{\text{CLOB}}}{2})}^{\text{loss from informed}} - \overbrace{\lambda_j \frac{N-1}{N} (J - \frac{s_{\text{CLOB}}}{2})}^{\text{loss from arbitrageur}}}_{\text{profit for liquidity provider}} \overset{!}{=} \underbrace{\lambda_j \frac{1}{N} (J - \frac{s_{\text{CLOB}}}{2})}_{\text{profit for arbitrageur}}$$

Rearranging this equation yields the equilibrium bid-ask spread $s_{\text{CLOB}} = 2J \frac{\lambda_i + \lambda_j}{\lambda_i + \lambda_j + \lambda_n}$.

Price impact. The equilibrium price impact is given by the expected change in the fundamental value upon a non-arbitrage trade.[33] For a buy order, we have

$$\begin{aligned}
&\mathbb{E}[V_t | \mathcal{H}_t, \text{buy}, i \vee n] \\
&= \mathbb{E}[V_t | \mathcal{H}_t, \text{buy}, i] \, \mathbb{P}(i | \mathcal{H}_t, \text{buy}, i \vee n) + \mathbb{E}[V_t | \mathcal{H}_t, \text{buy}, n] \, \mathbb{P}(n | \mathcal{H}_t, \text{buy}, i \vee n) \\
&= \Big(\mathbb{E}[V_t | \mathcal{H}_t] + J \Big) \frac{\mathbb{P}(i, \text{buy} | \mathcal{H}_t, i \vee n)}{\mathbb{P}(\text{buy} | \mathcal{H}_t, i \vee n)} + \mathbb{E}[V_t | \mathcal{H}_t] \frac{\mathbb{P}(n, \text{buy} | \mathcal{H}_t, i \vee n)}{\mathbb{P}(\text{buy} | \mathcal{H}_t, i \vee n)} \\
&= \underbrace{\mathbb{E}[V_t | \mathcal{H}_t]}_{= P_t} + \underbrace{J \frac{\lambda_i}{\lambda_i + \lambda_n}}_{=: \Delta}
\end{aligned}$$

which gives the price impact $\Delta = \frac{\lambda_i}{\lambda_i + \lambda_n}$.

Markup. In the CLOB market, the half-spread on every transaction can be decomposed into the rational price impact and an additional markup as follows:

$$\frac{s_{\text{CLOB}}}{2} = \Delta + x_{\text{CLOB}}$$

Rearranging yields equilibrium markup $x_{\text{CLOB}} = J \frac{\lambda_j \lambda_n}{(\lambda_i + \lambda_j + \lambda_n)(\lambda_i + \lambda_n)}$.

[33] Note that the liquidity provider *can* distinguish between arbitrage trades (any order upon a public news event) and non-arbitrage trades: With independent Poisson arrivals in continuous time, the probability that a public news event occurs at exactly the same time as a trader arrival is zero.

Expected markup flow per unit of time is given by $\mathbb{E}[x^f_{\text{CLOB}}] = (\lambda_i + \lambda_n)\, x_{\text{CLOB}}$. ■

A.2 Proof of Theorem 2 (No Zero Markups in FBA)

Consider the liquidity provider LP of the first ask quote. We show that the deviation of LP from $a_1 = \Delta$ to $\tilde{a}_1 = \Delta + y_1$ with $y_1 \in \left(0, \frac{p_2}{p_1+p_2}\Delta\right)$ constitutes a *strictly profitable robust deviation*. The deviation is clearly profitable for LP as it increases expected profits. It remains to be shown that there does not exist a strictly profitable *underbidding* reaction of any other trading firm TF in response to LP's deviation.

Let $\omega = (\omega_1, ..., \omega_{Q+1})$ where ω_q denotes TF's number of active quotes at or below a_q (before reacting to LP's deviation). Since LP owns a_1, we have $\omega_1 = 0$. Furthermore, it is most profitable for TF to underbid to $\tilde{a}_1 - \varepsilon$ for some small $\varepsilon > 0$. Two cases of underbidding need to be considered.

(A) TF inserts an additional quote.

$$\Pi^{TF}_{(A)}(\omega) \;=\; p_1(\omega_1{+}1)[y_1{-}\varepsilon] + p_2(\omega_1{+}1)[y_1{-}\Delta] - \sum_{k=3}^{Q+1} p_k(\omega_{k-1}{+}1)\Delta$$

The incentive of TF to underbid the deviation of LP is highest if TF was previously inactive, i.e. for $\omega^* := \arg\max \Pi^{TF}_{(A)}(\omega) = (0, ..., 0)$. But even then, we have

$$\Pi^{TF}_{(A)}(\omega^*) \;=\; y_1(p_1{+}p_2) - \Delta \sum_{k=2}^{Q+1} p_k \quad < \quad 0 \qquad \text{for } y_1 < \frac{\Delta}{p_1{+}p_2} \sum_{k=2}^{Q+1} p_k$$

(B) TF shifts down the k^{th} best existing quote.

$$\Pi^{TF}_{(B)}(\omega) \;=\; p_1(\omega_1{+}1)[y_1{-}\varepsilon] + p_2(\omega_1{+}1)[y_1{-}\Delta] - \sum_{j=3}^{k} p_j(\omega_{j-1}{+}1)\Delta$$

The incentive of TF to underbid the deviation of LP is highest if TF shifts down the second best quote, i.e. if $k = 2$ and $\omega_2 = 1$. But even then, we have

$$\Pi^{TF}_{(B)}(\omega) \;=\; y_1(p_1{+}p_2) - p_2\Delta \quad < \quad 0 \qquad \text{for } y_1 < \frac{p_2}{p_1{+}p_2}\Delta$$

In sum, underbidding LP's deviation is not profitable for any trading firm and hence

the zero-markup schedule cannot constitute an OBE in the one-shot auction game.

Likewise, in the finitely or infinitely repeated (with discounting) version of the auction game, zero-markup quoting does not constitute a Nash equilibrium. Given the zero-markup schedule, any liquidity provider has a strictly profitable one-shot deviation: In some period t, increase the quote by some small enough $\varepsilon > 0$ and remain inactive thereafter. Thereby, the liquidity provider makes strictly positive expected profits in t while making zero expected profits in all other periods – a strict improvement. ■

A.3 Proof of Theorem 3 (FBA Equilibrium)

In preparation of the main proof, we first establish several lemmas.

Lemma A.1 (FBA Markups and No-Entry Condition)**.**
Let excess demand in an auction be truncated to $|Z_\tau| \leq Q+1$ for some $Q \in \mathbb{N}$. Then the markups $\{x_q\}_{q=1,...,Q+1}$ as defined in (4.5) are the largest markups such that inactive trading firms have no incentive to enter the book and undercut existing quotes. If some markup x_q is larger than given in (4.5), then there exists a strict incentive to undercut the corresponding quote.

Proof:
Consider a liquidity provider offering the q^{th} best ask quote at $a_q^{(\tau)} = P_\tau + q\Delta + x_q$ as specified in (4.4). We consider the incentive of an inactive trading firm to enter the book and undercut the ask. The proof proceeds by induction over q.

Base case ($q = Q+1$): In order to have an incentive to undercut the $(Q+1)^{th}$ ask by some sufficiently small $\varepsilon > 0$ to markup $x_{Q+1} - \varepsilon$, an inactive trading firm must earn higher expected profits from undercutting compared to remaining inactive.

$$\underbrace{p_{Q+1}(x_{Q+1} - \varepsilon)}_{\text{undercutting}} \overset{!}{\geq} \underbrace{0}_{\text{inactive}} \qquad \Leftrightarrow \qquad x_{Q+1} \geq \varepsilon > 0$$

Clearly, $x_{Q+1} = 0$ is the largest markup such that an inactive trading firm has no incentive to enter the book and undercut while, if $x_{Q+1} > 0$, there exists a strict incentive to undercut by some sufficiently small $\varepsilon > 0$.

