(very preliminary and incomplete)

Queuing Uncertainty

Bart Zhou Yueshen*

current version: September 30, 2013
(first version: September 30, 2013)

work in progress
please do not circulate

* Bart Zhou Yueshen is currently visiting MIT Economics and is affiliated to Tinbergen Institute, VU University Amsterdam, and Duisenberg School of Finance. E-mail: b@yueshen.me. Contact address: Department of Economics, Massachusetts Institute of Technology, 77 Massachusetts Avenue E19-750 Cambridge, MA 02142, the U.S.A. This paper benefits tremendously from very helpful discussions with Mark van Achter, Johan Hombert, Bernard Hosman, Vincent van Kervel, Katya Malinova, Albert Menkveld, Sophie Moinas, Ioanid Roşu, Spyros Skouras, Kumar Venkataraman, Adrien Verdelhan, Mao Ye, Haoxiang Zhu, and many others. There are no competing financial interests that might be perceived to influence the analysis, the discussion, and/or the results of this article.

(very preliminary and incomplete)

Queuing Uncertainty

Abstract

In a low-latency trading environment, traders react to a public information event together but they do not know the sequence by which their decisions will take effect. The traders fiercely compete, leading to liquidity oversupply, as such queuing uncertainty weakens the strategic substitution effect of competitors' decisions. "Flickering depth"—new limit orders followed by immediate, massive cancellation—manifests in equilibrium. The phenomena are amplified by the reduction of trading latency. The model speaks to ongoing regulatory debates; e.g. "quote stuffing", quote-to-trade ratio, queue randomization, etc. It is shown that there exists a welfare-maximizing level of queuing uncertainty to which the exchange can steer by carefully randomizing order queues.

JEL code: D40, D43, D47, G10, G18, G19, L10, L13

Keywords: market microstructure, limit order market, low latency trading, flickering quotes, fleeting orders, queue, liquidity, speed, competition

(There are no competing financial interests that might be perceived to influence the analysis, the discussion, and/or the results of this article.)

work in progress
please do not circulate

1 Introduction

Starting from the automation of securities trading in the late 1990s, speed has become the salient feature of modern financial markets. The rise of the machines—technological advancement marked by the upsurge of algorithmic and high-frequency trading—has remarkably reduced trading latency from minutes (Biais, Hillion, and Spatt, 1995) to milliseconds (Hasbrouck and Saar, 2013), with the latest record of picoseconds (Gai, Yao, and Ye, 2013).

As the trading speed approaches the limit imposed by Nature,[1] it is arguable whether market participants can still perfectly condition their decisions on the true market status in real-time. For example, when a trader sees an update of the limit order book, is this the true book status at that moment? Probably not: Among many potential delays, it takes time for the electronic message to travel from the exchange to the trader, and, during this period, the order book might evolve. In addition, even parsing and registering the electronic messages take time.

This paper analyzes such market imperfection formally: In an extremely low-latency trading environment, traders *cannot* perfectly condition on the true market status. The main contributions are three-folded. First, a novel theoretical framework is developed to analyze traders' optimal decisions, accounting for their imperfect conditioning on the market status. Second, the model yields rich testable predictions, echoing evidence from the existing literature, on equilibrium liquidity provision and liquidity dynamics; e.g. (transitory) liquidity "oversupply", "flickering depth", "quote stuffing", heavy electronic message traffic (in absence of fundamental shocks), and "holes" in limit order books. The model identifies a competition channel through which low latency trading positively affects market liquidity. Finally, market quality is examined through a welfare criterion embedded in the model. The generated policy implications anchor recent regulatory debates regarding high-frequency trading; e.g. minimum order resting time, capping order-to-trade ratio, batching orders and randomizing queues, and etc. It is shown that trading efficiency can be improved by adjusting, sometimes *increasing*, the uncertainty in the market.

[1] A photon travels, in vacuum, from New York City to Chicago in about 4 milliseconds, an eternity in a trading environment with latency at picoseconds.

Work in progress
Please do not circulate

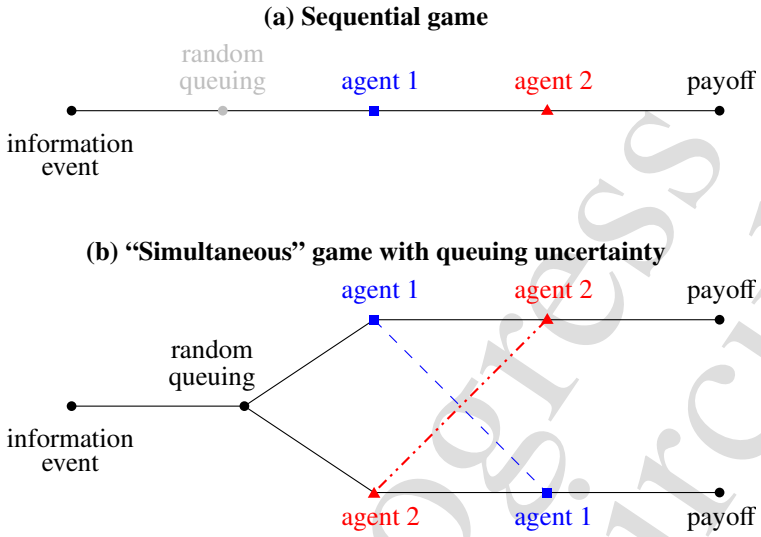


Figure 1: Game structure comparison. This figure illustrates the difference between the conventional view and the new view proposed in this paper of trading in a low-latency environment. In panel (a), agents arrive sequentially and they react to what previous agents have done. Panel (b) adds queuing uncertainty to panel (a). Nature queues the two agents randomly but the two agents do not know which queue position they each get. The agents in panel (b), effectively, play a simultaneous game, though it should be emphasized that the simultaneity does not require, in real-world trading, agents make decisions exactly at the same moment. Instead, it is merely a modeling technique to capture the enriched information set, in which no agents observes others' actions (and hence the market condition cannot be perfectly conditioned on).

The theoretical contribution lies in the novelty of the model's game structure, which differentiates itself from a large body of the existing literature.[2] Figure 1 illustrates the difference. Consider a very short time period right after the occurrence of an information event. There are two agents, labeled 1 and 2. Panel (a) presents the extensive form of the classical treatment of such a scenario. Agent 1 reacts first to the information event (e.g. submitting a new order). Agent 2 then arrives and reacts to agent 1's action. Finally, the payoffs realize. The game is sequential in nature and in particular, each agent can perfectly condition on the latest market status by observing what previous agents have done. Panel (b) contrasts this classical view with "queuing uncertainty", a

[2] See, most relevantly, the limit order market literature led by Glosten (1994), Seppi (1997), Parlour (1998), Foucault (1999), Sandås (2001), Foucault, Kadan, and Kandel (2005), Goettler, Parlour, and Rajan (2005, 2009), Foucault and Menkveld (2008), Roşu (2009), Back and Baruch (2013) and many others for comparison. A comprehensive survey of limit order market models can be found in Parlour and Seppi (2008).

work in progress
please do not circulate

specific form of the market imperfection discussed above. Nature queues the agents *randomly* and the agents do *not* observe their realized queue position (note the information sets, the dashed lines). This way, when making their decisions, neither agents can observe the other's action, and hence, the true market status cannot be perfectly conditioned on.

Model-wise, the game described in panel (b) is simultaneous, though it should be emphasized that the agents are not required to move simultaneously in the real world. In fact, what matters is the sequence in which the agents' decisions *take effect* but not the sequence in which the decisions are made. This is a legitimate concern in reality as time priority is enforced in almost all trading platforms and no two orders can be processed exactly at the same time. Even agents submit orders at the same time (as they observe and react simultaneously to the same information event), the orders arrive at the exchange server randomly due, for example, to idiosyncratic delays in computation, transmission, and etc, forming a stochastic queue.

The model develops from the seminal work by Glosten (1994) by adding the friction of queuing uncertainty to it. Glosten (1994) characterizes a "stable" equilibrium limit order book where at each price level, the marginal (last) limit order breaks even, earning zero profit, in expectation. With queuing uncertainty, however, the equilibrium liquidity provision strictly *exceeds* the prediction of Glosten (1994): Instead of breaking even, the marginal (last) limit order at each price level *loses*, a phenomenon referred to as "liquidity oversupply". This is because, with queuing uncertainty, no trader knows exactly where his order is going to be queued in the book (though the price level can be specified, he has no control over the order's position on that price level). The equilibrium condition becomes that each trader's marginal order breaks even *individually* in expectation, where this expectation is taken *over all possible queue positions*. If the order ends in the top of the queue, it earns positive profit and it loses if in the bottom as execution probability lowers and adverse-selection cost increases along the queue.[3] Because for any queue realizations, there must be one trader whose order ends in the bottom of the queue (and therefore loses), the last limit order

[3] See, for example, Hollifield, Miller, and Sandås (2004), Hasbrouck and Saar (2009), Liu (2009), and, more recently, Raman and Yadav (2013) for empirical evidence.

work in progress
please do not circulate

in the queue must be losing, hence resulting in liquidity oversupply. Note that such oversupply is gauged only among the liquidity providers. Welfare implications, accounting for also liquidity demanders, are discussed below.

Effectively, the simultaneity puts the traders in an oligopolistic competition game, as each trader's order size negatively affects the expected profitability of all others'. The queuing uncertainty, however, weakens the strategic substitution effects and, consequently, amplifies the competition among the traders, leading to liquidity oversupply. To this extent, queuing uncertainty governs the level of competition.

It should be emphasized that such liquidity oversupply is transitory and "unstable" in nature because if the oversupplied orders are not yet executed, some liquidity providers will eventually realize that their orders are queued at the bottom and losing, and therefore will have the oversupplying part canceled. In a dynamic extension, where liquidity providers are allowed to revise their unexecuted orders repeatedly, the model shows that such a revision option *exacerbates* the short-run oversupply: The liquidity providers compete for the profitability in the order book more recklessly by submitting larger orders, as later there will be possibility to cancel (revise) the orders that realize in the bottom of the queue.

The optimal order submission and revision strategies describe the dynamics of the *displayed* book depth and how the exacerbated short-run oversupply gradually dies out in the long-run. To this extent, this paper complements Glosten (1994) by showing how the "stable" equilibrium book depth is achieved over time. The model further reveals that the amount of oversupply in the short-run is amplified by trading latency drops, due to better revision option, and it takes many more rounds of revisions for the book depth to revert to its "stable" level.

Recent empirical studies, in general, find that low-latency trading technology improves market liquidity. See, for example, Hendershott, Jones, and Menkveld (2011) and Hasbrouck and Saar (2013) among many others. The above model prediction, therefore, provides a new explanation to such empirical findings as lowered latency exacerbates short-run liquidity oversupply (compe-

work in progress
please do not circulate

tition channel) and prolongs the revision process (revision option channel). The limit order book, therefore, deepens on average.

Transitory though it is, liquidity oversupply is nevertheless important as they typically occur right after an information event (e.g. a news announcement or a trade potentially with private information) followed by market orders that want to trade on such information. Liquidity demanders (fundamental investors) benefit from the oversupplied liquidity as they can meet their needs at lower cost. This leads to the question how queuing uncertainty affects overall market efficiency. In general, intuitively, welfare should improve with market liquidity, suggesting that *adding queuing uncertainty* to the market might improve efficiency, as the liquidity oversupply will be amplified by the elevated level of competition.

The above welfare concern echoes recent debate on how to regulate trading speed. Larry Harris suggests that "[r]egulatory authorities could require that all exchanges delay the processing of every posting, canceling and taking instruction they receive by a random period of between 0 and 10 milliseconds."[4] Correspondingly, it is reported that EBS, a leading foreign exchange trading platform, recently implemented queue randomization: "[M]essages transmitting orders...will be bundled into batches and then run through a process that randomizes their place in the queue" and the randomization takes "between one and three milliseconds".[5]

The last part of the paper lays a formal theoretical framework to examine the above proposal, esp. how welfare is affected by queue randomization in the order book. It is shown that welfare is not necessarily monotone-increasing in the aggregate liquidity supply: Too much liquidity provision might indulge inefficient risk transfer.[6] By properly randomizing the queues—adjusting the degree of competition among liquidity providers essentially—the exchange can steer the aggregate

[4] Larry Harris. December 27, 2012. "Stop the high-frequency trader arms race", *Financial Times*.

[5] Wanfeng Zhou and Nick Olivari. August 23, 2013. "Exclusive: EBS take new step to rein in high-frequency traders", *Reuters*.

[6] Consider a risk-averse investor facing a sector of risk-neutral market makers. Suppose that the book is very deep and the investor receives a private, *imperfect* signal that the asset value is very high. In this case, the investor buys all the asset offered in the market and yet, from a social planner's point of view, this purely information-driven trade is inefficient because 1) the investor's information rent offsets market makers' adverse-selection cost and 2) risky asset positions are transfered from the less risk-averse market making sector to the risk-averse investor.

work in progress
please do not circulate

liquidity supply toward the socially optimal level. However, such randomization should be carried out with care because 1) the effectiveness of randomizing queues is bounded by the ex ante speed heterogeneity of all traders (and can be very limited); 2) the optimal randomness in queuing is a delicate measure and can be hard to set in real-time trading; and 3) by adding queuing uncertainty, the exchange essentially deprives the market power of fast traders, and in the long-run, it might disincentivize traders' participation.

Queuing uncertainty is similar to the "contact-order uncertainty" in opaque over-the-counter (OTC) markets, as modeled in Zhu (2012). Despite the difference in the market setting (limit order book v.s. OTC market), in both models, liquidity providers do not know their queue positions. In the OTC setting, dealers compete strategically on the quote price basis. In the limit order book setting, the quote prices are given by the book and the market makers compete on the size of their orders. The depth of the order book is, therefore, the focal liquidity measure of the current paper.

The model's prediction on book *depth* dynamics recalls Baruch and Glosten (2013), who focus on the dynamics of *quote prices* in a limit order book with zero tick size, due to which liquidity providers can always undercut each other, resulting in a mixed-strategy equilibrium with "flickering quotes" (see Hasbrouck (2013) for an empirical analysis). Under the mixed-strategy equilibrium of their model, consistent with the current paper, agents' decisions cannot be perfectly conditioned on by one another.[7] Instead of quote prices, the current paper features short-run depth overshoot (i.e. liquidity oversupply) followed by immediate cancellation. Such "flickering depth" prediction agrees with Baruch and Glosten (2013) in that, the so-called "quote-stuffing" behavior by high-frequency traders, rather than gaming the trading system, might in fact be an equilibrium pattern as liquidity providers cancel their oversupplying, losing order, the occurrence of which is nothing but an "unlucky" realization of the orders' queue position. In particular, this paper shows that, given sufficient time for revision, the limit order book converges to the stable equilibrium

[7] The drivers, however, are different: It is the zero tick size assumption in Baruch and Glosten (2013), while in the current paper it is the queuing uncertainty resulting from the non-zero latency at which any trader can register new market updates. The current paper, hence, does not have direct implication on quote prices, while Baruch and Glosten (2013) do not generate predictions on order book depth (rigorously speaking, depth is not properly defined in a zero tick size limit order market).

work in progress
please do not circulate

described in Glosten (1994) (in contrast to Back and Baruch (2013), such convergence does not require infinite market makers).

The key result of this paper, liquidity oversupply, is evidenced in Sandås (2001), who shows that the break-even condition of Glosten (1994) does not hold on average and, instead, the marginal limit order loses. Such losses are attributed to the *negative* order processing cost (see tables 5 and 6 in Sandås, 2001), which is interpreted as a result of traders' heterogeneous valuations for the asset. Queuing uncertainty provides an alternative equilibrium explanation that does not require heterogeneity in traders.

The model prediction that the number of order revisions multiplies with trading latency drops is consistent with Gai, Yao, and Ye (2013) who show that an increase in trading speed in NASDAQ has led to a dramatic increase in the cancellation/execution ratio. A popular exchange fee structure is to impose surcharges on orders exceeding certain order-to-trade ratio thresholds among world's leading exchange markets (e.g. London Stock Exchange, Eurex Germany, and NASDAQ among others). This paper argues that such caps on order-to-trade ratios limit the revision options of liquidity providers and therefore reduce their willingness to supply liquidity in the first place. Friederich and Payne (2012) empirically analyze the effect of capping order-to-trade ratio on Borsa Italiana and find evidence that liquidity reduces, agreeing the model prediction of the current paper.

The current paper also adds to the debate on the impact of trading speed, e.g. high-frequency traders, on market quality. See, among others, Pagnotta and Philippon (2012), Riordan and Storkenmaier (2012), Biais, Foucault, and Moinas (2013), Hoffmann (2013), and Menkveld and Zoican (2013) for related studies. Notably, the current model examines market quality only in terms of the gains from trade (competition on the intensive margin) and does not rely on the technology arm race (competition on the extensive margin), which is potentially socially costly.

The paper is organized as follows. A benchmark model is developed in section 2. Book depth dynamics is analyzed in section 3. Section 4 studies liquidity oversupply deep in the limit order book. Section 5 generates policy implications on market quality. Section 6 then concludes.

work in progress
please do not circulate

Appendix A provides a tractable micro-foundation for queuing uncertainty based on speed heterogeneity. Appendix B summarizes the notations. All proofs and collated in appendix C.

2 Benchmark model

This section presents a fairly simple benchmark model to illustrate the key results of the paper. The focus is on the liquidity provision on one price level *in a very short period*, right after some public information event. Such information event can be macro news, earnings announcement, or simply an update in the order book. More concretely, for example, right after a sell market order hits the best bid price, the updated asset value typically drops and market makers might want to add ask order at the best ask price. The sell market order is the information event and the best ask price is the battlefield analyzed in this section. The *one-price focus* is motivated by the empirical fact that there are few market order that bite deep into the book and most of the order activity concentrates on the top of the limit order book. The dynamic extension is studied in section 3 and the extension to multiple price levels is considered in section 4.

2.1 Model setup

The model is consistent with Glosten (1994). The new element is queuing uncertainty, as explained in details below. This benchmark model only studies the liquidity supply side but leaves the demand side exogenous. Sections 4 endogenizes liquidity demanders.

Agents. There are n risk-neutral agents indexed by $i \in \mathcal{N} := \{1, \ldots, n\}$, where n is a positive integer (the special case of a continuum of agents is considered in section 3). These agents represent the market's liquidity supply side, which is the focus of this section, and they will be referred to as market makers (to be consistent with the literature), though in reality these agents can be any traders who want to use limit orders at such occasions (e.g. endogenous liquidity providers, ELPs, as studied in Anand and Venkataraman, 2013).

work in progress
please do not circulate

Timing. The benchmark model is a one-period simultaneous game. After an information event, each market maker "simultaneously" and independently chooses his limit order size, $q_i \geq 0$.[8] These orders are then queued by Nature according to a known distribution (see "queuing uncertainty" below; appendix A micro-founds such distribution). Finally, a market order arrives (exogenously), trades occur, and market makers consume. The extensive form of the game with $n = 2$ market makers is illustrated in panel (b) of figure 1.

Limit order profitability. The limit orders are perfectly divisible. Let $\dot{\pi}(y)$ be the (expected) *marginal* profit of the y-th limit order at the given price level.[9] Then the (expected) profit of the first limit order of size $y \geq 0$ at the given price is $\pi(y) = \int_0^y \dot{\pi}(x)\mathrm{d}x$. For example, if a limit order of size 10 is added to a price level with existing depth of 5 units, it earns $\pi(10 + 5) - \pi(5)$ in expectation.

Assumption 1 (Quasi-top-of-queue advantage)**.** The (expected) profit $\pi(y)$ is twice-differentiable and is quasi-concave on $y \in [0, \infty)$. Define its first order derivative to be $\dot{\pi}(y)$. The two terminal conditions hold: 1) $\dot{\pi}(0) > 0$ and 2) $\lim_{y\to\infty} \dot{\pi}(y) < 0$. Finally, normalize $\pi(0) = 0$.

For this section and the next, such a $\pi(y)$ is assumed to be exogenously given. In section 4 and later in the welfare analysis, $\pi(y)$ will be endogenized and it will be shown that assumption 1 holds exactly. Some remarks are in order.

Remark 1. Quasi-concavity implies the marginal profit, $\dot{\pi}(y)$, is, loosely speaking, first positive and then negative (but never positive again) as y increases. That is, if one thinks of a limit order book as a set of queues of orders (at various prices), orders on top of each queue earn profit while those in the bottom lose, hence the term "(quasi-)top-of-queue advantage". If $\pi(y)$ is strengthened to be concave, the top-of-queue advantage is exact: the marginal profitability always decreases along the queue.

[8] The simultaneity of order submission is a modeling device to capture that the market makers do not observe each other's action. The model does *not* require market makers to move exactly at the same moment in reality. See panel (b) of figure 1 for the illustration.

[9] This paper adopts Newton's notation, variables with dot(s) overhead, to indicate derivatives.

work in progress
please do not circulate

Remark 2. A quasi-concave $\pi(y)$ is reasonable because orders in the bottom of the queue have lower execution probability and are subject to higher adverse-selection costs (large market orders tend to carry strong signals). See the literature listed in footnote 3 for empirical evidence of decreasing marginal profitability.

Remark 3. The two terminal conditions are assumed to rule out triviality. First, $\dot{\pi}(0) > 0$ guarantees participation: Otherwise, by quasi-concavity, $\dot{\pi}(y) \leq 0$ for all $y \geq 0$ and each limit order at the price always loses. Second, $\dot{\pi}(\infty) < 0$ rules out the uninteresting case where the marginal profit is always strictly positive (all market makers submit infinitely large limit orders).

Queuing uncertainty. Let $\mathcal{K}$ be the collection of all possible queues.[10] Ties, e.g. both market makers 1 and 2 are at the same position in queue, are ruled out. Each queue is represented by an n-by-1 vector $\boldsymbol{k} \in \mathcal{N}^n$, where the i-th value $k_i \in \mathcal{N}$ is the queue position of market maker i. For example, if $n = 3$, a queue $\boldsymbol{k} = [2, 3, 1]^\intercal$ means market maker 1 is the second in queue, market maker 2 the third, and market maker 3 the first. Fix a probability measure that defines queue distribution $\mathbb{P}(\boldsymbol{K} = \boldsymbol{k})$ for all $\boldsymbol{k} \in \mathcal{K}$. The distribution of the random vector $\boldsymbol{K}$ is known by all market makers.[11]

2.2 Equilibrium

Consider market maker i. He chooses his order size q_i, given all others' order sizes, to maximize his expected profit. He cares about the queue position of his order because of the top-of-queue advantage (assumption 1). For a given queue $\boldsymbol{k}$, write the aggregate size of the orders that queue *before* market maker i's order by

$$Q_i^{\not\in}(\boldsymbol{k}) = \sum_{j \in \mathcal{N}} q_j \mathbb{1}_{\{k_j < k_i\}}, \tag{1}$$

[10] $\mathcal{K}$ is a finite set because there are at most $n!$ possible queues.

[11] Note that the specification allows speed heterogeneity: Some market makers can be faster than others. For example, in a two-agent game, market-maker 1 is said to be faster than market-maker 2 if $\mathbb{P}(\boldsymbol{K} = [1, 2]^\intercal) > \mathbb{P}(\boldsymbol{K} = [2, 1]^\intercal)$. Appendix A characterizes such speed heterogeneity in terms of stochastic dominance.

work in progress
please do not circulate

where the superscript "$\notin$" emphasizes that $Q_i^{\notin}$ excludes market maker i's own order. A (pure-strategy) Nash equilibrium is a set $\boldsymbol{q} = \{q_1, \ldots, q_n\}$ such that $\forall i \in \mathcal{N}$, given $\boldsymbol{q}_{-i} = \boldsymbol{q} \backslash \{q_i\}$,

$$q_i \in \arg\max_{q_i} \mathbb{E}\left[\pi\left(Q_i^{\notin}(\boldsymbol{K}) + q_i\right) - \pi\left(Q_i^{\notin}(\boldsymbol{K})\right)\right].$$

The existence of such a Nash equilibrium is established by the following lemma.

Lemma 1 (Existence of Nash equilibrium). *There exists a (pure-strategy) Nash equilibrium $\boldsymbol{q} \in [0, y^{\equiv}]^n$ where $y^{\equiv} > 0$ is the unique solution to $\dot{\pi}(y) = 0$ and at least one q_i is strictly positive.*

Lemma 1 gives an upper bound, $y^{\equiv}$, for each individual's equilibrium limit order size. The intuition is as follows. Note that $y^{\equiv}$ is the break-even point at which the marginal unit of limit order earns zero profit in expectation: $\dot{\pi}(y^{\equiv}) = 0$; the superscript "$\equiv$" emphasizes that the quantity breaks even. Therefore, no market maker will post a limit order exceeding $y^{\equiv}$ units because the part beyond the break-even point always loses (by quasi-concavity of $\pi(\cdot)$), regardless of the order's queue position.

Clearly neither $q_i = 0$ for all $i \in \mathcal{N}$ or $q_i = y^{\equiv}$ for all $i \in \mathcal{N}$ ($n \geq 2$) is an equilibrium. At least one market maker's order size is an interior solution, and the first-order condition holds for him:

$$\mathbb{E}\dot{\pi}\left(Q_i^{\notin}(\boldsymbol{K}) + q_i\right) = 0. \tag{2}$$

This first-order condition implies the following theorem.

Theorem 1 (Liquidity oversupply). *In equilibrium, the market makers oversupply in the sense that the last unit of the limit order earns negative expected profit, $\dot{\pi}\left(\sum_i q_i\right) \leq 0$. The equality holds if and only if there is one market maker who is almost surely the first in queue, i.e. $\exists i \in \mathcal{N}$ such that $\mathbb{P}(K_i = 1) = 1$.*

The intuition of this theorem is sketched in figure 2 with $n = 2$ market makers, whose orders are equally likely to be the first or the second in queue. The downward sloping curve is the marginal profit $\dot{\pi}(\cdot)$. It crosses the horizontal axis at $y^{\equiv}$, which maximizes the profit $\pi(\cdot)$. If there were only one market maker (who almost surely arrives first in queue), this $y^{\equiv}$ would be the optimal quantity

work in progress
please do not circulate

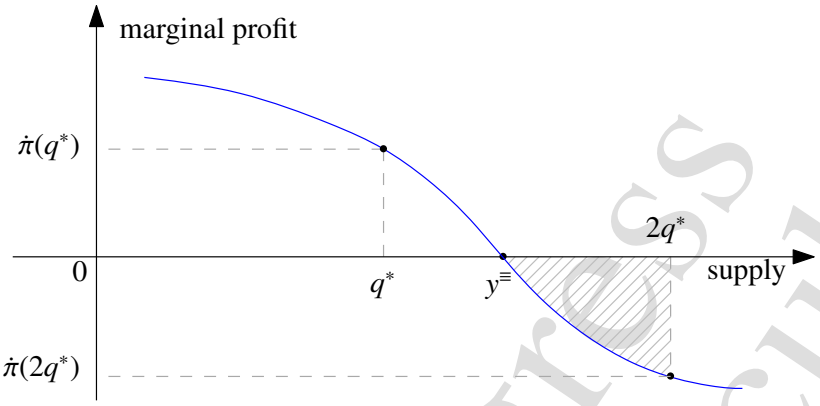


Figure 2: Equilibrium illustration with $n = 2$ market makers. This graphs illustrates the optimal supply decisions with two market makers, each having probability half of being the first in queue. The downward sloping curve is the marginal profit function $\dot{\pi}(q)$. It crosses $\dot{\pi}(q) = 0$ at $q = y^{\equiv}$. In a symmetric-strategy equilibrium, the first-order condition (2) requires that the *expected* marginal profit be zero. That is, $\frac{1}{2}\dot{\pi}(q^*) + \frac{1}{2}\dot{\pi}(2q^*) = 0$, which implies that $\dot{\pi}(q^*) = \dot{\pi}(2q^*)$. The shaded area represents the loss that the last-in-queue market maker suffers.

he would choose (c.f. the break-even condition of Glosten (1994), proposition 2). The two market makers, however, choose a symmetric equilibrium strategy q^* such that each market maker has a *zero expectation*, taken over all possible queue realizations, for his marginal profit. The first-order condition (2) simplifies to $\frac{1}{2}\dot{\pi}(q^*) + \frac{1}{2}\dot{\pi}(2q^*) = 0$ in a symmetric equilibrium. That is, they choose q^* such that the absolute value of $\dot{\pi}(q^*)$ and $\dot{\pi}(2q^*)$ are the same. Clearly, by quasi-concavity of $\pi(\cdot)$, the marginal profit of the very last unit order in the book must be negative. In aggregate, the two market makers thus oversupply.

The shaded area in figure 2 shows the size of the oversupply inefficiency. The key friction that creates such inefficiency is *queuing uncertainty*. If the queue is deterministic (panel (a) of figure 1), then the first market maker in queue will submit an order of size $y^{\equiv}$ and the maximum aggregate profit is achieved. It should be noted that such inefficiency, however, is not necessarily socially costly: The oversupply of liquidity might benefit the demand side, which so far remains exogenous. Welfare analysis is deferred until section 5.