Induction Step ($q+1 \Rightarrow q$): Suppose $x_{q+1}, ..., x_{Q+1}$ satisfy (4.5).

Again, we consider the incentive of an inactive trading firm to undercut the q^{th} quote by some sufficiently small $\varepsilon > 0$ to markup $x_q - \varepsilon$. In case of undercutting, the new quote at markup $x_q - \varepsilon$ becomes the q^{th} best quote, the old previously q^{th} best quote at markup x_q becomes the new $(q+1)^{th}$ best quote, and so on. That means all quotes above the new undercutting quote shift positions by one; the undercutting firm profits if the order balance is $Z_\tau = q$ but incurs expected losses when $Z_\tau = q+1, ..., Q+1$.

$$\underbrace{p_q(x_q - \varepsilon) + \sum_{k=q+1}^{Q+1} p_k(x_{k-1} - \Delta)}_{\text{undercutting}} \overset{!}{\geq} \underbrace{0}_{\text{inactive}}$$

$$\Leftrightarrow \qquad x_q(p_q + p_{q+1}) - p_q\,\varepsilon + \sum_{k=q+1}^{Q} x_k p_{k+1} \;\geq\; \Delta \sum_{k=q+1}^{Q+1} p_k$$

Clearly, the incentive to undercut increases with x_q. To find the largest x_q such that there is no incentive to undercut, we let $\varepsilon \to 0$ and consider the equality

$$x_q(p_q + p_{q+1}) + \sum_{k=q+1}^{Q} x_k p_{k+1} \;=\; \Delta \sum_{k=q+1}^{Q+1} p_k \tag{A.1}$$

It remains to be shown that x_q in (A.1) satisfies (4.5) as well: Consider (A.1) for x_q and x_{q+1}. Subtracting the equation for x_{q+1} from the one for x_q and rearranging gives

$$x_q \;=\; \underbrace{\frac{p_{q+1}}{p_q + p_{q+1}}}_{=:\alpha_q} (\Delta + x_{q+1}) \;=\; \alpha_q \Delta + \alpha_q x_{q+1} \tag{A.2}$$

Finally, inserting the induction hypothesis (4.5) for x_{q+1} into (A.2) yields

$$\begin{aligned}
x_q \;&=\; \alpha_q \Delta + \alpha_q x_{q+1} \overset{(4.5)}{=} \alpha_q \Delta + \alpha_q \left(\Delta \sum_{k=0}^{Q-(q+1)} \prod_{s=k}^{Q-(q+1)} \alpha_{Q-s} \right) \\
&=\; \Delta \left\{ \alpha_q + \alpha_q \left[\alpha_{q+1} + \alpha_{q+1}\alpha_{q+2} + \alpha_{q+1}\alpha_{q+2}\alpha_{q+3} + \cdots + \alpha_{q+1} \cdots \alpha_Q \right] \right\} \\
&=\; \Delta \left\{ \alpha_q + \alpha_q\alpha_{q+1} + \alpha_q\alpha_{q+1}\alpha_{q+2} + \alpha_q\alpha_{q+1}\alpha_{q+2}\alpha_{q+3} + \cdots + \alpha_q\alpha_{q+1} \cdots \alpha_Q \right\} \\
&=\; \Delta \sum_{k=0}^{Q-q} \prod_{s=k}^{Q-q} \alpha_{Q-s}
\end{aligned}$$

proving that x_q satisfies (4.5). ■

Corollary A.1 (FBA Markups and Recursive Relations)**.**
The FBA markups defined in (4.5) satisfy the following recursive relations

$$x_q = \alpha_q(\Delta + x_{q+1}) \tag{A.3}$$

$$x_q(p_q + p_{q+1}) = p_{q+1}(\Delta + x_{q+1}) \tag{A.4}$$

$$p_{q+1}(x_q - \Delta) = x_{q+1}p_{q+1} - x_q p_q \tag{A.5}$$

Lemma A.2 (FBA Markups and Competitive Pressure)**.**
Let excess demand in an auction be truncated to $|Z_\tau| \leq Q+1$ for some $Q \in \mathbb{N}$ and let markups $x_{q+1}, ..., x_{Q+1}$ be given as in (4.5). Then inactive trading firms have the (weakly) strongest incentive to enter the book and undercut the q^{th} quote, i.e. exert the highest competitive pressure on markup x_q.

Proof:
Let markups $x_{q+1}, ..., x_{Q+1}$ be given by (4.5) (Proposition 1 will show that these are strictly smaller than Δ) while $\tilde{x}_1, ..., \tilde{x}_q$ do not necessarily satisfy (4.5). We consider expected profits from undercutting the q^{th} quote, i.e. from offering a quote at markup $\tilde{x}_q - \varepsilon$ for some small $\varepsilon > 0$. Three cases of undercutting need to be considered.

(A) *Inactive* trading firm *adding* a new quote.

$$\begin{aligned}
\mathbb{E}[\pi_{\text{inactive before}}] &= 0 \\
\mathbb{E}[\pi_{\text{inactive adding}}] &= p_q(\tilde{x}_q - \varepsilon) + p_{q+1}(\tilde{x}_q - \Delta) + p_{q+2}(x_{q+1} - \Delta) + \cdots + p_{Q+1}(x_Q - \Delta)
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \mathbb{E}[\Delta\pi_{\text{inactive adding}}] &= \mathbb{E}[\pi_{\text{inactive adding}}] - \mathbb{E}[\pi_{\text{inactive before}}] \\
&\overset{(A.5)}{=} p_q(\tilde{x}_q - \varepsilon) + p_{q+1}(\tilde{x}_q - \Delta) + p_{q+2}\, x_{q+2} - p_{q+1}\, x_{q+1} \pm \cdots + p_{Q+1} \underbrace{x_{Q+1}}_{=0} - p_Q x_Q \\
&= \tilde{x}_q(p_q + p_{q+1}) - p_{q+1}(\Delta + x_{q+1}) - p_q\varepsilon
\end{aligned}$$

To specify profits of *active* trading firms, let $\omega_q \geq 0$ denote the number of active quotes (before undercutting) at or below the q^{th} quote, and $\omega = (\omega_1, ..., \omega_{Q+1})$. For a trading firm to be active, it must own some quote in the book, say at least the k^{th} quote, i.e. $\omega_k = \omega_{k-1}+1$ and $\omega_{k+l} \geq 1$ for all $l \geq 0$. Furthermore, for undercutting to be meaningful, the firm must *not* own the q^{th} quote, i.e. $\omega_q = \omega_{q-1}$.

(B) *Active* trading firm *adding* a new quote.

$$\begin{aligned}\mathbb{E}[\pi_{\text{active before}}(\omega)] \;=\; & p_1\omega_1\tilde{x}_1 + \cdots + p_{q-1}\,\omega_{q-1}\,\tilde{x}_{q-1} + p_q\omega_q\tilde{x}_q \\ & + p_{q+1}\,\omega_{q+1}\,x_{q+1} + \cdots + p_{k-1}\,\omega_{k-1}\,x_{k-1} + p_k\underbrace{\omega_k}_{=\omega_{k-1}+1}x_k + \cdots + p_{Q+1}\underbrace{\omega_{Q+1}}_{\geq 1}x_{Q+1}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\pi_{\text{active adding}}(\omega)] \;=\; & p_1\omega_1\tilde{x}_1 + \cdots + p_{q-1}\,\omega_{q-1}\,\tilde{x}_{q-1} + p_q\underbrace{(\omega_{q-1}+1)}_{=\omega_q+1}(\tilde{x}_q-\varepsilon) \\ & + p_{q+1}(\omega_q+1)(\tilde{x}_q-\Delta) + \cdots + p_{Q+1}(\omega_Q+1)(x_Q-\Delta)\end{aligned}$$

$$\begin{aligned}\Rightarrow \mathbb{E}[\Delta\pi_{\text{active adding}}(\omega)] \;=\; & p_q\tilde{x}_q - p_q(\omega_q+1)\varepsilon + p_{q+1}(\omega_q+1)(\tilde{x}_q-\Delta) \\ & + p_{q+2}(\omega_{q+1}+1)\underbrace{(x_{q+1}-\Delta)}_{<0} + \cdots + p_{Q+1}(\omega_Q+1)\underbrace{(x_Q-\Delta)}_{<0} \\ & - p_{q+1}\,\omega_{q+1}\,x_{q+1} - \cdots - p_{Q+1}\,\omega_{Q+1}\,x_{Q+1}\end{aligned}$$