To better understand the intuition of theorem 1, note that the game is reminiscent of oligopolis-

work in progress
please do not circulate

tic competition among producers of a homogeneous good: The supply of one producer (market maker) negatively affects the profit of his competitors. Indeed, the liquidity provision in limit order market can be viewed as strategic substitutes, as defined in Bulow, Geanakoplos, and Klemperer (1985):

Proposition 1 (Strategic substitution). *Suppose $\pi(\cdot)$ is concave on $[0, \sum_i q_i^*]$ where $\{q_i^*\}$ is an equilibrium with $n \geq 2$ market makers. Then the market makers' limit orders are strategic substitutes and, further, the substitution rate is lower than unity in absolute value. Mathematically, $\partial^2 \mathbb{E}\pi(Q_i^{\not q} + q_i)/(\partial q_i \partial q_j) \leq 0$ and $\partial q_i^*/\partial q_j \in (-1, 0]$ for all $i, j \in \{1, 2, \ldots, n\}$ and $i \neq j$.*

Remark. Note that quasi-concavity is strengthened to concavity in the proposition. This is necessary in order to guarantee a monotone decreasing $\dot{\pi}(\cdot)$ on the relevant support, akin to a monotone decreasing demand function.

Strategic substitutes means whenever a market maker i increases his supply, he "threatens" to capture the others' profit (when he queues to the front). The queuing uncertainty, however, bounds the substitution rate. This is because such threats are under-weighed by other market makers as there is always non-zero probability for market maker i's order to queue behind his competitor's, in which case the "threat" is ineffective. As such threats are under-weighed, the market makers engage in fierce competition which results in oversupply of liquidity, i.e. the last unit of the limit order no longer breaks even but loses. The key friction is queuing uncertainty, which leads to the intensified competition.

3 Book depth dynamics

This section builds extensions on the benchmark model to analyze the book depth dynamics over time: The market makers are allowed to revise (cancel and resubmit) their existing (unexecuted) limit orders. As before, the focus remains on the depth level at a single price level in a short time interval immediately after some information event.

work in progress
please do not circulate

3.1 Two-period model

To illustrate the idea, consider the following two-period extension of the benchmark model.

Timing. The model setup follows section 2.1, except the "timing" part. There are two periods, a short-run and a long-run. Nature decides when the market order arrives: With probability $\phi \in (0, 1]$, it arrives in period 0 (the short-run) and with probability $1 - \phi$, it arrives in period 1 (the long-run). Upon arrival, the market order executes against the limit orders in the book and then consumption takes place. The n market makers "simultaneously" submit limit orders in period 0 and these orders form a stochastic queue, $\boldsymbol{K}$, in the book. If the market order does not arrive in period 0, the n market makers in period 1 *observe the queue realization* and are allowed to cancel (part of) existing and/or submit new orders.[12] For simplicity, it is assumed that if the market order arrives in either period, it arrives after all market makers' submission/cancellation have been processed.

Equilibrium analysis. A market maker's decision in the long-run (period 1) is simple: He keeps his order if it is earning (positive) expected profit and cancels it if it is losing. If only part of the order is losing, he cancels that losing part. He never submits new orders because, as will be proved below in corollary 1, in equilibrium, there is always liquidity oversupply in period 0; hence his new orders in period 1 would always lose in expectation.

Now turn to the short-run (period 0). Consider the optimization problem of market maker i, who chooses his period 0 limit order size, q_i, to maximize his expected profit, given all others' limit order size $\boldsymbol{q}_{-i}$:

$$\max_{q_i} \ \phi\mathbb{E}\left[\pi\left(Q_i^{\notin} + q_i\right) - \pi\left(Q_i^{\notin}\right)\right] + (1-\phi)\mathbb{E}\left[\mathbb{1}_{\{Q_i^{\notin}+q_i<y^{\equiv}\}}\left(\pi\left(Q_i^{\notin} + q_i\right) - \pi\left(Q_i^{\notin}\right)\right)\right.$$
$$\left. + \mathbb{1}_{\{Q_i^{\notin}<y^{\equiv}\leq Q_i^{\notin}+q_i\}}\left(\pi\left(y^{\equiv}\right) - \pi\left(Q_i^{\notin} + q_i\right)\right) + \mathbb{1}_{\{Q_i^{\notin}\geq y^{\equiv}\}} \cdot 0\right]$$

[12] Cancellation of part of an existing order is typically allowed in real-world trading. In particular, the order's queue position is not affected. See, for example, *Conditions for Trading at Eurex Deutschland and Eurex Zürich*.

work in progress
please do not circulate

(For notation simplicity, the argument $\boldsymbol{K}$ in $Q_i^{\not\Leftarrow}(\cdot)$ is omitted.) Clearly, if $\phi = 1$, the optimization problem degenerates to that of the benchmark model. When $\phi < 1$, this optimization problem has the extra term led by $(1-\phi)$, reflecting the market maker's expected profit in period 1. Depending on the queue realization, there are three possible sources of expected profit in period 1:

- If the order queues before the break-even point, i.e. if $Q_i^{\not\Leftarrow} + q_i < y^{\equiv}$, the order will always earn strictly positive profit in expectation and the market maker will not revise the order. (Neither will he want to submit new orders because there is liquidity oversupply and any new orders will only lose in expectation.)

- If the order crosses the break-even point, i.e. if $Q_i^{\not\Leftarrow} < y^{\equiv} \leq Q_i^{\not\Leftarrow} + q_i$, the market maker cancels the losing part of the order.

- If the order queues behind the break-even point, i.e. if $Q_i^{\not\Leftarrow} \geq y^{\equiv}$, it is always losing in expectation and the market maker will cancel the entirety of it.

These three possibilities correspond to the three terms in the second square brackets in the above formula.

The optimization problem has the following first-order condition

$$\phi \mathbb{E}\dot{\pi}\left(Q_i^{\not\Leftarrow} + q_i\right) + (1-\phi)\mathbb{E}\left[\mathbb{1}_{\{Q_i^{\not\Leftarrow} + q_i < y^{\equiv}\}}\dot{\pi}\left(Q_i^{\not\Leftarrow} + q_i\right)\right] = 0. \tag{3}$$

Compared to its benchmark version (equation 2), this first-order condition implies $\mathbb{E}\dot{\pi}\left(Q_i^{\not\Leftarrow} + q_i\right) < 0$ because the second term is strictly positive (by quasi-concavity, $\dot{\pi}(y) > 0$ for $0 \leq y < y^{\equiv}$). That is, instead of breaking even, the market maker is willing to incur losses in period 0 thanks to his revision option in period 1. When the revision option shuts off ($\phi = 1$), the first-order condition reduces to the benchmark version.

Corollary 1 (Liquidity oversupply in short-run). *Suppose $n \geq 2$. In equilibrium, there is liquidity oversupply in the short-run (period 0) but the oversupply vanishes in the long-run (period 1).*

work in progress
please do not circulate

The corollary follows theorem 1. Intuitively, as the first-order condition suggests, the option to revise the losing order in period 1 makes the market makers more reckless in submitting their orders in period 0. This way, their competition becomes more fierce, hence leading to larger oversupply.

Without further restrictions on, for example, the shape of $\pi(\cdot)$, comparative statics in such a setting are difficult to derive. The difficulty inherits from the class of competition models where oligopolistic agents play actions that are strategic substitutes to each other.[13] In order to gain additional insights, the following assumption is made and maintained in the rest this section.

Assumption 2. There is a continuum of homogeneous market makers of mass 1.

Remark 1. The continuum of market makers has two implications. First, each market maker is of infinitesimally small market power, and therefore any individual's strategy does not affect the aggregate depth. Second, because there are infinitely many market makers, the liquidity supply business is essentially in perfect competition: The aggregate expected profit, as will be shown, is zero in equilibrium.

Remark 2. The homogeneity in market makers implies a specific distribution for each individual's order queue position: K_i is i.i.d. uniform on $[0,1]$ for all $i \in [0,1]$. Further, it allows the analysis to focus only on symmetric equilibrium.

With assumption 2, the first-order condition (3) becomes

$$\phi\mathbb{E}\dot{\pi}\left(K_i q\right) + (1-\phi)\mathbb{E}\left[\mathbb{1}_{\{K_i q < y^{\equiv}\}}\dot{\pi}\left(K_i q\right)\right] = \phi\mathbb{E}\dot{\pi}\left(K_i q\right) + (1-\phi)\mathbb{P}\left(K_i q < y^{\equiv}\right)\mathbb{E}\left[\dot{\pi}\left(K_i q\right) \mid K_i q < y^{\equiv}\right] = 0$$

where the first equality follows as all market makers (except i) use the same strategy of submitting a limit order of size q in period 0 (symmetric equilibrium). (Because the individual market maker is of zero mass, his order does not affect the aggregate book depth and that is why q_i does not appear

[13] More specifically, the stability of the equilibrium is of great importance (Samuelson, 1941, 1942). Literature on the stability of equilibrium in oligopolistic competition games dates back as early as to Fisher (1961) and Hahn (1962). The comparative statics techniques developed in Milgrom and Shannon (1994) does not work, in a setting with strategic substitutes, if there are more than 2 agents.

work in progress
please do not circulate

in the above first-order condition.) Evaluating the expectations in the above first-order condition (the proof of proposition 2 details the steps) gives the following *zero-profit* condition:

$$\phi\pi(q) + (1-\phi)\pi(y^{\equiv}) = 0. \tag{4}$$

The first term on the left-hand side is the expected aggregate profit if the market order arrives in period 0 and the second term is the expected aggregate profit if the market arrives in period 1 (with losing order canceled). As remark 1 of assumption 2 preludes, this condition implies that, in equilibrium, the expected aggregate profit of all market makers is zero.[14]

Though it might be clear already, the difference between the *zero-profit* condition (4) and the *break-even* condition $\dot{\pi}(q) = 0$ (Glosten, 1994) is worth emphasizing: The zero-profit condition, due to the continuum of market makers, says that the expected profit for any of the market makers is zero, while the break-even condition is in the *marginal* sense that the last unit of the limit order in the book earns zero in expectation. The two conditions are not contradictory.

Proposition 2 (Comparative statics of liquidity oversupply, part I). *In equilibrium, the liquidity oversupply is more pronounced if the market order is less likely to arrive in the short-run. That is, if denote the period 0 equilibrium book depth by $y(0) = 1 \cdot q(0)$, where $q(0)$ is the equilibrium individual order size, then $\partial y(0)/\partial\phi < 0$.*

If the market order is likely to arrive only in the long-run, the market makers will have sufficient time to revise their orders. Thus, when they first submit the limit orders to compete for top-of-queue advantage, each order will be very large as the later, losing orders can be canceled safely.

The proposition speaks to the speed competition in electronic markets. If the fundamental investors, who use market orders to trade their liquidity needs and/or private information, are relatively slow compared to the market makers, the liquidity oversupply spews. This is likely

[14] Note that the zero-profit condition does not necessarily hold. If $-(1-\phi)\pi(y^{\equiv})/\phi < \lim_{y\to\infty}\pi(y)$, the left-hand side is always positive, implying there is infinite rent in market making. In this case, there is no equilibrium book depth. To avoid this technicality, in the rest of this section, it is assumed that $\lim_{y\to\infty}\pi(y) < 0$ so that the zero-profit condition (4) holds.

work in progress
please do not circulate

to be the reality: High-frequency market makers, colocated to the exchange server, can access the market at much lower latency than fundamental investors. If there is no significant speed advantage of market makers over fundamental investors, the market makers submit limit orders more cautiously, resulting in less oversupply. Consistent with this model prediction, Brogaard et al. (2013) find that a system upgrade which make colocated traders even faster than others, on NASDAQ OMX Stockholm improves market depth.

3.2 Multiple revisions

This subsection develops a continuous time model to provide additional insights to the short-run order book depth dynamics. Market makers are allowed to revise their orders multiple times and, further, two additional real-world trading frictions are recognized. First, the exchange has finite processing capacity. It takes a non-zero delay to parse and react to incoming orders. Second, similarly, market makers cannot submit orders continuously in time (even time is continuous as is in the real world). Though, admittedly, the trading speed is ever accelerating, it takes each trader a strictly positive amount of time to update the market condition and evaluate the optimal decisions. Even the travel time of electrons matters. Such gap plays a key role in the order book dynamics.

Setup. The model setup follows the benchmark model, except the time line, modified as the following. Time is continuous and runs from 0 to infinity. There is a gap $\delta > 0$ for each market maker to submit two consecutive orders (either addition or cancellation) to the exchange. That is, all market makers submit new orders only at time $t \in \{0, \delta, 2\delta, \dots\}$. The arrival time of the market order, T, follows exponential distribution with mean $\mathbb{E}T = 1/\lambda$, where $\lambda > 0$ is the arrival intensity. Upon its arrival, the market order executes against the existing limit orders and then the trading game ends. Denote by ϕ the probability of the market order arriving within the h-th interval of δ units of time, i.e. $\phi(\delta) := \mathbb{P}(h\delta \leq T < (h+1)\delta)$ where $h \in \{0, 1, \dots\}$.

Normalize by 1 the mass of orders that the exchange can process in each unit of time; that is, if all market makers submit a new order at the same time, these messages queue up randomly at

work in progress
please do not circulate

the exchange server and it takes exactly 1 unit of time to process all these orders. Such a "message queue" is crucial to the analysis below. The unit of the time is arbitrary; in modern electronic markets, it can be within a few nanoseconds or even picoseconds (Gai, Yao, and Ye, 2013). The book depth is disseminated to all market makers in real-time and all market makers have the same information at each $t \in \{0, \delta, 2\delta, ...\}$.

Both assumptions 1 and 2 are maintained. In addition, for the comparative statics, assumption 1 needs to be slightly strengthened:

Assumption 3 (Top-of-queue advantage, strengthened)**.** Assumption 1 holds and, further, $\pi(y)$ is concave on $[0, y^{\equiv}]$ where $\dot{\pi}(y^{\equiv}) = 0$.

This assumption is not very restrictive. Later in section 4, it will be shown that this strengthened version of top-of-queue advantage is implied by various commonly used frameworks in the literature of limit order market models.

Order book dynamics: Oversupplying followed by massive cancellation. Before proceeding to the formal analysis, it is helpful to heuristically consider a very low-latency environment where δ is small. At $t = 0$, all market makers have incentive to place an order as $\dot{\pi}(0) > 0$. Suppose the equilibrium order size by each market maker at $t = 0$ is $q(0)$, exceeding $y^{\equiv}$ (see corollary 1). This implies a "true" book depth of $q(0)$ (after the exchange processes all the orders), while the displayed book depth at this instance of time is still 0, because the exchange has not processed any of these orders.