(C) *Active* trading firm *shifting* downward an existing quote.

$$\begin{aligned}\mathbb{E}[\pi_{\text{active shifting}}(\omega)] \;=\; & p_1\omega_1\tilde{x}_1 + \cdots + p_{q-1}\,\omega_{q-1}\,\tilde{x}_{q-1} + p_q\underbrace{(\omega_{q-1}+1)}_{=\omega_q+1}(\tilde{x}_q-\varepsilon) \\ & + p_{q+1}(\omega_q+1)(\tilde{x}_q-\Delta) + \cdots + p_k(\omega_{k-1}+1)(x_{k-1}-\Delta) \\ & + p_{k+1}\,\omega_{k+1}\,x_{k+1} + \cdots + p_{Q+1}\,\omega_{Q+1}\,x_{Q+1}\end{aligned}$$

$$\begin{aligned}\Rightarrow\; \mathbb{E}[\Delta\pi_{\text{active shifting}}(\omega)] \;=\; & p_q\tilde{x}_q - p_q(\omega_q+1)\varepsilon + p_{q+1}(\omega_q+1)(\tilde{x}_q-\Delta) \\ & + p_{q+2}(\omega_{q+1}+1)\underbrace{(x_{q+1}-\Delta)}_{<0} + \cdots + p_k(\omega_{k-1}+1)\underbrace{(x_{k-1}-\Delta)}_{<0} \\ & - p_{q+1}\,\omega_{q+1}\,x_{q+1} - \cdots - p_k\omega_k x_k\end{aligned}$$

We show that the incentive to undercut is (weakly) maximal in case (A), which follows in three steps.

Step 1: Case (C) dominates case (B): For active firms, undercutting by shifting an existing quote is strictly more profitable than undercutting by adding a new quote.

We clearly have $\mathbb{E}[\Delta\pi_{\text{active shifting}}(\omega)] > \mathbb{E}[\Delta\pi_{\text{active adding}}(\omega)]$ as the latter is largely identical but contains additional strictly negative terms ($x_k - \Delta < 0$ due to Proposition 1).

Step 2: Maximal incentive in case (C): The incentive for an active firm to undercut by shifting is maximal if the firm owns only one single quote, the k^{th} quote to be shifted.

Since $\omega_{q+1}, \dots, \omega_{Q+1}$ only occur in negative terms, maximization of $\mathbb{E}[\Delta\pi_{\text{active shifting}}(\omega)]$ requires these to be minimal. The only other occurring weight, ω_q, is now considered separately. Recall that $\tilde{x}_q \leq \Delta + x_{q+1}$, otherwise it would not be the q^{th} markup. Further recall that ω_q is contained in all $\omega_{q+1}, \dots, \omega_{Q+1}$ or, in other words, increasing ω_q necessarily increases all subsequent weights. Therefore, we get

$$\begin{aligned}
\frac{\partial\mathbb{E}[\Delta\pi_{\text{active shifting}}(\omega)]}{\partial\omega_q} &= -p_q\,\varepsilon + p_{q+1}\underbrace{(\tilde{x}_q - \Delta)}_{\leq x_{q+1}} \\
&\quad + p_{q+2}(x_{q+1} - \Delta) + \dots + p_k(x_{k-1} - \Delta) - p_{q+1}\,x_{q+1} - \dots - p_k x_k \\
&\overset{(A.5)}{\leq} -p_q\,\varepsilon + p_{q+1}\,x_{q+1} + p_{q+2}\,x_{q+2} - p_{q+1}\,x_{q+1} \pm \dots + p_k x_k - p_{k-1}\,x_{k-1} \\
&\quad - p_{q+1}\,x_{q+1} - \dots - p_{k-1}\,x_{k-1} - p_k x_k \\
&= -p_q\,\varepsilon - p_{q+1}\,x_{q+1} - \dots - p_{k-1}\,x_{k-1} \quad < \quad 0
\end{aligned}$$

which shows that maximization of $\mathbb{E}[\Delta\pi_{\text{active shifting}}(\omega)]$ requires ω_q to be minimal as well. Overall, we have

$$\underset{0\leq\omega_1\leq\dots\leq\omega_{Q+1}\leq Q+1}{\arg\max} \mathbb{E}[\Delta\pi_{\text{active shifting}}(\omega)] = \omega^*$$

with $\omega_1^* = \dots = \omega_{k-1}^* = 0$ and $\omega_k^* = \dots = \omega_{Q+1}^* = 1$.

Step 3: Incentive in case (A) is *equal* to the maximal incentive in case (C): To complete

the proof, we now show that $\mathbb{E}[\Delta\pi_{\text{inactive adding}}] = \mathbb{E}[\Delta\pi_{\text{active shifting}}(\omega^*)]$.

$$\begin{aligned}
&\mathbb{E}[\Delta\pi_{\text{active shifting}}(\omega^*)] \\
&\quad = p_q(\tilde{x}_q - \varepsilon) + p_{q+1}(\tilde{x}_q - \Delta) + p_{q+2}(x_{q+1} - \Delta) + \cdots + p_k(x_{k-1} - \Delta) - p_k x_k \\
&\quad \overset{(A.5)}{=} p_q\tilde{x}_q - p_q\,\varepsilon + p_{q+1}\tilde{x}_q - p_{q+1}\Delta + p_{q+2}\,x_{q+2} - p_{q+1}\,x_{q+1} \\
&\qquad \pm \cdots + p_k x_k - p_{k-1}\,x_{k-1} - p_k x_k \\
&\quad = \tilde{x}_q(p_q + p_{q+1}) - p_{q+1}(\Delta + x_{q+1}) - p_q\varepsilon \quad = \quad \mathbb{E}[\Delta\pi_{\text{inactive adding}}]
\end{aligned}$$

■

Lemma A.3 (FBA Markups and No Robust Deviation)**.**
Let excess demand in an auction be truncated to $|Z_\tau| \leq Q+1$ for some $Q \in \mathbb{N}$. Let markups $x_q, ..., x_{Q+1}$ be given by (4.5). Then, there does not exist a strictly profitable robust deviation for the liquidity provider (LP) of the q^{th} quote at markup x_q.

Proof:
Let markups $x_q, ..., x_{Q+1}$ be given by (4.5) (Proposition 1 will show that these are strictly smaller than Δ) while $\tilde{x}_1, ..., \tilde{x}_{q-1}$ do not necessarily satisfy (4.5). Let ω_k denote the number of active quotes that LP owns at or below the k^{th} best ask quote. We show that any (profitable) deviation of LP *can* be rendered unprofitable by a strictly profitable safe price improvement of some previously inactive trading firm. LP's expected profits before deviating are:

$$\mathbb{E}[\pi^{\text{LP}}_{\text{before}}] \;=\; p_1\omega_1\tilde{x}_1 + \cdots + p_q\omega_q x_q + p_{q+1}\,\omega_{q+1}\,x_{q+1} + \cdots + p_{Q+1}\,\omega_{Q+1}\,x_{Q+1}$$

Now suppose LP increases his quote from $a_q = q \cdot \Delta + a_q$ to $\hat{a}_k := k \cdot \Delta + \hat{x}_k$ with $k \geq q$ and $\hat{x}_k \in (x_k, \Delta + x_{k+1})$. In other words, LP shifts his quote upward s.t. he offers the k^{th} best ask quote post deviation. Any such deviation has the form:

$$\begin{aligned}
\mathbb{E}[\pi^{\text{LP}}_{\text{deviate}}] \;=\; & p_1\omega_1\tilde{x}_1 + \cdots + p_{q-1}\,\omega_{q-1}\,\tilde{x}_{q-1} + p_q(\omega_{q+1} - 1)(\Delta + x_{q+1}) \\
& + p_{q+1}(\omega_{q+2} - 1)(\Delta + x_{q+2}) \;+\; \cdots \;+\; p_{k-1}(\omega_k - 1)(\Delta + x_k) \\
& + p_k\omega_k\hat{x}_k + p_{k+1}\,\omega_{k+1}\,x_{k+1} + \cdots + p_{Q+1}\,\omega_{Q+1}\,x_{Q+1}
\end{aligned}$$

However, this deviation is *not* robust: Consider the reaction of a previously inactive trading firm (TF) who enters the market and quotes $a'_q = q \cdot \Delta + x'_q$ with $x'_q = x_q$.