Then at $t = \delta$, the *displayed* book depth becomes $\delta q(0)$ (assuming δ very small). Only a fraction of δ of the initial orders have been processed and only these market makers observe their orders' queue position. These market makers do not want to modify their orders because these $\delta q(0)$ units of the orders are queuing in the front, earning positive profit in expectation. The other $(1-\delta)$ market makers, however, still unsure about their orders' position in queue, have incentive to *reduce* their order sizes. This is because, knowing his order has not been processed yet, the probability it ends

work in progress
please do not circulate

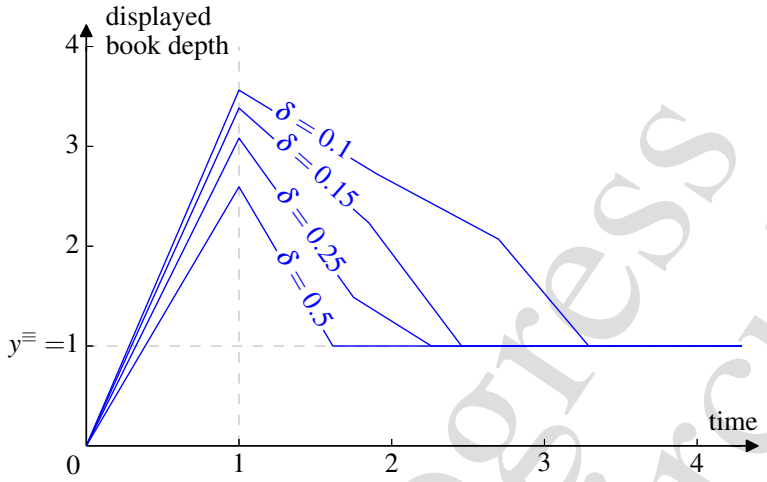


Figure 3: Order book depth dynamics (numerical example). This graph plots the numerically solved order book depth over time based on the numerical example described in section 3.3. Four levels of latencies are plotted: $\delta \in \{0.50, 0.25, 0.15, 0.10\}$. The long-run, stable break-even book depth (as in Glosten, 1994) $y^{\equiv}$ and the market order arriving intensity λ are both normalized to 1.

up in the bottom of the queue increases and if the market maker does nothing, he is more likely to lose in expectation. Thus, the remaining $1-\delta$ market makers will cancel part of their orders to a new equilibrium level, $q(t=\delta)$. The "true" new book depth level will become $\delta q(0) + (1-\delta)q(\delta)$, while the displayed book depth is still $\delta q(0)$.

The iteration goes on. Right after $t = 2\delta$, the "true" new book depth becomes $\delta q(0)+\delta q(\delta)+(1-2\delta)q(2\delta)$; and so forth. The iteration ends upon $t = \bar{h}\delta$ at which time, the queued agents' aggregate "true" order size exceeds the break-even depth level $y^{\equiv}$. At this point of time, any yet-to-be-queued orders are for sure losing in expectation and will be fully canceled.

The resulting book depth dynamics, therefore, exhibits short-run oversupply followed by massive cancellations. Figure 3 illustrates the dynamics under the numerical example of section 3.3. The displayed book depth first increases and exceeds the break-even level, $y^{\equiv}$. The market makers revise their orders as they observe carefully the updates in the meantime, and the book depth gradually reduces and converges to the break-even level. When the market makers' latency δ reduces, the

work in progress
please do not circulate

oversupply exacerbates and it takes more revisions and longer time for the oversupply to die out. This is because with lower latency, the market makers can revise more frequently. The improved accessibility to the market intensifies the competition and it leads to the increased oversupply. The formal analysis continues below.

Equilibrium analysis. Suppose after the initial batch of new orders at $t = 0$, there are $\bar{h}$ (≥ 1) revisions and the order size, right after the h-th revision, is $q(h)$, where $h \in \{0, 1, \ldots, \bar{h}\}$ (if $h = 0$, the "revision" is the first round submission of new orders). Right before the h-th revision, the *displayed* book depth is $h\delta q(0)$, while the "true" book depth is $\delta \sum_{i=0}^{h-1} q(i)$. (This is because the exchange must first process the initial batch of orders, submitted at $t = 0$ with size $q(0)$, before processing the later cancellation orders.) There are $(1 - h\delta)$ mass of market makers who, unsure about their initial orders' queue positions, want to revise. Each of these market makers has an expected profit function of

$$\pi_h(y) = \pi\left(y + \delta \sum_{i=0}^{h-1} q(i)\right) - \pi\left(\delta \sum_{i=0}^{h-1} q(i)\right)$$

for his existing order. $\pi_h(y)$ is simply the residual expected profit function given that the top part of the queue has already been taken. It is a left-down parallel shift of $\pi(y)$ and inherits from $\pi(y)$ a break-even point $y_h^{\equiv} = y^{\equiv} - \delta \sum_{i=0}^{h-1} q(i)$. Further, $\pi_h(y)$ satisfies assumptions 1 and 3.

By symmetry, each of these market makers holds the same belief that his own initial order will be queued at position K_h, a uniformly distributed random variable on $[0, 1 - h\delta]$ for the remaining, unprocessed orders. If $h < \bar{h}$, the zero-profit condition (due to the continuum of market makers) can be written as

$$\phi \mathbb{E}\dot{\pi}_h\left(K_h q(h)\right) + (1 - \phi)\left(\mathbb{P}\left(K_h \leq k_h^*\right)\mathbb{E}\left[\dot{\pi}_h\left(K_h q(h)\right) \middle| K_h < k_h^*\right] + \mathbb{P}\left(K_h > k_h^*\right)\dot{u}_{h+1}\right) = 0,$$

where $\phi = \mathbb{P}\left(h\delta < T < (h+1)\delta\right) = 1 - e^{-\delta\lambda}$ is the probability of the market order arriving before the next revision and $\dot{u}_{h+1}$ is the partial derivative of the continuation value, u_{h+1}, with respect to $q(h)$. The first term on the left-hand side is the expected profit if the market order arrives before

work in progress
please do not circulate

the next revision. The second term has two components, depending on whether the market maker's initial order will be queued to the front or to the rear. The cut-off point k_h^* is defined, for $h < \bar{h}$, as

$$k_h^* = \begin{cases} \delta, & \text{if } \delta \sum_{i=0}^{h} q(h) < y^{\equiv}; \\ \left(y^{\equiv} - \delta \sum_{i=0}^{h-1} q(i)\right) / q(h), & \text{otherwise.} \end{cases}$$

If $h = \bar{h}$, i.e. if this is the last revision, all orders queued behind $y^{\equiv}$ are canceled because it is the last revision if and only if the true depth satisfies $\delta \sum_{i=0}^{h-1} q(i) < y^{\equiv} \leq \delta \sum_{i=0}^{h} q(i)$.

The set of zero-profit conditions can be solved recursively. A significant simplification to the recursion is that $\dot{u}_h = 0$ for all $h \in \{0, 1, ..., \bar{h}\}$, thanks to the perfect competition among the continuum of market makers, who earn zero profit in expectation *at every revision*. Thus, after evaluating the expectations, the zero-profit condition can be simplified to

$$\phi \pi_h \left((1 - h\delta) q(h)\right) + (1 - \phi) \pi_h \left(k_h^* \cdot q(h)\right) = 0 \tag{5}$$

for $0 \leq h < \bar{h}$, a system of difference equation which recursively updates $\pi_h(y)$ and k_h^*. Note that the zero-profit condition (4) derived in the two-period model is a special case, where $\bar{h} = 1$, of the current generalization.

Proposition 2 (Comparative statics of liquidity oversupply, continued). *The oversupply exacerbates 1) if the market order arrival intensity decreases; or 2) if the market makers' latency reduces. That is,* $y(0) = 1 \cdot q(0) > y^{\equiv}$, $\partial y(0)/\partial \lambda < 0$, *and* $\partial y(0)/\partial \delta < 0$.

Note that the first "if" of the above proposition is just a restatement of its two-period version on page 17, as $\phi = \mathbb{P}\left(h\delta < T < (h+1)\delta\right)$ is monotone increasing in the arriving intensity λ. The second "if" is new, and yet, the intuition is the same: As the latency reduces, market makers' option to revise their orders improve—they can access the market more frequently. This improvement escalates the competition among market makers, leading to aggravated oversupply.

work in progress
please do not circulate

Corollary 2 (Number of revisions). *There exists a positive, strictly decreasing series* ($\delta_0 = \infty >$) $\delta_1 > \delta_2 > ... > 0$ *such that in equilibrium, after the submission at* $t = 0$*, it takes* $\bar{h} = h$ (≥ 1) *times of revisions before the book depth reverts to the break-even level of* $y^{\equiv}$*, if and only if* $\delta_h \leq \delta < \delta_{h-1}$.

An implication of corollary 2 is that, fixing the exchange's order processing capacity, the lower is market makers' latency, the longer it takes for the book to "stabilize", reverting to the stable equilibrium as predicted in Glosten (1994).

The limiting case. In the limiting case of $\delta \to 0$, the difference equation system (5) becomes an ordinary differential equation:

$$\pi\left((1-t)q(t) + r(t)\right) - \pi\left(r(t)\right) + \frac{1}{\lambda} q(t)\dot{\pi}\left(r(t)\right) = 0,$$

where $r(t) := \int_0^t q(s)\mathrm{d}s$. The terminal condition at $t = 0$ also holds: $\pi(q(0)) + q(0)\dot{\pi}(0)/\lambda = 0$, from which $q(0)$ can be solved. Corollary 2 implies that as $\delta \to 0$, the total number of revisions, $\bar{h}$, approaches infinite; that is, it takes forever for the book to stablize. This result follows intuitively as, on one hand, the exchange server's capacity is fixed at mass 1 orders per unit of time, while, on the other hand, each market maker operates at zero latency: The message queue at the exchange server becomes infinitely long in an instance of time.

Policy discussion. The above model predicted order book depth dynamics captures some stylized empirical facts, e.g. clustered order submission followed by immediate, massive cancellation. Such behavior is similar, in appearance, to the ill-purposed trading strategy known as "quote stuffing". For example, Egginton, van Ness, and van Ness (2012) refer to quote stuffing as "a practice in which a large number of orders to buy or sell securities are placed and then canceled almost immediately" and similar definitions are found in the media.[15]

[15] Jessica Holzer and Brett Philbin. September 7, 2010. "SEC Is Looking at 'Quote Stuffing"', *The Wall Street Journal*.

work in progress
please do not circulate

The model lends a rather innocuous explanation to the so-called quote stuffing behavior. The oversupply and the following immediate cancellation might simply be an equilibrium outcome due to queuing uncertainty, a market imperfection of low latency trading. Baruch and Glosten (2013)'s model predicts "flickering quotes" in a zero-tick order book with a different mechanism, while the current paper focuses on the book depth dynamics, or "flickering depth".

Along this line of implication, this paper argues that the recent regulation on limiting order-to-trade ratio (also cancellation-to-execution ratio) and on minimum quote life is worth debating. For example, Oslo Børs has imposed a bound on each member's monthly order-to-executed-order ratio of 70, exceeding which results in additional charge. Borsa Italiana has also implemented similar a order-to-trader policy in 2012. London Stock Exchange has imposed and been adjusting a "high usage surcharge" on orders and quotes since 2010.

Such regulations limit the revision opportunity of market makers. In response, the market makers will reduce their order sizes in the short-run, leading to a thinner, less liquid order book. The MiFID II proposal rightfully expresses the concern that "to set out the maximum ratio of unexecuted orders to transactions" should take "into account the liquidity of the financial instrument" (Article 51.7). The U.K. Foresight Projects also feature articles by Brogaard (2011) and by Friederich and Payne (2012) concerning similar negative effects on market quality. In particular, preliminary evidence from Friederich and Payne (2012) finds that book depth in Italian stocks falls after an order-to-trade ratio limit is imposed.

3.3 Numerical example

Consider the following parametrization: $\lambda = 1$, and $\pi(y) = y - y^2/2$ (for $y \geq 0$) which implies $y^{\equiv} = 1$. The quadratic expected profit function can be thought of as a coarse approximation of any other more realistic candidates. For any $\delta \in [0, 1]$, the zero-profit conditions give the equilibrium

work in progress
please do not circulate

order sizes

$$q(0) = \frac{2(e^{\delta} - 1 + \delta)}{e^{\delta} - 1 + \delta^2},$$

and iteratively, for $1 \leq h < \bar{h}$,

$$q(h) = \frac{2(1 - h\delta + \eta\delta)(1 - \delta \sum_{i=1}^{h-1} q(i))}{(1 - h\delta)^2 + \eta\delta^2}$$

where $\eta := 1/(e^{\delta} - 1)$. Note that $q(0) > 1 = y^{\equiv}$, i.e. there is (short-run) liquidity oversupply. The number of revision is $\bar{h} = \lceil y^{\equiv}/(q_0\delta) \rceil$, where $\lceil x \rceil$ is the ceiling integer of a real number x. The terminal condition is always $q(\bar{h}) = 0$ as all unqueued orders will be fully canceled in the last revision. Figure 3 plots the resulting displayed book depth dynamics for four latencies: $\delta \in \{0.5, 0.25, 0.15, 0.1\}$. As the latency drops, the short-run oversupply increases and it takes longer for the book depth to revert to its long-run stable level, $y^{\equiv}$ (see proposition 2 and corollary 1).

4 Deep in the book

This section studies the consequence of queuing uncertainty deep in the book. To do so, the liquidity-demand side is endogenized using the framework developed by Biais, Martimort, and Rochet (2000, 2013). In particular, assumptions 1 and 3 are shown to be implied under the framework (and two alternative specifications, frequently used in the market microstructure literature; see section 4.3). The fully endogenized setup of Biais, Martimort, and Rochet is helpful as it allows welfare analysis later pursued in section 5.

Setup. The market settings are exactly the same as in Biais, Martimort, and Rochet (2000), except that the prices are discrete (see below). There are two types of agents: One investor with CARA utility and $n \geq 1$ risk-neutral market makers, described in section 2.1. The traded asset pays off V units of the numéraire good upon consumption. The payoff V is defined as $V = v_0 + S + \varepsilon$, where v_0 is the unconditional expected payoff, S is the investor's private signal with $\mathbb{E}S = 0$, and

work in progress
please do not circulate

ε is the unobservable innovation, assumed to be normally distributed with mean zero. Denote by ρ (> 0) the product of the investor's constant absolute risk-aversion coefficient and the variance of ε. Timing of the game is as described in section 2.1.[16] The investor observes the signal $S = e$ and also suffers from an endowment shock $E = e$, where both realizations s and e are her private information. Without loss of generality, let $\mathbb{E}E = 0$. The joint distribution of E and S is commonly known. The stochastic queues of orders at each price level, the innovation ε, and the joint distribution of E and S are all independent.