First, this price improvement is strictly profitable for TF: $\mathbb{E}[\pi^{\text{TF}}_{\text{inactive}}] = 0$ and

$$\begin{aligned}\mathbb{E}[\pi^{\text{TF}}_{\text{enter}}](\hat{x}_k) \;=\; & p_q x'_q + p_{q+1}x_{q+1} + p_{q+2}x_{q+2} \;+\; \cdots \;+\; p_k x_k \;+\; p_{k+1}(\hat{x}_k - \Delta) \\ & + p_{k+2}(x_{k+1} - \Delta) + \;\cdots\; + \; p_{Q+1}(x_Q - \Delta)\end{aligned}$$

$\mathbb{E}[\pi^{\text{TF}}_{\text{enter}}](\hat{x}_k)$ is increasing in $\hat{x}_k$. To show that the reaction of TF is profitable for *any* $\hat{x}_k \in (\hat{x}^{\min}_k, \hat{x}^{\max}_k) = (x_k, \Delta + x_{k+1})$, it is sufficient to consider $\hat{x}^{\min}_k = x_k + \varepsilon$:

$$\begin{aligned}\mathbb{E}[\pi^{\text{TF}}_{\text{enter}}](\hat{x}^{\min}_k) \;=\; & p_q x_q + p_{q+1}x_{q+1} \;+\; \cdots \;+\; p_{k-1}x_{k-1} + p_k x_k \;+\; p_{k+1}(x_k + \varepsilon - \Delta) \\ & + p_{k+2}(x_{k+1} - \Delta) \;+\; \cdots \;+\; p_{Q+1}(x_Q - \Delta) \\ =\; & p_{k+1}\varepsilon + p_q x_q + p_{q+1}x_{q+1} \;+\; \cdots \;+\; p_{k-1}x_{k-1} + p_k x_k \\ & + \underbrace{p_{k+1}(x_k - \Delta) + p_{k+2}(x_{k+1} - \Delta) \;+\; \cdots \;+\; p_{Q+1}(x_Q - \Delta)}_{\text{use (A.5) for these terms}} \\ =\; & p_{k+1}\varepsilon + p_q x_q + p_{q+1}x_{q+1} \;+\; \cdots \;+\; p_{k-1}x_{k-1} + p_{Q+1}x_{Q+1} \;>\; 0\end{aligned}$$

Second, TF's price improvement renders any initial deviation of LP to $\hat{x}_k$ unprofitable:

Denoting by $\mathbb{E}[\pi^{\text{LP}}_{\text{react}}](\hat{x}_k)$ the expected profits of LP after TF has reacted, we get:

$$\begin{aligned}\mathbb{E}[\pi^{\text{LP}}_{\text{react}}](\hat{x}_k) \;=\; & p_1\omega_1\tilde{x}_1 + \cdots + p_{q-1}\omega_{q-1}\tilde{x}_{q-1} + p_q(\omega_q - 1)\underbrace{x'_q}_{=x_q} + p_{q+1}(\omega_{q+1} - 1)x_{q+1} \\ & + \;\cdots\; + p_{k-1}(\omega_{k-1} - 1)x_{k-1} + p_k(\omega_k - 1)x_k + p_{k+1}\omega_k(\hat{x}_k - \Delta) \\ & + p_{k+2}\omega_{k+1}(x_{k+1} - \Delta) \;+\; \cdots \;+\; p_{Q+1}\,\omega_Q(x_Q - \Delta)\end{aligned}$$

$\mathbb{E}[\pi^{\text{LP}}_{\text{react}}](\hat{x}_k)$ is increasing in $\hat{x}_k$. To show that the reaction of TF renders LP unprof-

itable for *any* $\hat{x}_k \in (x_k, \Delta + x_{k+1})$, it is sufficient to consider $\hat{x}_k^{\max} = \Delta + x_{k+1} - \varepsilon$:

$$\begin{aligned}
\mathbb{E}[\pi^{\text{LP}}_{\text{react}}](\hat{x}_k^{\max}) \;=\; & p_1\omega_1\tilde{x}_1 + \cdots + p_{q-1}\,\omega_{q-1}\,\tilde{x}_{q-1} + p_q(\omega_q - 1)x_q + p_{q+1}(\omega_{q+1} - 1)x_{q+1} \\
& + \;\cdots\; + p_{k-1}(\omega_{k-1} - 1)x_{k-1} + p_k(\omega_k - 1)x_k + p_{k+1}\omega_k(x_{k+1} - \varepsilon) \\
& + p_{k+2}\omega_{k+1}(x_{k+1} - \Delta) \;+\; \cdots \;+\; p_{Q+1}\omega_Q(x_Q - \Delta) \\
\\
=\; & -p_{k+1}\omega_k\,\varepsilon + p_1\omega_1\tilde{x}_1 + \cdots + p_{q-1}\omega_{q-1}\tilde{x}_{q-1} + p_q(\omega_q - 1)x_q + p_{q+1}(\omega_{q+1} - 1)x_{q+1} \\
& + \;\cdots\; + p_{k-1}(\omega_{k-1} - 1)x_{k-1} + p_k(\omega_k - 1)x_k + p_{k+1}\omega_k x_{k+1} \\
& + p_{k+2}\omega_{k+1}(x_{k+1} - \Delta) \;+\; \cdots \;+\; p_{Q+1}\omega_Q(x_Q - \Delta)
\end{aligned}$$

The profitability difference of LP is:

$$\begin{aligned}
\mathbb{E}[\Delta\pi^{\text{LP}}_{\text{react}}] \;=\; & \mathbb{E}[\pi^{\text{LP}}_{\text{react}}] - \mathbb{E}[\pi^{\text{LP}}_{\text{before}}] \\
\\
=\; & -p_{k+1}\omega_k\varepsilon + p_q\Big[\underbrace{(\omega_q - 1)x_q \;-\; \omega_q x_q}_{<0 \text{ since } (\omega_q - 1) < \omega_q}\Big] \;+\; p_{q+1}\Big[\underbrace{(\omega_{q+1} - 1)\,x_{q+1} \;-\; \omega_{q+1}\,x_{q+1}}_{<0}\Big] \\
& + \;\cdots\; + p_{k-1}\Big[\underbrace{(\omega_{k-1} - 1)\,x_{k-1} \;-\; \omega_{k-1}\,x_{k-1}}_{<0}\Big] \;+\; p_k\Big[\underbrace{(\omega_k - 1)x_k - \omega_k x_k}_{<0}\Big] \\
& + \;p_{k+1}\Big[\underbrace{\omega_k x_{k+1} \;-\; \omega_{k+1}x_{k+1}}_{\leq 0 \text{ since } \omega_k \leq \omega_{k+1}}\Big] \;+\; \underbrace{p_{k+2}\Big[\omega_{k+1}\underbrace{(x_{k+1} - \Delta)}_{<0 \text{ by } \textit{Proposition 1}} - \omega_{k+2}x_{k+2}\Big]}_{<0} \\
& + \;\cdots\; + \underbrace{p_Q\Big[\omega_{Q-1}\underbrace{(x_{Q-1} - \Delta)}_{<0} - \omega_Q x_Q\Big]}_{<0} \;+\; \underbrace{p_{Q+1}\Big[\omega_Q\underbrace{(x_Q - \Delta)}_{<0} - \omega_{Q+1}x_Q\Big]}_{<0} \\
\\
<\; & 0
\end{aligned}$$

The profitable reaction of TF hence renders any deviation of LP unprofitable. ■

Lemma A.4 (Probability Distribution of Excess Demand)**.**
Let excess demand in an auction be truncated to $|Z_\tau| \leq Q{+}1$ *for some* $Q \in \mathbb{N}$*. Then, excess demand* Z_τ *follows a truncated symmetric Skellam distribution with parameter*

$\lambda := \frac{1}{2}\tau(\lambda_i + \lambda_n)$. *Its probability mass function* $\{p_k\}_{k=-(Q+1),\ldots,Q+1}$ *is given by*

$$p_k \;:=\; \mathbb{P}(Z_\tau = k \mid |Z_\tau| \leq Q{+}1) \;=\; \frac{\mathbb{P}(Z_\tau = k)}{\mathbb{P}(Z_\tau \leq Q{+}1) \;-\; \mathbb{P}(Z_\tau \leq -(Q{+}1))} \tag{A.6}$$

where

$$\mathbb{P}(Z_\tau = k) \;=\; e^{-2\lambda}\, I_{|k|}(2\lambda) \tag{A.7}$$

and where $I_{|k|}(\,\cdot\,)$ *denotes the modified Bessel function of the first kind. The probability mass function is symmetric* ($p_k = p_{-k}$) *and maximized at* $k = 0$. *It holds that*

$$p_i \;>\; p_j \;>\; 0 \tag{A.8}$$

for any $0 \leq i < j \leq Q{+}1$ *and*

$$\lim_{k\to\infty} p_k = 0$$

at an exponential rate.

Proof:
Excess demand $Z_\tau := D_\tau - S_\tau$ is the difference between total demand D_τ and total supply S_τ in a given auction interval of length τ. In equilibrium, buy and sell orders can stem from informed and uninformed investors arriving at Poisson intensities of λ_i and λ_n per unit of time, respectively. Both types of investors issue buy and sell orders with equal probability. Letting λ_D and λ_S denote the Poisson intensities of demand and supply orders within a given auction interval, those intensities are given by

$$\lambda_D \;=\; \lambda_S \;=\; \frac{1}{2}\tau(\lambda_i + \lambda_n) \;=:\; \lambda$$

As Z_τ is the difference between two statistically independent random variables, each Poisson distributed with the same intensity λ, excess demand Z_τ follows a symmetric Skellam distribution (truncated, if $Q{+}1 < \infty$) with probability mass function

$$\mathbb{P}(Z_\tau = k) \;=\; e^{-(\lambda_D+\lambda_S)}\Big(\frac{\lambda_D}{\lambda_S}\Big)^{\frac{k}{2}}\, I_{|k|}\Big(2\sqrt{\lambda_D\cdot\lambda_S}\Big) \;=\; e^{-2\lambda}\, I_{|k|}(2\lambda)$$

for $k \in \mathbb{Z}$, where $I_k(z) = \sum_{v=0}^{\infty} \frac{1}{v!\,\Gamma(v+k+1)} \left(\frac{z}{2}\right)^{2v+k}$ denotes the modified Bessel function

of the first kind. It holds for the symmetric Skellam distribution that it is maximized at $Z_\tau = 0$ and that $\mathbb{P}(|Z_\tau| = i) > \mathbb{P}(|Z_\tau| = j)$ for any $i, j \in \mathbb{N}$ with $i < j$.

For small arguments $0 < z \ll \sqrt{k+1}$ and $k \in \mathbb{N}$, the Bessel function has the following asymptotic form

$$I_k(z) \;\simeq\; \frac{1}{\Gamma(k+1)}\Big(\frac{z}{2}\Big)^k$$

where $\Gamma(m)$ is the Gamma function given by $\Gamma(m) = (m-1)!$ for any $m \in \mathbb{N}$. Using this asymptotic result of the Bessel function, we can approximate the symmetric probability mass function of the Skellam distribution by

$$\mathbb{P}(Z_\tau = k) \;=\; e^{-\tau(\lambda_i+\lambda_n)}\, I_{|k|}\Big(\tau(\lambda_i+\lambda_n)\Big) \;\simeq\; e^{-\tau(\lambda_i+\lambda_n)} \frac{1}{|k|!}\Big(\frac{\tau(\lambda_i+\lambda_n)}{2}\Big)^{|k|}$$

Finally, Skellam probabilities tend to zero exponentially as $k \to \infty$, i.e.

$$\lim_{k\to\infty} \mathbb{P}(Z_\tau = k) \;\simeq\; \lim_{k\to\infty} e^{-\tau(\lambda_i+\lambda_n)} \frac{1}{k!}\Big(\frac{\tau(\lambda_i+\lambda_n)}{2}\Big)^k \;\overset{(\star)}{=}\; 0$$

where $(\star)$ follows from the standard analytical result that the factorial sequence will asymptotically grow faster than any exponential function with constant base. This immediately implies $\lim_{k\to\infty} p_k = 0$ for the truncated Skellam probabilities. ■

Lemma A.5 (FBA Markup Flow)**.**
Let excess demand in an auction be truncated to $|Z_\tau| \leq Q+1$ for some $Q \in \mathbb{N}$. For batch auctions of length τ with markups $\{x_k\}_{k=1,\dots,Q+1}$ and probabilities $\{p_k\}_{k=1,\dots,Q+1}$ of excess demand, the expected markup flow for investors per unit of time is given by

$$\mathbb{E}[x^f_{FBA}] = \frac{2}{\tau}\sum_{k=1}^{Q} k\, p_k\, x_k \;<\; \infty.$$

For $\lambda_i > 0$ (existence of informed trade), we have $\mathbb{E}[x^f_{FBA}] \in (0, \infty)$.

Proof:
Compute total expected markup payments to market makers within an auction interval and scale by $\frac{1}{\tau}$, the inverse of the auction interval. We exploit the symmetry

$x_k = x_{-k}$ of markups on the buy and sell side to get

$$\mathbb{E}[x^f_{\text{FBA}}] \;=\; 2\,\frac{1}{\tau}\sum_{k=1}^{\infty} k\,p_k\,x_k \;=\; \frac{2}{\tau}\sum_{k=1}^{Q} k\,p_k\,x_k$$

Q is finite and $p_k \in (0,1)$. Further, it holds that $x_k < \Delta$ for all $k = 1, ..., Q$ which will be shown in Proposition 1. Therefore, $\mathbb{E}[x^f_{\text{FBA}}] < \infty$. Finally, when $\lambda_i > 0$, it holds that $x_k > 0$ for all $k = 1, ..., Q$ which will also be shown in Proposition 1. Hence $\mathbb{E}[x^f_{\text{FBA}}] \in (0, \infty)$ in this case. ■

Proof of Theorem 3:

We show that if the provider of the q^{th} ask quote deviates to a larger markup than the one given in (4.5), then (a) some other trading firm can perform a *strictly profitable safe price improvement* and (b) the price improvement renders the original deviation to a larger markup *unprofitable*.[34] We proceed by induction over q.

Base case ($q = Q+1$): For the last unit, no markup can be charged, i.e. $x_{Q+1} = 0$.