Discrete prices. In order to make sense of queuing uncertainty, prices in the order book must be discrete. Let $\mathcal{P}$ denote the collection of all possible price levels in the limit order book: $\mathcal{P} = \{p_{\min}, ..., p_{\max}\}$, where $-\infty < p_{\min} < v_0 < p_{\max} < \infty$. The book depth is described by $y(p) : \mathcal{P} \mapsto \mathbb{R}$, which gives the (signed) depth at price level p. As the analysis is symmetric, this paper only considers the ask side. Denote the j-th ask price by a_j ($j \in \{1, 2, ...\}$) which is the j-th smallest price in $\{p \mid p > v_0, p \in \mathcal{P}, y(p) > 0\}$.[17] Let the depth at the j-th best price be denoted by y_j and define $a_0 = v_0$ with $y_0 = 0$. (It will be justified that the ask side indeed corresponds to the prices $\{p \mid p > v_0, p \in \mathcal{P}\}$ in equilibrium.) Finally, denote the cumulative depth up to and including (excluding) the j-th level by $y_{\leq j} := \sum_{i \leq j} y_i$ (and $y_{<j} := \sum_{i<j} y_i$, respectively).

4.1 The investor's order size

Consider the investor's optimization problem given a book described by $y(p)$. Suppose the investor wants to buy the asset offered at the j-th ask price, with price a_j and depth y_j (> 0). Note that, the optimization is iterative: Due to price priority, the investor can only buy the offers at the j-th ask price if she has already bought everything offered below the j-th price. That is, she must have

[16] To focus on the short-run order submission and to keep the model analytically tractable, unlike section 3, the game is only one-period as in Biais, Martimort, and Rochet (2000). The dynamic nature, however, arises as market makers optimize their order sizes at different price levels. Equilibrium of a fully dynamic model with both time and price dimension concerns can be numerically solved by using the method developed by Goettler, Parlour, and Rajan (2005, 2009).

[17] Perhaps surprisingly, under queuing uncertainty, it is not necessary that the equilibrium ask prices are consecutive; i.e. it is possible that $a_{j+1} - a_j > \rho$. See proposition 3 for details.

work in progress
please do not circulate

bought $y_{<j}$ units of the asset, accumulating her endowment to $e + y_{<j}$ units, before she could choose her optimal demand at price a_j. Then, conditional on her private information of $S = s$ and $E = e$, the investor optimizes:

$$\max_{0 \le x_j \le y_j} \mathbb{E}\left[-\exp\left\{ -\left[\left(E + y_{<j}\right) V + \left(V - a_j\right) x_j \right] \right\} \mid E = e, S = s \right]$$

which is equivalent to maximizing her certainty equivalent:

$$\max_{0 \le x_j \le y_j} m_j + \left(v_0 + \theta - a_j\right) x_j - \rho y_{<j} x_j - \frac{\rho}{2} x_j^2$$

where $m_j := (s + v_0)(e + y_{<j}) - \rho(e + y_{<j})^2/2$ is the certainty equivalent for her endowment (after buying everything offered below a_j) and $\theta := s - \rho e$ is her *endowment-risk-adjusted* signal for the asset. (Note that the transfer, $\sum_{i<j} y_i a_i$, that the investor has already paid does not come into the optimization.) With the lower and the upper bound of $0 \le x_j \le y_j$, the potentially cornered solution is

$$x_j(\theta) = \min\left\{ y_j, \max\left\{ 0, \frac{1}{\rho}\left(v_0 + \theta - a_j\right) - y_{<j} \right\} \right\}. \tag{6}$$

It can be seen that the investor's optimal order size truthfully reveals θ, an imperfect signal (due to the liquidity/endowment shock; see also Vayanos and Wang, 2012), up to the lower and the upper bounds, which give the following "hurdle" that does not exist if the price is continuous:

Lemma 2 (Hurdle in signal). *After the book depth at the j-th best price depletes, the investor further bites to the $(j+1)$-th price level if and only if her (endowment-risk-adjusted) private signal exceeds a strictly positive hurdle.*

Mathematically, for each ask price a_j, there exist a floor $\lfloor \theta_j \rfloor = (a_j - v_0) + \rho y_{<j}$ and a ceiling $\lceil \theta_j \rceil = (a_j - v_0) + \rho y_{\le j}$ such that $x_j(\theta)$ is off-corner if and only if $\lfloor \theta_j \rfloor < \theta \le \lceil \theta_j \rceil$. The hurdle between two consecutive price levels is $\lfloor \theta_{j+1} \rfloor - \lceil \theta_j \rceil = a_{j+1} - a_j > 0$.

Figure 4 illustrates the investor's optimal cumulative order size as a function of her endowment-risk-adjusted signal θ and the book $\{y_j\}$. This cumulative order size is essentially her demand

work in progress
please do not circulate

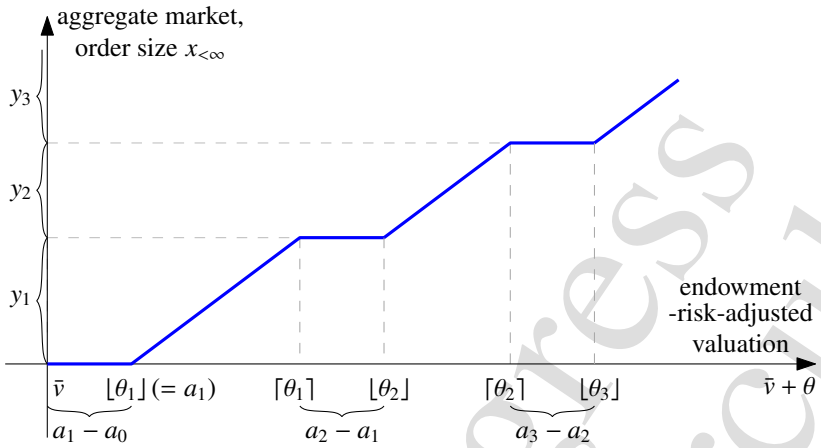


Figure 4: The investor's optimal market order size. The investor's cumulative order size, $x_{\leq\infty}$, is plotted against her endowment-risk-adjusted signal, θ, given a limit order book with depths y_j at the j-th best ask price. The curve is piece-wise increasing in θ. The investor bites into price level a_j only if θ exceeds a floor threshold $\lfloor\theta_j\rfloor$ and depletes that level only if θ exceeds a ceiling $\lceil\theta_j\rceil$ as given in lemma 2. The hurdle size is $a_{j+1} - a_j$.

for the asset (given the book depth), which is piece-wise linearly increasing with slope $1/\rho$ (the risk-adjustment).

4.2 Depths in the order book

Consider the j-th ask price level at $a_j = p > v_0$, with the previous cumulative depth $y_{<j}$. Market makers do not know the investor's (endowment-risk-adjusted) signal, Θ, but they know $\Theta := S - \rho E$, hence also its distribution (the joint density of S and E are common knowledge). Suppose the depth at this price a_j is $y \geq 0$. Then the expected profitability for these y units of the limit order is

$$\pi_j(y) = \int_{\lfloor\theta_j\rfloor}^{\lfloor\theta_j\rfloor+\rho y} \Big(a_j - v(\theta)\Big)\, x_j(\theta) f(\theta) \mathrm{d}\theta + \int_{\lfloor\theta_j\rfloor+\rho y}^{\infty} \Big(a_j - v(\theta)\Big)\, y f(\theta) \mathrm{d}\theta \tag{7}$$

where the signal floor $\lfloor\theta_j\rfloor$ is defined in lemma 2, $x_j(\theta)$ is the investor's market order size at this price level (equation 6), $v(\theta) := \mathbb{E}[V \mid \Theta = \theta]$ is the conditional expectation of the asset payoff, and $f(\theta)$ is the density of Θ. Following Biais, Martimort, and Rochet (2000), two regularity conditions

work in progress
please do not circulate

are needed. First, the joint distribution of S and E satisfies that $v(\theta)$ is weakly increasing in θ, i.e. $\dot{v}(\theta) \geq 0$. Second, the support of Θ is continuous.

The marginal expected profit is

$$\dot{\pi}_j(y) = \int_{\lfloor\theta_j\rfloor+\rho y}^{\infty} \Big(a_j - v(\theta)\Big) f(\theta)\mathrm{d}\theta = \mathbb{P}\Big(\Theta > \lfloor\theta_j\rfloor + y\Big)\Big(a_j - \mathbb{E}\Big[v(\Theta)\,\Big|\,\Theta > \lfloor\theta_j\rfloor + y\Big]\Big). \tag{8}$$

The second equality gives the intuitive interpretation of the expected profit of the marginal order: Conditional on its execution (with probability $\mathbb{P}\Big(\Theta > \lfloor\theta_j\rfloor + y\Big)$), the sales revenue is the ask price a_j and the adverse-selection cost is $\mathbb{E}\Big[v(\Theta)\,\Big|\,\Theta > \lfloor\theta_j\rfloor + y\Big]$. The following proposition shows that "top-of-queue advantage" holds (and, in fact, is further strengthened).

Proposition 3 (Top-of-queue advantage). *There is top-of-queue advantage in the limit orders at each price level. In particular, both assumptions 1 and 3 hold. Further, $\dot{\pi}_j(y)$ is quasi-convex in $y \geq 0$.*

Now turn to the market makers. Each market maker submits a set of limit orders, one order for each price level, to maximize his expected profit. Because the cumulative depth at any price level affects the profitability of orders deep in the book, the optimization is dynamic (along the price dimension) in nature. Consider market maker i's order size at price level j, given all other market makers' order sizes. Denote by $u_{i,j}$ his continuation value for price level j and the Bellman equation can be written as

$$u_{i,j} = \max_{q_{i,j}} \ \mathbb{E}\Big[\pi_j\Big(Q^{\not\in}_{i,j} + q_{i,j}\Big) - \pi_j\Big(Q^{\not\in}_{i,j}\Big)\Big] + u_{i,j+1}, \tag{9}$$

where for notation simplicity, the arguments of $u_{i,j+1}$ are omitted but it should be understood that it depends on market maker i's order size $q_{i,j}$ at price level j. With some manipulation, it can be shown that the first-order condition can be written as (see the proof of proposition 4 in appendix)

$$\mathbb{E}\dot{\pi}_j\Big(Q^{\not\in}_{i,j} + q_{i,j}\Big) - \mathbb{E}\dot{\pi}_{j+1}\Big(Q^{\not\in}_{i,j+1}\Big) = 0. \tag{10}$$

Compared with the first-order condition in the benchmark model (equation 2), there appears a

work in progress
please do not circulate

new term, $-\mathbb{E}\dot{\pi}_{j+1}\left(Q^{\notin}_{i,j+1}\right)$, in the equation. This new term measures the expected profit of the first marginal unit of the order on price level $j+1$. It is negative because if market maker i increases his order size at level j, the profitability at level $j+1$ reduces (lower execution probability and higher adverse-selection cost). The interpretation of this first-order condition is, instead of myopically choosing the order size to break even the expected profit at the current price level, market maker i also accounts for his expected profitability on the next price level. This effect, however, is not large enough to curb the fierce competition among market makers and the aggregate liquidity at each price level is still oversupplied, as proved in the following proposition.

Proposition 4 (Oversupply deep in the book). *Suppose the investors' endowment-risk-adjusted signal Θ exhibits (weakly) increasing hazard rate. Then, in equilibrium, there is liquidity oversupply at each price level. Mathematically, $\dot{\pi}_j(\sum_i q_{i,j}) \leq 0$ for all j and the equality holds if and only if there is one market maker who is almost surely the first in queue at price level j.*

Remark. The technical condition that Θ exhibits (weakly) increasing hazard rate, i.e. $f(\theta)/(1-F(\theta))$ is weakly increasing in θ, is a sufficient (but not necessary) condition for the oversupply result. As discussed in Biais, Martimort, and Rochet (2000), who also assume such a condition (equations 16 and 17), this is *not* a restrictive assumption. Some commonly used distributions (to name a few, uniform, normal, and exponential) all satisfy increasing hazard rate. The hazard rate essentially reflects the severity of market makers' adverse-selection problem. Therefore, the increasing hazard rate assumption can also be understood, intuitively, as a condition to ensure the regularity of the problem; see Baruch and Glosten (2013) for more details on the importance of sufficient adverse-selection in the limit order market.

The oversupply results suggests that instead of break-even, the last unit limit order at each price level loses in expectation, at least in very short run. This gives an equilibrium prediction about "holes" in the limit order book (in short-run):

work in progress
please do not circulate

Corollary 3 (Holes in book). *For sufficiently small tick size, the equilibrium limit order book has, possibly interior, "holes", or empty price levels on which no market makers are willing to post any orders.*

To understand the intuition of this result, it is helpful to look at again the expression of the expected marginal profit (equation 10). When the price level goes from j to $j+1$, the (expected) marginal revenue jumps from a_j to a_{j+1} and, yet, the expected marginal cost, which only depends on the cumulative book depth, does not change. Liquidity oversupply implies that the expected marginal cost of the last limit order at price level j exceeds a_j. Hence, if the next price level $a_j + \rho$ (where ρ is the tick size) is not large enough, the marginal profit at price $a_j + \rho$ will never be positive and no market maker will place order there.

Empirical evidence of holes in limit order book has been found by, for example, Biais, Hillion, and Spatt (1995). Queuing uncertainty gives an equilibrium explanation. In particular, equilibrium holes are not supported by the equilibrium of Glosten (1994), whose break-even condition asserts that the last unit of limit orders at each price level has expected revenue equating expected cost. This is because, as the price level moves to the next tick, the expected revenue jumps up but the expected cost does not change, implying strictly positive expected marginal profit for the first unit of limit orders at the new price. Market makers will compete for such profit by submitting orders to the new level, and there will be no holes.

It should be emphasized that the predicted holes in the book are *transitory* in nature. If the market makers are given sufficient time to revise/cancel their orders, the long-run equilibrium reverts to the break-even level described by Glosten (1994) and the holes will disappear; see section 3 for the book depth dynamics.

A second corollary follows proposition 4. It compares the cumulative liquidity in the book under queuing uncertainty and under the break-even condition of Glosten (1994).

work in progress
please do not circulate

Corollary 4 (Cumulative liquidity oversupply). *Denote by $y^*_{\leq j}$ and $y^{\equiv}_{\leq j}$, respectively, the equilibrium cumulative depths until (inclusive) the j-th price level under queuing uncertainty and under the break-even condition (Glosten, 1994). If there are no holes in the book under queuing uncertainty, then $y^*_{\leq j} \geq y^{\equiv}_{\leq j}$ for all price levels.*

This corollary is not a restatement of proposition 4 because the proposition focuses on the depth at each price level while the corollary addresses the *cumulative* depth. In particular, the oversupply at price level j affects the break-even quantity at price level $j+1$.