1. $x_{Q+1} = 0$ is an OBE markup. Suppose to the contrary that the corresponding liquidity provider deviates to $\hat{x}_{Q+1} > 0$. Then some other trading firm could undercut and offer at $x'_{Q+1} = \hat{x}_{Q+1} - \varepsilon > 0$ for some small $\varepsilon > 0$ which constitutes a price improvement.
   (a) Undercutting is safe and profitable on expectation: For $Z_\tau < Q+1$, the last unit is not traded, but for $Z_\tau = Q+1$, profit $x'_{Q+1} > 0$ will be realized.
   (b) Undercutting renders the original deviation to markup $\hat{x}_{Q+1} > 0$ unprofitable: The corresponding ask would become the $(Q+2)^{th}$ best ask quote and never execute, thus yielding zero profit.
2. $\hat{x}_{Q+1} \neq 0$ cannot be an OBE markup.
   (1) $\hat{x}_{Q+1} > 0$ cannot be an OBE markup: By Lemma A.1, some inactive trading firm can undercut the markup which constitutes a safe profitable price improvement.

[34] Note that in our setup, *liquidity withdrawals* lead to fewer quotes in the order book and can hence only induce (weakly) larger clearing prices. Therefore, liquidity withdrawals can never render deviations unprofitable and the "*safe*" requirement of any safe profitable price improvement is always fulfilled.

(2) $\hat{x}_{Q+1} < 0$ cannot be an OBE markup: By Lemma A.1 and Lemma A.2, the liquidity provider of this unit can increase her quote until $x_{Q+1} = 0$ which constitutes a profitable robust deviation.

Induction Step ($q+1 \Rightarrow q$): Suppose $x_{q+1}, \ldots, x_{Q+1}$ are given by (4.5). We show that the unique OBE markup x_q must also adhere to (4.5).

1. x_q as in (4.5) is an OBE markup. First, by Lemma A.1 and Lemma A.2, there do not exist profitable price improvements of other firms. Second, by Lemma A.3, there does not exist a profitable robust deviation of the liquidity provider of the q^{th} quote.

2. $\hat{x}_q \neq x_q$ cannot be an OBE markup.

   (1) $\hat{x}_q > x_q$ cannot be an OBE markup by Lemma A.1.

   (2) $\hat{x}_q < x_q$ cannot be an OBE markup by Lemma A.1 and Lemma A.2.

This completes the proof of Theorem 3. ■

A.4 Proof of Proposition 1 (Properties of FBA Markups)

(1) $x_{Q+1} = 0$ **(no markup on last unit)**

See Theorem 3.

(2) $0 < x_q < \Delta$ **(positivity and boundedness)**

Note that $\Delta > 0$ and, for all $q = 1, ..., Q$, we have $p_q > p_{q+1} > 0$ due to (A.8). Thus

$$\alpha_q \;=\; \frac{p_{q+1}}{p_q + p_{q+1}} \;\in \left(0, \frac{1}{2}\right)$$

The proof proceeds by induction over q.

Base case ($q = Q$): We have $x_Q \overset{(A.3)}{=} \underbrace{\alpha_Q}_{\in(0,\frac{1}{2})} \Delta \;\in (0, \Delta)$.

Induction Step $(q+1 \Rightarrow q)$*:* Suppose $x_{q+1} \in (0, \Delta)$ holds for some q. Then

$$x_q \overset{(A.3)}{=} \underbrace{\alpha_q}_{\in(0,\frac{1}{2})} \underbrace{(\Delta + x_{q+1})}_{\in(\Delta,2\Delta)} \in (0, \Delta)$$

(3) $x_q > x_{q+1}$ **(monotonically decreasing)**

We again use (A.3) and proceed by induction over q.

Base case $(q = Q)$*:* We have $x_Q \overset{(A.3)}{=} \alpha_Q(\Delta + x_{Q+1}) > 0 = x_{Q+1}$.

Induction Step $(q+1 \Rightarrow q)$*:* Suppose $x_{q+1} > x_{q+2}$ holds for some q. Then

$$\frac{x_q}{x_{q+1}} \overset{(A.3)}{=} \underbrace{\frac{\alpha_q}{\alpha_{q+1}}}_{>1 \text{ from } (\star)} \underbrace{\frac{(\Delta + x_{q+1})}{(\Delta + x_{q+2})}}_{>1} > 1$$

implying $x_q > x_{q+1}$.

$(\star)$: $\frac{\alpha_q}{\alpha_{q+1}} > 1$ stems from the following considerations.

$$\frac{\alpha_q}{\alpha_{q+1}} = \frac{\frac{p_{q+1}}{p_q + p_{q+1}}}{\frac{p_{q+2}}{p_{q+1} + p_{q+2}}} = \frac{p_{q+1}^2 + p_{q+1}p_{q+2}}{p_q p_{q+2} + p_{q+1}p_{q+2}}$$

Since the second summand is the same in the numerator and denominator, it follows that $\frac{\alpha_q}{\alpha_{q+1}} > 1$ if and only if $\frac{p_{q+1}^2}{p_q p_{q+2}} > 1$. Looking at this term in isolation and using the Skellam probability mass function from (A.7) yields

$$\frac{p_{q+1}^2}{p_q p_{q+2}} = \frac{\left(e^{-2\lambda} I_{q+1}(2\lambda)\right)^2}{e^{-2\lambda} I_q(2\lambda)\, e^{-2\lambda} I_{q+2}(2\lambda)} = \frac{I_{q+1}^2(2\lambda)}{I_q(2\lambda)\, I_{q+2}(2\lambda)} > 1$$

where the last inequality follows from the Turán-type inequalities of the modified Bessel function $I_k(z)$ of the first kind (see Lorch, 1994, eqn. (5.1)).

(4) $x_q(Q+1) > x_q(Q)$ **(increasing in order book depth)**

(5) $\lim_{Q\to\infty} \left(x_q(Q+1) - x_q(Q)\right) = 0$ **(convergence in order book depth)**

We show the last two properties jointly.

$$
\begin{aligned}
x_q(Q{+}1) - x_q(Q) &\overset{(4.5)}{=} \Delta \sum_{k=0}^{Q+1-q} \prod_{s=k}^{Q+1-q} \alpha_{Q+1-s} - \Delta \sum_{k=0}^{Q-q} \prod_{s=k}^{Q-q} \alpha_{Q-s} \\
&= \Delta \prod_{s=0}^{Q+1-q} \alpha_{Q+1-s} + \Delta \sum_{k=1}^{Q+1-q} \prod_{s=k}^{Q+1-q} \alpha_{Q+1-s} - \Delta \sum_{k=0}^{Q-q} \prod_{s=k}^{Q-q} \alpha_{Q-s} \\
&= \Delta \prod_{s=0}^{Q+1-q} \alpha_{Q+1-s} + \Delta \sum_{k=0}^{Q-q} \prod_{s=k}^{Q-q} \alpha_{Q-s} - \Delta \sum_{k=0}^{Q-q} \prod_{s=k}^{Q-q} \alpha_{Q-s} \\
&= \underbrace{\Delta \prod_{s=0}^{Q+1-q} \underbrace{\alpha_{Q+1-s}}_{\in(0,\frac{1}{2})}}_{>0} \quad \searrow \quad 0 \qquad \text{for } Q \to \infty
\end{aligned}
$$

In the above calculation, we have used that α_q does not depend on Q. To see this, denote by $\mathbb{P}(Z_\tau = q)$ and p_q the untruncated and truncated Skellam probabilities of excess demand, respectively. It holds that

$$
\begin{aligned}
p_q &:= \mathbb{P}(Z_\tau = q \mid |Z_\tau| \leq Q{+}1) \overset{(A.6)}{=} \frac{\mathbb{P}(Z_\tau = q)}{\mathbb{P}(Z_\tau \leq Q{+}1) - \mathbb{P}(Z_\tau \leq -(Q{+}1))} \\
&= \underbrace{\frac{1}{2\,\mathbb{P}(Z_\tau \leq Q{+}1) - 1}}_{=:\eta_Q} \mathbb{P}(Z_\tau = q)
\end{aligned}
$$

It follows that α_q is indeed independent of Q because

$$
\alpha_q = \frac{p_{q+1}}{p_q + p_{q+1}} = \frac{\eta_Q\,\mathbb{P}(Z_\tau = q{+}1)}{\eta_Q\,\mathbb{P}(Z_\tau = q) + \eta_Q\,\mathbb{P}(Z_\tau = q{+}1)} = \frac{\mathbb{P}(Z_\tau = q{+}1)}{\mathbb{P}(Z_\tau = q) + \mathbb{P}(Z_\tau = q{+}1)}
$$

■

A.5 Proof of Proposition 2 (FBA Quotes for $Q \to \infty$)

Convergence of FBA markups, and hence of equilibrium supply and demand schedules, follows from the last property of Proposition 1. As excess demand Z_τ is no longer bounded, its Skellam probability distribution is no longer truncated. Finally, the total expected markup flow is finite since

$$
\mathbb{E}[x^f_{\text{FBA}}] = \frac{2}{\tau} \sum_{k=1}^{\infty} k\,\mathbb{P}(Z_\tau = k)\,x_k \overset{(\star)}{\leq} \frac{2\Delta}{\tau} \sum_{k=1}^{\infty} k\,\mathbb{P}(Z_\tau = k) \overset{(\star\star)}{<} \infty
$$

where $(\star)$ uses $x_k < \Delta$ from Proposition 1 and $(\star\star)$ uses the fact that Skellam probabilities tend to zero exponentially by Lemma A.4. ■

A.6 Proof of Lemma 1 (Expected Markup Flow)

(i) First, we have $\frac{\partial}{\partial \lambda_j}\Big(\mathbb{E}[x^f_{\text{CLOB}}]\Big) = \frac{\partial}{\partial \lambda_j}\left(\frac{\lambda_j \lambda_n J}{(\lambda_i+\lambda_j+\lambda_n)}\right) = \frac{\lambda_n(\lambda_i+\lambda_n)J}{(\lambda_i+\lambda_j+\lambda_n)^2} > 0.$

Second, since FBA markups are independent of public news, $\frac{\partial}{\partial \lambda_j}\Big(\mathbb{E}[x^f_{\text{FBA}}]\Big) = 0.$

(ii) First, we have $\mathbb{E}[x^f_{\text{CLOB}}]\Big|_{\lambda_j=0} = \frac{\lambda_j \lambda_n J}{(\lambda_i+\lambda_j+\lambda_n)}\Big|_{\lambda_j=0} = 0.$

Second, by L'Hospital's Rule, we get $\lim_{\lambda_j \to \infty} \mathbb{E}[x^f_{\text{CLOB}}] = \lim_{\lambda_j \to \infty} \frac{\lambda_j \lambda_n J}{(\lambda_i+\lambda_j+\lambda_n)} = \lambda_n J.$

■

A.7 Proof of Theorem 4 (Inefficiency Comparison)

(i) $\lambda_i = 0$ implies $\Delta = J\frac{\lambda_i}{\lambda_i+\lambda_n} = 0$, which leads to $x_q = \Delta \sum_{k=0}^{Q-q} \prod_{s=k}^{Q-q} \alpha_{Q-s} = 0$ and thus $\mathbb{E}[x^f_{\text{FBA}}] = 0$. Due to $\lambda_j, \lambda_n > 0$, we have

$$\mathbb{E}[x^f_{\text{CLOB}}] \;=\; J\frac{\lambda_j \lambda_n}{(\lambda_i+\lambda_j+\lambda_n)} \;>\; 0 \;=\; \mathbb{E}[x^f_{\text{FBA}}]$$

(ii) $\lambda_j = 0$ implies $\mathbb{E}[x^f_{\text{CLOB}}] = J\frac{\lambda_j \lambda_n}{(\lambda_i+\lambda_j+\lambda_n)} = 0$. Since $\lambda_i > 0$, we have $\Delta = J\frac{\lambda_i}{\lambda_i+\lambda_n} > 0$, which leads to $x_q = \Delta \sum_{k=0}^{Q-q} \prod_{s=k}^{Q-q} \alpha_{Q-s} > 0$ and thus $\mathbb{E}[x^f_{\text{FBA}}] > 0$. Overall,

$$0 \;=\; \mathbb{E}[x^f_{\text{CLOB}}] \;<\; \mathbb{E}[x^f_{\text{FBA}}]$$

■

A.8 Proof of Corollary 1 (Markup Flow Boundary)

As shown in Lemma 1, the markup flow under CLOB increases in λ_j, becomes maximal for $\lambda_j \to \infty$ and approaches $\lim_{\lambda_j \to \infty} \mathbb{E}[x^f_{\text{CLOB}}] = \lambda_n J$. On the other hand, the markup flow under FBA are independent of λ_j.

Case 1: If $\mathbb{E}[x^f_{\text{FBA}}] \geq \lambda_n J$, then $\mathbb{E}[x^f_{\text{FBA}}] > \mathbb{E}[x^f_{\text{CLOB}}]$ for all $\lambda_j \geq 0$.

Case 2: If $\mathbb{E}[x^f_{\text{FBA}}] < \lambda_n J$, we can solve for the unique $\lambda^*_j > 0$ that induces equality.

$$\mathbb{E}[x^f_{\text{CLOB}}] \;=\; J\frac{\lambda^*_j \lambda_n}{(\lambda_i + \lambda^*_j + \lambda_n)} \;\stackrel{!}{=}\; \mathbb{E}[x^f_{\text{FBA}}] \qquad \Leftrightarrow \qquad \lambda^*_j \;=\; \frac{\mathbb{E}[x^f_{\text{FBA}}]\left(\lambda_i + \lambda_n\right)}{\lambda_n J - \mathbb{E}[x^f_{\text{FBA}}]}$$

For $\lambda_j < \lambda^*_j$, we have $\mathbb{E}[x^f_{\text{CLOB}}] < \mathbb{E}[x^f_{\text{FBA}}]$, and vice versa. ■

References

Aldrich, E. M. and K. López Vargas (2020): "Experiments in High-Frequency Trading: Comparing Two Market Institutions," *Experimental Economics*, 23, 322–352.

Ausubel, L. M., P. Cramton, M. Pycia, M. Rostek, and M. Weretka (2014): "Demand Reduction and Inefficiency in Multi-Unit Auctions," *Review of Economic Studies*, 81, 1366–1400.

Back, K. and S. Baruch (2007): "Working Orders in Limit Order Markets and Floor Exchanges," *Journal of Finance*, 62, 1589–1621.

Baldauf, M. and J. Mollner (2020): "High-Frequency Trading and Market Performance," *Journal of Finance*, 75, 1495–152.

Baruch, S. and L. R. Glosten (2013): "Fleeting Orders," *Columbia Business School Research Paper 13-43*.

Bellia, M., L. Pelizzon, M. G. Subrahmanyam, J. Uno, and D. Yuferova (2020): "Low-Latency Trading and Price Discovery without Trading: Evidence from the Tokyo Stock Exchange in the Pre-Opening Period and the Opening Batch Auction," *University Ca'Foscari of Venice, Dept. of Economics Research Paper Series No. 9*.

Biais, B., L. R. Glosten, and C. Spatt (2005): "Market Microstructure: A Survey of Microfoundations, Empirical Results and Policy Implications," *Journal of Financial Markets*, 8, 217–264.

Budish, E. B., P. Cramton, and J. J. Shim (2014): "Implementation Details for Frequent Batch Auctions: Slowing Markets Down to the Blink of an Eye," *American Economic Review: Papers & Proceedings*, 104, 418–424.

——— (2015): "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," *Quarterly Journal of Economics*, 130, 1547–1621.

Budish, E. B., R. S. Lee, and J. J. Shim (2020): "A Theory of Stock Exchange Competition and Innovation: Will the Market Fix the Market?" *Working Paper*.