4.3 Alternative setups

So far this section has been building on the fully-endogenized framework of Biais, Martimort, and Rochet (2000). This subsection gives two other examples, following Back and Baruch (2013), to show that the top-of-queue advantage (assumptions 1 and 3) holds generically.

Example 1 (Informed or liquidity dichotomy)**.** This example adopts the classical dichotomy that treats the incoming market order to be driven by either private information or liquidity shock. Such dichotomy is seen in, for example, Glosten and Milgrom (1985), Kyle (1985), Easley and O'Hara (1992), and many others.

Suppose that with probability ψ the arriving investor observes the true value of the asset, v, while with probability $1-\psi$ the arriving investor is purely liquidity driven. If the investor is information driven, she buys all units of the asset offered at prices $p \leq v$ (again, the example only deals with the ask side). If the investor is liquidity driven, she submits a market order of size X, which has density $f(x)$ defined on $\mathbb{R}$. Then market makers' expected profit of a limit order of size y at price level a_j is

$$\pi_j(y) = \psi \cdot \left(a_j - \mathbb{E}\left[V \,\middle|\, V \geq a_j\right]\right) y + (1-\psi)\left(a_j - v_0\right)\left[\int_{y_{<j}}^{y_{<j}+y} \left(x - y_{<j}\right) f(x)\mathrm{d}x + \int_{y_{<j}+y}^{\infty} y f(x)\mathrm{d}x\right]$$

where, following the notations developed in this section, $y_{<j}$ is the cumulative depth excluding the j-th price level. It is easy to evaluate that $\ddot{\pi}_j(y) = -(1-\phi)\left(a_j - v_0\right) f\left(y_{<j} + y\right) < 0$; that is, $\pi_j(y)$

work in progress
please do not circulate

is strictly concave at any price level $a_j > v_0$. Therefore, assumption 1 holds under this setup, with the quasi-concavity strengthened to concavity, and hence follows assumption 3.

Example 2 (Inelastic demand)**.** In this example, the incoming market order size follows an exogenous distribution that is independent of the order book depth. Such models are seen in, for example, Sandås (2001), Foucault and Menkveld (2008), and Kervel (2013).

Denote the market order size by X, a random variable with density function $f(x)$. Denote the price impact by $v(x) := \mathbb{E}[V \mid X = x]$ and assume $\dot{v}(x) \geq 0$. Then given the cumulative depth $y_{<j}$, market makers' expected profit of a limit order of size y at price a_j is

$$\pi_j(y) = \int_{y_{<j}}^{y_{<j}+y} \Big(a_j - v(x)\Big)\Big(x - y_{<j}\Big) f(x)\mathrm{d}x + \int_{y_{<j}+y}^{\infty} \Big(a_j - v(x)\Big) y f(x)\mathrm{d}x.$$

The expression is akin to equation (7), the fully-endogenized version. Indeed, following the proof of proposition 3, it can be shown that $\pi_j(y)$ is quasi-concave and, further, $\dot{\pi}_j(y) > 0$ implies $\ddot{\pi}_j(y) < 0$ as in proposition 3, satisfying assumptions 1 and 3.

The other results derived in section 4.2 (proposition 4 and its two corollaries) are still valid under the setups of these two examples. The derivation and proofs are omitted for brevity.

5 Welfare

The queuing uncertainty escalates the level of competition among market makers and hence pauperizes them. However, such oversupply benefits the liquidity demand side as the fundamental investors can trade to meet their liquidity needs at lower costs. This section uses the fully endogenized framework of section 4 to evaluate the overall effect of queuing uncertainty on social welfare. The analysis clarifies 1) how welfare depends on the liquidity in a limit order market and 2) under what conditions can the exchange improve welfare by randomizing the order queues.

Setup. The framework is the same as section 4 with some additional structures to keep the exposition clear. To define "liquidity" precisely, let there be only one possible (ask) price: $\{p|p >$

work in progress
please do not circulate

$v_0, p \in \mathcal{P}\} = \{a\}$. This way, the liquidity measure can be summarized in one scalar variable, the book depth. (With multiple prices, a complex "liquidity" measure, covering variations in both price and depth, is needed.) The one-price focus can be motivated by the fact that in reality, there are few market orders that are large enough to bite into the book.

5.1 Liquidity and welfare

Define welfare as the sum of the (ex ante) expected gains from trade of all participants, the one investor and n market makers, measured by their respective ex ante certainty equivalents. Suppose the book depth is y. (Since there is only one price, subscripts indicating price levels are dropped for clarity.) The investor's certainty equivalent is $\mathrm{ce}(y) = \mathbb{E}[M + (v_0 + \Theta - a)\, x(\Theta; y) - \rho x(\Theta; y)^2/2]$, where $M := E \cdot (S + v_0) - \rho E^2/2$ is the certainty equivalent of the investor's endowment, $\Theta := S - \rho E$ is the endowment-risk-adjusted signal, and $x(\Theta; y)$ is the investor's optimal market order size (which might be capped; see equation 6). The expected gains from trade is

$$\mathrm{ce}(y) - \mathbb{E}M = \mathbb{E}\left[(\Theta - \lfloor\theta\rfloor)\, x(\Theta; y) - \frac{\rho}{2} x(\Theta; y)^2\right],$$

where $\lfloor\theta\rfloor := a - v_0$ is the floor which Θ must exceed so that the investor will trade (see lemma 2).

The market makers are risk-neutral and their aggregate expected certainty equivalent is just the sum of their expected profits.

Lemma 3 (Market makers' aggregate profit). *The aggregate profit of all market makers is queue-position irrelevant. Only the aggregate book depth matters:* $\forall \boldsymbol{k} \in \mathcal{K}$, $\sum_i \pi\left(Q_i^{\notin}(\boldsymbol{k}) + q_i\right) = \pi\left(\sum_i q_i\right)$.

With lemma 3, the market makers' aggregate expected gains from trade, fixing a book depth y, can be computed as (c.f. equation 7)

$$\pi(y) = \mathbb{E}\left[(\lfloor\theta\rfloor + v_0 - v(\Theta))\, x(\Theta; y)\right].$$

work in progress
please do not circulate

Therefore, the social welfare, $(\text{ce}(y) - \mathbb{E}M) + \pi(y)$, is a function of the book depth, i.e. the liquidity level in this market:

$$w(y) = \mathbb{E}\left[(v_0 + \Theta)\, x(\Theta; y) - \frac{\rho}{2} x(\Theta; y)^2 - v(\Theta) x(\Theta; y)\right]. \tag{11}$$

Note that the transfer, $\lfloor\theta\rfloor x(\Theta; y)$, from the market makers (adverse selection cost) to the investor (information rent) offsets in the aggregation. The remaining terms have intuitive interpretations. The first term is the investor's expected valuation (after adjusting her endowment risk) of the position she buys from the market makers. The second term corrects her acquired risk. The last term is the value of the position seen from market makers' perspective (recall that $v(\theta) := \mathbb{E}[V \mid \Theta = \theta]$).

Lemma 4 (Shape of welfare as a function of liquidity). *For all $y \geq 0$, $w(y) > \pi(y)$, $\dot{w}(y) > \dot{\pi}(y)$, and $\ddot{w}(y) < \ddot{\pi}(y)$. Further, $w(y)$ is concavely monotone increasing on $y \in [0, y^{\equiv}]$.*

Lemma 4 only says that welfare is initially increasing. There might be times when too much liquidity hurts social welfare (if $y^* < \infty$). To see this, write the derivative of $w(y)$ as

$$\dot{w}(y) = \mathbb{P}(\Theta > \lfloor\theta\rfloor + \rho y)\left(\mathbb{E}[v_0 + \Theta - \rho y \mid \Theta > \lfloor\theta\rfloor + \rho y] - \mathbb{E}[v(\Theta) \mid \Theta > \lfloor\theta\rfloor + \rho y]\right).$$

The difference between the two conditional expectations is the wedge between the investor's and market makers' valuation of the marginal unit of the asset. The difference is not necessarily positive as, from the social planner's perspective, the investor might buy "too aggressively". This is because though the investor enjoys her information rent, the social planner does not, as such rent offsets with the market makers' adverse-selection cost. Consider the extreme case where the investor is very likely to draw a positive signal but no endowment shock. In this case, large liquidity provision in the market is socially suboptimal because it "indulges" the risk-averse agent to buy risky asset from the risk-neutral market makers. Figure 5 shows an initially increasing welfare function eventually decreases if there is too much liquidity. The parametrization for the numerical procedure is described in section 5.3.

work in progress
please do not circulate

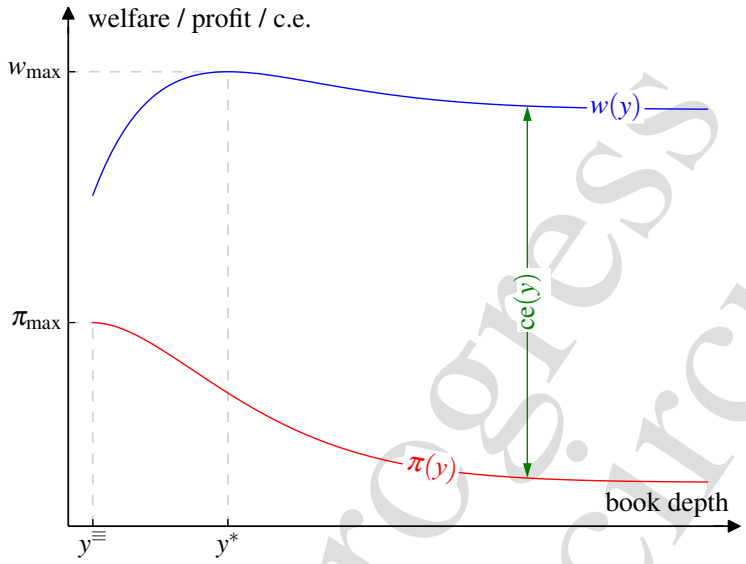


Figure 5: Welfare as a function of liquidity supply. This figure numerically illustrates the shape of welfare as a function of aggregate liquidity supply (i.e. aggregate book depth). The parameters are as described in section 5.3. The vertical axis shows welfare, market makers' expected profit, and the investor's certainty equivalent. The horizontal axis shows the book depth, y. The upper curve (blue) is welfare, $w(y)$. The lower curve (red) is market makers' aggregate expected profit, $\pi(y)$. The difference, indicated by the (green) vertical line, is the investor's certainty equivalent, ce(y).

Proposition 5 (Maximum social welfare). *Social welfare maximizes at y^*, where y^* is strictly larger than $y^{\equiv}$ (market makers' break-even depth) and is possibly finite.*

To see this, compare the cost-benefit analyses of a social planner and of a market maker (using equation 8):

$$\begin{aligned}\dot{w}(y) - \dot{\pi}(y) &= \int_{\lfloor\theta\rfloor+\rho y}^{\infty} \left[(v_0 + \theta - \rho y) - a\right] f(\theta)\mathrm{d}\theta \\ &= \mathbb{P}\left(\Theta > \lfloor\theta\rfloor + \rho y\right)\left(\mathbb{E}\left[v_0 + \Theta - \rho y \,|\, \Theta > \lfloor\theta\rfloor + \rho y\right] - a\right) > 0.\end{aligned}$$

The difference is exactly the investor's expected benefit of buying a marginal unit of the asset. (If she could buy the marginal unit, she gains the conditional expectation term and she pays the ask

work in progress
please do not circulate

price a.) That is, compared to a representative market maker who chooses the book depth to maximize the aggregate expected profit, a social planner in addition takes into account the investor's additional gain. The socially optimal depth, therefore, always exceeds the market making optimal depth. The result is analogous to the scenario of a monopolist producer who, facing an exogenous demand function, chooses the production quantity to maximize his profit. There is always some consumer surplus cannot be realized because of the monopolist under-supplies in order to sell at a high price.

Recall that with queuing uncertainty, the order book deepens (liquidity oversupply) and hence there is room for welfare improvement by potentially *adding* queuing uncertainty to the market. This leads to the following subsection, which introduces queue randomization and analyzes how welfare, i.e. aggregate expected gains from trade, is affected.

5.2 Policy implication

In this subsection, consider $n = 2$ (representative) market makers, a fast one and a slow one. The two-agent assumption is an abstraction to reflect speed heterogeneity in the real-world trading environment, where some participants, having invested heavily in technology, have relatively low latency accessing the market compared to others.

Label the two market makers 1 and 2, and there are two possible queues of the limit orders: $\boldsymbol{k} \in \mathcal{K} = \{[1,2]^\intercal, [2,1]^\intercal\}$. Denote by $\alpha := \mathbb{P}(\boldsymbol{K} = [1,2]^\intercal)$ the probability of market maker 1's order queuing in the front of market maker 2's. Without loss of generality, let market maker 1 be the (weakly) faster one: $\alpha \in [1/2, 1)$. In equilibrium, the limit order sizes, q_1 and q_2, should satisfy the following first-order condition system (c.f. equation 2):

$$\begin{cases} \alpha\dot{\pi}(q_1) + (1-\alpha)\dot{\pi}(q_2+q_1) = 0, & \text{for market maker 1;} \\ (1-\alpha)\dot{\pi}(q_2) + \alpha\dot{\pi}(q_1+q_2) = 0, & \text{for market maker 2.} \end{cases} \tag{12}$$

work in progress
please do not circulate

The following technical condition is needed for comparative statics:[18]

$$\ddot{\pi}(2y^{\equiv}) \leq 0. \tag{13}$$

Then a connection between the aggregate book depth and speed heterogeneity can be established:

Proposition 6 (Speed heterogeneity and liquidity oversupply). *The liquidity oversupply is more pronounced when the market makers have more heterogeneous speed.*

Mathematically, if denote the equilibrium aggregate book depth by y, then $y(\alpha)$ has a local maximum at $\alpha = 1/2$.

Corollary 5. *If $\dot{\pi}(y)$ is "regular", the equilibrium aggregate book depth y is monotone increasing (decreasing) on $\alpha \in (0, 1/2)$ (on $\alpha \in (1/2, 1)$, respectively) and has its unique maximum at $\alpha = 1/2$.*

Remark. A "regular" $\dot{\pi}(\cdot)$ is important for the purpose of comparative statics. This is because the market makers' problem, as alluded in proposition 1, resembles firms' competition game and inherits the difficulty of signing comparative statics from such oligopolistic competition games where agents' decisions are strategic substitutes of each other. A convenient (sufficient) condition is that $\dot{\pi}(\cdot)$ is concavely decreasing (c.f. a concavely decreasing demand function as in Kreps and Scheinkman, 1983) on the relevant support, as shown in the proof. When concavity does not hold, the proof shows that if the higher-order effects are small, the proposition is still valid.