Copeland, T. E. and D. Galai (1983): "Information Effects on the Bid-Ask Spread," *Journal of Finance*, 38, 1457–1469.

Dorre, E. (2020): "Hold on... this doesn't make sense - Frequent Batch Auctions," *Highbury Associates, LLC, Blog*, https://highburyassociates.com/blog/wait-a-0100-second-this-doesnt-make-sense-frequent-batch-auctions-part-i, accessed: Aug 10, 2021.

Du, S. and H. Zhu (2017): "What Is the Optimal Trading Frequency in Financial Markets?" *Review of Economic Studies*, 84, 1606–1651.

Easley, D., N. M. Kiefer, M. O'Hara, and J. B. Paperman (1996): "Liquidity, Information, and Infrequently Traded Stocks," *Journal of Finance*, 51, 1405–1436.

Economides, N. and R. A. Schwartz (1995): "Electronic call market trading," *Journal of Portfolio Management*, 21, 10–18.

Farmer, J. D. and S. Skouras (2012): "Review of the Benefits of a Continuous Market vs. Randomized Stop Auctions and of Alternative Priority Rules (Policy Option 7 and 12)," *UK Government's Foresight Report, The Future of Computer Trading in Financial Markets, Economic Impact Assessment EIA11*.

Foucault, T., O. Kadan, and E. Kandel (2013): "Liquidity Cycles and Make/-Take Fees in Electronic Markets," *Journal of Finance*, 68, 299–341.

Foucault, T., R. Kozhan, and W. W. Tham (2017): "Toxic Arbitrage," *Review of Financial Studies*, 30, 1053–1094.

Foucault, T., A. Röell, and P. Sandås (2003): "Market Making with Costly Monitoring: An Analysis of the SOES Controversy," *Review of Financial Studies*, 16, 345–384.

Fricke, D. and A. Gerig (2018): "Too Fast or Too Slow? Determining the Optimal Speed of Financial Markets," *Quantitative Finance*, 18, 519–532.

Fudenberg, D. and J. Tirole (1991): *Game Theory*, MIT Press Books.

Glosten, L. R. and P. R. Milgrom (1985): "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders," *Journal of Financial Economics*, 14, 71–100.

Haas, M., M. Khapko, and M. Zoican (2020): "Speed and Learning in High-Frequency Auctions," *Journal of Financial Markets*, forthcoming.

Harris, L. E. (2003): *Trading and Exchanges: Market Microstructure for Practitioners*, Oxford University Press.

Ibikunle, G. and Z. Zhang (2021): "Latency Arbitrage and Frequent Batch Auctions," *Working Paper*.

Jagannathan, R. (2020): "On Frequent Batch Auctions for Stocks," *Journal of Financial Econometrics*, 1–17.

Jones, C. M. (2002): "A Century of Stock Market Liquidity and Trading Costs," *Working Paper*.

——— (2013): "What Do We Know About High-Frequency Trading?" *Columbia Business School Research Paper No. 13-11*.

Jovanovic, B. and A. J. Menkveld (2021): "Equilibrium Bid-Price Dispersion," *Journal of Political Economy (forthcoming)*.

Kandel, E., B. Rindi, and L. Bosetti (2012): "The Effect of a Closing Auction on Market Quality and Trading Strategies," *Journal of Financial Intermediation*, 21, 23–49.

Kyle, A. S. (1985): "Continuous Auctions and Insider Trading," *Econometrica*, 53, 1315–1335.

Kyle, A. S. and J. Lee (2017): "Toward a Fully Continuous Exchange," *Oxford Review of Economic Policy*, 33, 650–675.

Lorch, L. (1994): "Monotonicity of the Zeros of a Cross Product of Bessel Functions." *Methods and Applications of Analysis*, 1, 75–80.

Madhavan, A. (1992): "Trading Mechanisms in Securities Markets," *Journal of Finance*, 47, 607–641.

——— (2000): "Market Microstructure: A Survey," *Journal of Financial Markets*, 3, 205–258.

Maskin, E. and J. Tirole (2001): "Markov Perfect Equilibrium I: Observable Actions," *Journal of Economic Theory*, 100, 191–219.

Menkveld, A. J. and M. A. Zoican (2017): "Need for Speed? Exchange Latency and Liquidity," *Review of Financial Studies*, forthcoming.

O'Hara, M. (1995): *Market Microstructure Theory*, Blackwell.

——— (2015): "High Frequency Market Microstructure," *Journal of Financial Economics*, 116, 257–270.

Pagano, M. S., L. Peng, and R. A. Schwartz (2013): "A Call Auction's Impact on Price Formation and Order Routing: Evidence from the Nasdaq Stock Market," *Journal of Financial Markets*, 16, 331–361.

Pagano, M. S. and R. A. Schwartz (2003): "A Closing Call's Impact on Market Quality at Euronext Paris," *Journal of Financial Economics*, 68, 439–484.

Riccò, R. and K. Wang (2020): "Frequent Batch Auctions vs. Continuous Trading: Evidence from Taiwan," *Working Paper*.

Securities, E. and M. Authority (2019): "Final Report: Call for Evidence on Periodic Auctions," *ESMA70-156-1035*.

Treynor, J. L. (1981): "What Does It Take to Win the Trading Game?" *Financial Analysts Journal*, 37, 55–60.

Varian, H. R. (1980): "A Model of Sales," *American Economic Review*, 70, 651–659.

Wah, E., D. R. Hurd, and M. P. Wellman (2016): "Strategic Market Choice: Frequent Call Markets vs. Continuous Double Auctions for Fast and Slow Traders," *EAI Endorsed Transactions on Serious Games*, 3.

Wah, E. and M. P. Wellman (2016): "Latency Arbitrage in Fragmented Markets: A Strategic Agent-Based Analysis," *Algorithmic Finance*, 5, 69–93.

LEIBNIZ INSTITUTE SAFE Sustainable Architecture for Finance in Europe

Recent Issues

No. 343	Alessandro Di Nola, Leo Kaas, Haomin Wang	Rescue Policies for Small Businesses in the COVID-19 Recession
No. 342	Alperen A. Gözlügöl, Wolf-Georg Ringe	Private Companies: The Missing Link on The Path to Net Zero
No. 341	Sandra Eckert	The Limits of Joint-Institutional Frameworks for Sectoral Governance in EU-Swiss Bilateral Relations: Lessons for Future Relations with the UK
No. 340	Anastasia Kotovskaia, Tobias Tröger	National Interests and Supranational Resolution in the European Banking Union
No. 339	Vincent Lindner, Sandra Eckert, Andreas Nölke	Political Science Research on the Reasons for the (non) Adoption and (non) Implementation of EMU Reform Proposals: The State of the Art
No. 338	Elsa Massoc	Fifty Shades of Hatred and Discontent Varieties of Anti-finance Discourses on the European Twitter (France, Germany, Italy, Spain and the UK)
No. 337	Elsa Massoc, Maximilian Lubda	Social Media, Polarization and Democracy: A Multi-Methods Analysis of Polarized Users' Interactions on Reddit's r/WallStreetBets
No. 336	Victor Klockmann, Marie Claire Villeval, Alicia von Schenk	Artificial Intelligence, Ethics, and Diffused Pivotality
No. 335	Victor Klockmann, Marie Claire Villevalc, Alicia von Schenk	Artificial Intelligence, Ethics, and Intergenerational Responsibility
No. 334	Ilya Dergunov, Christoph Meinerding, Christian Schlag	Extreme Inflation and Time-Varying Expected Consumption Growth
No. 333	Vincent R. Lindner	Solidarity without Conditionality. Comparing the EU Covid-19 Safety Nets SURE, Pandemic Crisis Support, and European Guarantee Fund
No. 332	Gyozo Gyöngyösi, Judit Rariga, Emil Verner	The Anatomy of Consumption in a Household Foreign Currency Debt Crisis