"Speed" in the context is the probability of queuing in the front of the queue, and it is reflected by parameter α for market maker 1 and by $1 - \alpha$ for market maker 2. When $\alpha = 1 - \alpha = 1/2$, there is no speed heterogeneity. As α increases towards 1 (or decreases towards 0), the speed heterogeneity increases. By symmetry, the proposition implies that the oversupply reaches its maximum

[18] Condition (13) ensures that $\pi(\cdot)$ is concave on a support large enough. Effectively, it ensures the equilibrium is a "stable" one. The importance of a stable equilibrium for comparative statics has been emphasized in Samuelson (1941, 1942).

work in progress
please do not circulate

at $\alpha = 1/2$ and then monotonically decreases. Conceptually, therefore, less speed heterogeneity is equivalent to more queuing uncertainty, which characterizes the degree of competition among market makers. When there is no speed heterogeneity (most queuing uncertainty), the competition is most fierce and hence the most liquidity oversupply follows. When $\alpha \to 1$, market maker 1 is almost surely to be the first in queue and, hence, he will submit a limit order of size $y^{\equiv}$ to maximize his expected profit. The competition vanishes in this extreme case and there is no oversupply (setting α to 1 in the first-order condition system 12).

Now consider adjusting the queuing uncertainty in the market by introducing a queue randomizer, $\boldsymbol{R}$, defined as a square random matrix

$$\boldsymbol{R} = \begin{bmatrix} B & 1-B \\ 1-B & B \end{bmatrix},$$

where B is a Bernoulli random variable with success probability $\beta \in [0, 1]$. For example, the exchange can apply the queue randomizer to every realized queue in a fixed time interval. Specifically, $\boldsymbol{R}$ applies to a realized queue, $\boldsymbol{k}$, by pre-multiplying it: The randomized queue becomes $\boldsymbol{K}_R = \boldsymbol{R}\boldsymbol{k}$, which is either $\boldsymbol{K}_R = \boldsymbol{k}$ (with probability β) or $\boldsymbol{K}_R \in \mathcal{K}\backslash\{\boldsymbol{k}\}$ (with probability $1 - \beta$). Therefore, when the exchange adopts a queue randomizer $\boldsymbol{R}$, the effective queue distribution can be characterized by a single parameter

$$\alpha_R(\beta) := \mathbb{P}(\boldsymbol{K}_R = [1, 2]^\intercal) = \alpha\beta + (1-\alpha)(1-\beta) \in [1-\alpha, \alpha],$$

thanks to the simplified assumption that $n = 2$. When there are $n \geq 3$ market makers, the queue distribution is more complex to characterize. Appendix A uses more structure to characterize the queue distribution, the micro-foundation of which is adopted in section 5.3 to numerically illustrate the policy implication.

Corollary 6 (Optimal queue randomizer). *There exists a randomizer $\boldsymbol{R}$ such that welfare can be (weakly) improved when $\boldsymbol{R}$ is applied.*

work in progress
please do not circulate

Effectively, the queue randomizer adjusts the queuing uncertainty (equivalently, speed heterogeneity) and also the degree of competition between the market makers. For example, consider the case where there is very severe speed heterogeneity, i.e. little queuing uncertainty. Without queue randomization, y is close to $y^{\equiv}$ (lack of competition). By proposition 5, welfare can be improved by increasing the equilibrium book depth in this case. One way to do so is to *add* more queuing uncertainty by, for example, setting $\beta \leq 1/2$ so that the originally slow market makers are more likely to be re-queued to the front. This randomization curtails the "market power" of the fast market makers and strengthens the competition, which results in larger oversupply and a deeper book in equilibrium, benefiting the investors and improving welfare.

It should be noted, however, such queue randomization must be implemented with care for at least three reasons. First, the effectiveness of queue randomization is bounded by the initial speed heterogeneity among market makers. If there is little speed heterogeneity ($\alpha \to 1/2$), the queue randomization is ineffective ($\alpha_R \to 1/2$).

Second, improperly adjusting the queuing uncertainty can destroy social welfare. For example, if the unrandomized equilibrium aggregate book depth $y(\alpha)$ (with $\alpha \geq 1/2$) already exceeds the socially optimal depth y^*: $y^* < y(\alpha)$. In this case, improper queue randomization will reduce welfare because by adding more queuing uncertainty, the oversupply exacerbates, $y(\alpha_R(\beta)) \geq y(\alpha)$, and the increased liquidity supply is not appreciated by the social planner (see lemma 4 and its discussion).

Finally, by adding queuing uncertainty, the exchange essentially deprives the market power from fast market makers (who invest heavily in technology to achieve such speed advantage). The reduced profitability might, over time, constrain the participation of market makers. Such concerns, though beyond the scope of the current model, are legitimate and should be accounted for in choosing optimal queue randomization schemes.

work in progress
please do not circulate

5.3 Numerical example

This subsection follows the CARA-normal example of Glosten (1994) and also example 1 of Back and Baruch (2013). Let the investor's private signal and endowment shock be independent and both normally distributed with zero means. The endowment-risk adjusted signal is $\Theta = S - \rho E$, which is then also normally distributed with zero mean and variance $\text{var}\Theta := \text{var}S + \rho^2 \text{var}E$. The market maker's inference, given a signal θ, about the asset value is $v(\theta) = v_0 + \gamma\theta$, where the price impact factor is $0 < \gamma = \text{var}S/\text{var}\Theta < 1$. In the numerical illustration below, the parameters are $a = v_0 + 1$, $\text{var}S = 0.2$, $\text{var}E = 0.1$, and $\rho = 0.9$.

The investor's certainty equivalent, the market makers' aggregate expected profit, and welfare can then be evaluated using the expressions derived in section 5.1. Figure 5 illustrates the shape of welfare as a function of the aggregate book depth, contrasting with the market maker's aggregate expected profit. It can be seen that welfare is *not* monotone increasing in liquidity supply in this example. Further, when market makers' expected profit maximizes, welfare can still be improved by increase the liquidity supply from $y^{\equiv}$ (≈ 0.2495) to y^* (≈ 0.5216).

When there is queuing uncertainty (and multiple market makers), the equilibrium book depth exceeds $y^{\equiv}$ and hence, there is room for welfare improvement by adding queuing uncertainty. However, when there is too much queuing uncertainty, the liquidity oversupply will be so excessive that welfare reduces. This would be the case when the equilibrium book depth exceeds y^*.

To visualize the effect directly, the following parametrization, which provides certain tractability for the general case with $n \geq 2$ (representative) market makers, is prepared for figure 6. Let each market maker be endowed with a type μ_i, such that the expected waiting time (latency) for market maker i's order to be processed by the exchange server is $\mathbb{E}L_i = \mu_i$, where L_i is exponentially distributed. The parameter μ_i is interpreted as the (expected) "latency" of market maker i; a faster market maker has a smaller μ_i. Assume L_i is independent of L_j if $i \neq j$. Appendix A gives the details and provides closed-form solutions to the queue distribution $\mathbb{P}(\boldsymbol{K})$. Finally, assume that the n market makers all have heterogeneous speed: $\mu_i - \mu_{i+1} = \Delta\mu > 0$ for all $i \in \{1, ..., n-1\}$; that is,

work in progress
please do not circulate

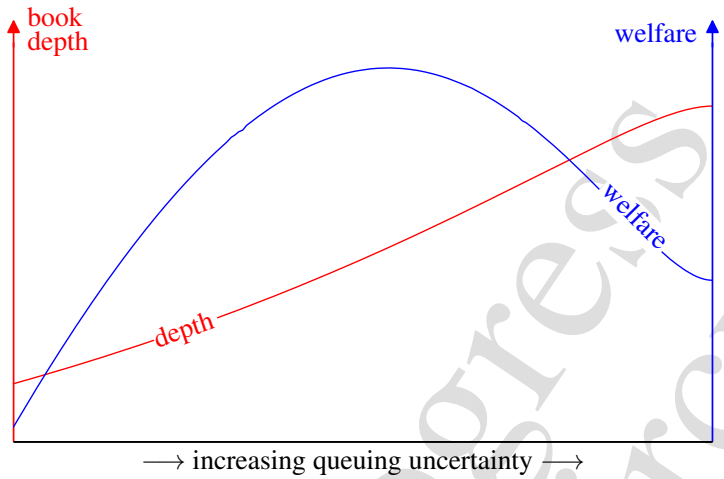


Figure 6: Queuing uncertainty and welfare. This figure illustrates how welfare and the equilibrium book depth respond to the level of queuing uncertainty in the market. The red curve is the equilibrium book depth (y, left axis) and the blue curve is welfare (w, right axis). The horizontal axis represents the level of queuing uncertainty, defined as the speed difference between market makers ($\Delta\mu$). The parameters are as described in section 5.3.

the speed difference is equally spaced. Note that, fixing n, the larger is $\Delta\mu$, the more heterogeneous are the market makers' speed and hence the more queuing uncertainty in the market.

Figure 6 plots the book depth (liquidity supply) and welfare in response of the level of queuing uncertainty in the market. The number of market makers is $n = 5$. The fastest market maker's speed is normalized to $\mu_1 = 1$ and $\Delta\mu$ ranges from 0 to 1.8. It can be seen that as queuing uncertainty increases, though the equilibrium book depth monotonically increases, welfare is first improved and then destroyed. The shape of the welfare curve is consistent with figure 5.

6 Conclusion

In a low-latency trading environment, traders cannot perfectly condition their decisions on real-time market status. This is because there is non-zero delay in between a market status update and a trader's registering of the update, and during this exact delay, the market status might evolve. This

work in progress
please do not circulate

paper acknowledges such market imperfection by introducing queuing uncertainty to the standard limit order market model. As agents do not react to the exact market status (i.e. the order book information), they essentially play a simultaneous game which leads to liquidity oversupply because the queuing uncertainty weakens the strategic substitution effects of others' decisions.

The model generates rich prediction on order book depth dynamics, characterized by short-run liquidity oversupply and high cancellation/execution ratio, both of which are amplified by the reduction in traders' latency accessing the market. The transitory nature of such liquidity oversupply suggests that massive very short-lived orders can be an equilibrium outcome owing to queuing uncertainty.

The model also speaks to recent debate on trading speed and related regulations. Under a fully endogenous extension, the model examines market quality through the expected gains from trade of all agents in the economy and shows that the exchange can improve market quality by cautiously randomizing the order queues in the limit order book, effectively adjusting the level of queuing uncertainty.

On the empirical side, this paper invites future works to test the model's prediction about order book depth dynamics and to verify the source of liquidity improvement due to the advancement in trading technology. The theoretical framework of this paper also leaves room for further exploration of, for example, how particular regulations (e.g. make/take fees, minimum order resting time, etc.) affect traders' order submission and market quality under queuing uncertainty.

work in progress
please do not circulate

Appendix

A Speed and queuing

This appendix models how the stochastic queue, $\boldsymbol{K}$, is realized according to market makers' speed. The probability masses of potential $n!$ queuing outcomes are fully characterized by n speed parameters of the market makers. The model specification captures a key feature that queue positions are *relative* in nature: Increasing all market makers' speed by a same factor should not affect the distribution of the queue. Hence, "speed" in this paper should always be understood in a relative sense.

Endow each market maker i with a speed $\mu_i^{-1} > 0$. In what follows, μ will be interpreted as the (expected) latency and, interchangeably, its inverse, μ^{-1}, will be referred to as the speed. Collate these speed parameters in a vector $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^\intercal$. Let Nature draw for each market maker a *latency*, L_i (> 0), from a commonly known distribution with c.d.f. $G(l_i; \mu_i)$ where l_i is defined on $(0, \infty)$. The latency draws are assumed to be independent. Order the realized latencies as $l_{i(1)} \leq \cdots \leq l_{i(n)}$, where $i(k)$ is the index of the market maker who draws the k-th smallest latency. Each market maker j's queue position then realizes to be $K_j = k$ if $i(k) = j$.

The latency (or, simply, delay) measures how soon the market maker can "effectuate" the order size she chose in time 1. Such effectuation delay, in a limit order market, can be the waiting time for the order to travel from a market maker's computer to the exchange server.

Assumption 4 (Latency distribution)**.** Latency L_i is exponentially distributed on $(0, \infty)$ with mean μ_i (> 0). That is, $\mathbb{P}(L_i \leq l) = 1 - e^{-l/\mu_i}$ for $l > 0$. L_i is independent of L_j for all $j \neq i$.

It is acknowledged that this assumption is not without loss of generality. First, it rules out simultaneity almost surely ($\mathbb{P}(L_i = L_j) = 0$ if $i \neq j$). Second, all market makers have strictly positive probability to be in any queue position, i.e. $\mathbb{P}(K_i = k) > 0$ for all i and k. This rules out, for example, the case where market maker 1 is never the first. In exchange, assumption 4 brings tractability and the intuitive property that only relative speed matters for the distribution of the queue.

Probability mass function for the queue vector $\boldsymbol{K}$ can be easily solved. Consider a queue realization $\boldsymbol{k} = \{k_1, \ldots, k_n\}$ determined by latencies $L_{i(1)} < \cdots < L_{i(n)}$. The probability of realizing such a queue is

$$\mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{K} = \boldsymbol{k}) = \mathbb{P}_{\boldsymbol{\mu}}(L_{i(1)} < \cdots < L_{i(n)}) = \int_0^\infty \int_{l_{i(1)}}^\infty \cdots \int_{l_{i(n-1)}}^\infty \mathrm{d}G(l_{i(n)}; \mu_{i(n)}) \ldots \mathrm{d}G(l_{i(1)}; \mu_{i(1)}),$$

which under assumption 4 has closed-form solution

$$\mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{K} = \boldsymbol{k}) = \prod_{j=1}^{n} \frac{\mu_{i(j)}^{-1}}{\sum_{h=j}^{n} \mu_{i(h)}^{-1}}. \tag{A.1}$$

work in progress
please do not circulate

The subscript $\boldsymbol{\mu}$ on $\mathbb{P}(\cdot)$ emphasizes the dependence of the probability measure $\mathbb{P}$ on market makers' speed $\boldsymbol{\mu}$. The following two lemmas will be useful in characterizing the queue distribution.

Lemma 5 (Speed relativity). *Under assumption 4, only market makers'* relative *speed matters for the queue distribution. Mathematically,* $\mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{K}) = \mathbb{P}_{c\boldsymbol{\mu}}(\boldsymbol{K})$ *for any* $c > 0$.

Lemma 6 (Speed increase and queue distribution). *Holding all other market makers' speed constant, an increase in the speed of a market maker pushes him to the front of the queue and all others to the back (in a probability sense).*

Mathematically, write C_i *as a diagonal matrix of size n with the i-th diagonal term equal to* $c > 1$ *and all other diagonal terms equal to* 1. *Then, for any* $i \neq j$, *the distribution* $\mathbb{P}_{\boldsymbol{\mu}}(K_i)$ *first-order stochastically dominates the distribution* $\mathbb{P}_{C_i\boldsymbol{\mu}}(K_i)$ *and the distribution* $\mathbb{P}_{\boldsymbol{\mu}}(K_j)$ *is first-order stochastically dominated* by *the distribution* $\mathbb{P}_{C_i\boldsymbol{\mu}}(K_j)$.

Rather intuitively, lemma 5 says if all market makers' speed are boosted by a same (strictly positive) factor, the resulting distribution of their queue does not change. That is, market makers' latencies are ordinal. Lemma 6 qualifies the effect on queue distribution of an increase in market maker i's speed. The vague idea that faster market makers are more likely to queue in the front is proved, under assumption 4, in terms of stochastic dominance.

B Notation summary

General notations:

- i: market maker index, $i \in \{1, ..., n\}$ (in section 3, $i \in [0, 1]$ under assumption 2).
- k: queue position index.
- $\boldsymbol{k}$: queue realization, a vector of size n.
- $\mathcal{K}$: the collection of all possible queues realizations.
- n: the total number of market makers.
- $\mathcal{N}$, the set of all market maker indices, $\mathcal{N} = \{1, ..., n\}$.
- π (also $\dot{\pi}$, $\ddot{\pi}$, etc.): the aggregate expected profit (and its derivatives) of all limit orders.
- q_i: the limit order size of market maker i.
- $\boldsymbol{q}$: the vector collecting all market makers limit order sizes, $\boldsymbol{q} = [q_1, ..., q_n]^\intercal$ and $\boldsymbol{q}_{-1} = \boldsymbol{q} \backslash \{q_i\}$.
- $Q_i^{\notin}(\boldsymbol{k})$: the aggregate depth before (excluding) market maker i's order, given the queue realization $\boldsymbol{k}$.
- y: the book depth.
- $y^{\equiv}$, the "break-even" book depth level at which $\dot{\pi}(y^{\equiv}) = 0$.

Additional notations in section 3:

- δ: the gap of time (latency) for a market maker to make two consecutive decisions.
- h: the number of order revisions.
- $\bar{h}$: the maximum number of order revisions; only in section 3.2.
- λ: the arrival intensity of the market order, $\mathbb{E}T = 1/\lambda$.
- h: the number of order revisions.
- $\bar{h}$: the maximum number of order revisions; only in section 3.2.
- ϕ: the probability that the market order arrives "early", i.e. arrives in the first period.
- T: the (exponentially distributed) stochastic arriving time of the market order.

Additional notations in section 4:

- a_j: the j-th best ask price; $a_j \in \{p | p > v_0, p \in \mathcal{P}\}$.

work in progress
please do not circulate

- E: the endowment shock to the investor (as her private knowledge) with zero mean.
- ε: the unobservable innovation in the asset value; normally distributed.
- $f(\theta)$: the density function of Θ.
- j: the j-th best (ask) price level.
- M: the investor's certainty equivalent of her endowment; $M := E \cdot (S + v_0) - \rho E^2/2$.
- $\mathcal{P}$: the collection of all possible price levels in the order book with $\min \mathcal{P} = p_{\min} < \max \mathcal{P} = p_{\max}$.
- ρ: the CARA investor's risk-adjustment, the product of her absolute risk-aversion coefficient and the variance of ε.
- S: the investor's private signal about the asset value with zero mean.
- Θ: the investor's endowment-risk-adjusted marginal valuation of the asset; $\Theta := S - \rho E$.
- $\lfloor \theta_j \rfloor$ and $\lceil \theta_j \rceil$: the floor and the cap that bind the investor's market order size at price level j (see lemma 2).
- V: the asset value (payoff).
- v_0: the unconditional expected asset value, $v_0 = \mathbb{E}V$; only in sections 4 and 5.
- x_j: the investor's market order size (net demand) at price level j.
- $y_{\leq j}$ (and $y_{<j}$, respectively): the cumulative depth up to and including (excluding) the j-th level.

Additional notations in section 5:

- a: the ask price.
- α: the probability of market maker 1 being first in queue; only in section 5.2.
- B: a Bernoulli random variable with success probability β; only in section 5.2.
- ce: the investor's (ex ante) certainty equivalent.
- $\boldsymbol{K}_R$: the randomized queue, $\boldsymbol{K}_R = \boldsymbol{R}\boldsymbol{K}$; only in section 5.2.
- $\boldsymbol{R}$: the queue randomizer, a 2-by-2 random matrix; only in section 5.2.
- w (and $\dot{w}$, $\ddot{w}$, etc.): the welfare, measured as the aggregate ex ante gains from trade of all agents (and its derivatives).

Additional notations in appendix A:

- μ_i: market maker i's (expected) latency; $\mathbb{E}L_i = \mu$.
- L_i: market maker i's random latency, exponentially distributed with mean μ_i.

C Proofs

Lemma 1

Proof. First, note that by quasi-concavity of $\pi(\cdot)$ and the terminal conditions, there exists a $y^{\equiv} > 0$ such that $\dot{\pi}(y^{\equiv}) = 0$, i.e. the $y^{\equiv}$-th marginal order breaks even (in expectation). Second, no market maker will submit a limit order larger than $y^{\equiv}$ because the part exceeding $y^{\equiv}$ always loses (negative marginal profit). By reducing the order size to $y^{\equiv}$ strictly improves his total profit. This reduces each market maker's strategy space to $[0, y^{\equiv}]$, and the best response correspondences implied by the first-order conditions can be summarized as a vector-valued function $f : [0, y^{\equiv}]^n \mapsto [0, y^{\equiv}]^n$, a nonempty, compact, and convex set. Note that the convexity of the value set of f is implied by the continuity of the first-order conditions ($\pi(\cdot)$ is twice differentiable by assumption 1). Then, the existence of a Nash equilibrium follows Kakutani's fixed point theorem. Finally, it is obvious that $\boldsymbol{q} = \boldsymbol{0}$ is not an equilibrium because at least one market maker wants to deviate ($\dot{\pi}(0) > 0$, assumption 1). Therefore, at least one q_i is strictly positive in equilibrium. □

Work in progress
Please do not circulate

Theorem 1

Proof. Suppose the opposite, $\sum_i q_i < y^{\equiv}$, holds in equilibrium. Then by the quasi-concavity of $\pi(\cdot)$, $\dot{\pi}(y) > 0 = \dot{\pi}(y^{\equiv})$, $\forall 0 \leq y \leq \sum_i q_i$. Note that for any queue realization $\boldsymbol{k}$, $Q_i^{\notin} + q_i \leq \sum_j q_j$ by construction (equation 1), and hence $\dot{\pi}(Q_i^{\notin} + q_i) > 0$. As the first-order condition (2) holds in equilibrium (lemma 1 guarantees the interior solution), this leads to a contradiction that the expectation of the product of strictly positive numbers equates zero. (The utility function is assumed to be strictly increasing.) Thus, the assumed inequality cannot hold in equilibrium. Instead, $\sum_i q_i \geq y^{\equiv}$ and by quasi-concavity, $\dot{\pi}(\sum_i q_i) \leq 0 = \dot{\pi}(y^{\equiv})$.

Clearly, the "if and only if" part holds true in the trivial case of $n = 1$. Consider $n \geq 2$ in what follows. First, the "only if" direction: Suppose there is an equilibrium with $\sum_i q_i = y^{\equiv}$ (the equality holds) but there is no $j \in \mathcal{N}$ such that $\mathbb{P}(K_j = 1) = 1$. Then for at least some $j \in \mathcal{N}$, $0 < \mathbb{P}(K_j = 1) < 1$ (at least someone might, but not almost surely, be the first in queue). Consider the first-order condition of such a market maker. A contradiction is derived from the first-order condition of such a market maker because the left-hand side is always strictly positive and it cannot equate zero: By the same argument as before, $\dot{\pi}(y) > 0$ for all $0 \leq y < y^{\equiv}$ and further, with probability $\mathbb{P}(K_j = 1) \in (0, 1)$, $\dot{\pi}(q_j) > 0$. Second, the "if" direction: Suppose for some $j \in \mathcal{N}$, $\mathbb{P}(K_j = 1) = 1$ but the equilibrium has the inequality strict, $\sum_i q_i < y^{\equiv}$. This simply implies that the first-order condition for such an almost-surely first-in-queue market maker does not hold, as he can always improve his expected utility by increase his supply to $y^{\equiv}$. Hence, it cannot be an equilibrium. □

Proposition 1

Proof. Differentiate the left-hand side of the first-order condition (2) with respect to q_j for some $j \neq i$ to get $\partial^2 \mathbb{E}\pi/(\partial q_i \partial q_j) = \mathbb{E}\left[\mathbb{1}_{\{K_j < K_i\}} \cdot \ddot{\pi}(Q_i^{\notin} + q_i)\right] \leq 0$, by concavity of $\pi(\cdot)$. This then establishes that q_i and q_j are strategic substitutes. Next, in equilibrium, the first-order condition (2) holds and it implies an implicit function of $q_i = q_i^*(q_j)$. The substitution rate is the first order derivative of q_i with respect to q_j, which by implicit function theorem is $\partial q_i^*/\partial q_j = -\mathbb{E}[\mathbb{1}_{\{K_j < K_i\}}\ddot{\pi}(Q_i^{\notin}+q_i)]/\mathbb{E}[\ddot{\pi}(Q_i^{\notin}+q_i)]$. The partial derivative is negative because of the concavity, i.e. $\ddot{\pi}(\cdot) < 0$. Note that there is at least one queue in which $k_j > k_i$ and the indicator function becomes zero. Hence, the numerator is always strictly less negative than the denominator and the substitution rate is bounded by $(-1, 0]$. □

Corollary 1

Proof. (The proof of the existence of the equilibrium is reminiscent to the proof of lemma 1 and is omitted for brevity.) Consider the equilibrium order submission in period 0. Expand the first term of the first-order condition (3) as $\phi\mathbb{E}\dot{\pi}\left(Q_i^{\notin} + q_i\right) = \phi\mathbb{P}(K_i = n)\mathbb{E}\left[\dot{\pi}\left(Q_i^{\notin} + q_i\right) | K_i = n\right] + \phi\mathbb{P}(K_i < n)\mathbb{E}\left[\dot{\pi}\left(Q_i^{\notin} + q_i\right) | K_i < n\right]$. Note that $\mathbb{E}\left[\dot{\pi}\left(Q_i^{\notin} + q_i\right) | K_i = n\right] = \dot{\pi}\left(\sum_j q_j\right)$. Hence, the first order

work in progress
please do not circulate

condition can be rewritten as

$$\mathbb{P}(K_i = n)\,\dot{\pi}\left(\sum_j q_j\right) = -\mathbb{P}(K_i < n)\mathbb{E}\left[\dot{\pi}\left(Q_i^{\notin} + q_i\right) \mid K_i < n\right] - \frac{1-\phi}{\phi}\mathbb{E}\left[\mathbb{1}_{\{Q_i^{\notin}+q_i<y^{\equiv}\}}\dot{\pi}\left(Q_i^{\notin} + q_i\right)\right].$$

Because $\sum_j q_j \geq Q_i^{\notin} + q_i$ for all realizations of $\boldsymbol{K}$ and because $\pi(\cdot)$ is quasi-concave, $\dot{\pi}\left(\sum_j q_j\right)$ must be negative (otherwise $\dot{\pi}\left(\geq Q_i^{\notin} + q_i\right) \geq 0$ and the above first-order condition cannot hold). That is, $\dot{\pi}(\sum_j q_j) \leq 0$ in equilibrium, i.e. there is oversupply in period 0. That there is no oversupply in the period 1 trivially follows as all oversupplying orders will be canceled. □

Proposition 2, part I

Proof. For completeness, this proof first evaluates the first-order condition with the continuum of market makers: $\phi\mathbb{E}\dot{\pi}(K_i q) + (1-\phi)\mathbb{P}(K_i q < y^{\equiv})\,\mathbb{E}\left[\dot{\pi}(K_i q) \mid K_i q < y^{\equiv}\right] = 0$. There are two cases, depending on whether $q \lesseqgtr y^{\equiv}$. Suppose $q \leq y^{\equiv}$, then $\mathbb{P}(K_i q < y^{\equiv}) = 1$ and the condition in the second expectation does not bind. The expression then simplifies to $\mathbb{E}\dot{\pi}(K_i q) = 0$. The expectations can be integrated to get $\mathbb{E}\dot{\pi}(K_i q) = \int_0^1 \dot{\pi}(K_i q)/\mathrm{d}K_i = \pi(K_i q)/q\big|_{K_i=0}^{K_i=1} = \pi(q) = 0$. However, $\pi(q) = 0$ implies $q > y^{\equiv}$ (or $q = 0$, which clearly is not an equilibrium as any market maker would deviate to supply more than zero) because of the quasi-concavity of $\pi(\cdot)$. Therefore, $q > y^{\equiv}$ and, similarly, evaluating the expectations gives the zero-profit condition (5). Next, by implicit function theorem, the zero-profit condition implies, in equilibrium, $\partial q^*/\partial\phi = \pi(y^{\equiv})/\left(\phi^2\dot{\pi}(q^*)\right)$. The partial derivative is negative because $\pi(y^{\equiv}) > 0$ (in fact, $\pi(\cdot)$ reaches is its global maximum at $y^{\equiv}$) and $\dot{\pi}(q^*) < 0$ as $q^* > y^{\equiv}$. □

Proposition 2, continued

Proof. Consider the first "if" stated in the proposition. If $\bar{h} = 1$, then by proposition 2, part I, $\partial q(0)/\partial\phi < 0$, and it then follows $\partial q(0)/\partial\lambda < 0$ because ϕ is strictly increasing in λ. If $\bar{h} > 1$, then by implicit function theorem, the first-order condition (5) at $h = 0$ gives

$$\frac{\partial q(0)}{\partial\lambda} = -\frac{(\pi(q(0)) - \pi(\delta q(0)))\partial\phi/\partial\lambda}{\phi\dot{\pi}(q(0)) + (1-\phi)\delta\dot{\pi}(\delta q(0))} < 0.$$

The "< 0" holds because the numerator is negative ($\pi(q(0)) < 0$, $\pi(\delta q(0)) > 0$, and ϕ strictly increases in λ) and because the denominator is also negative as it is the second-order condition for the optimality (otherwise the zero-profit condition does not give an equilibrium).

Consider the second "if" stated in the proposition. If $\bar{h} = 1$, then by implicit function theorem, the first-order condition (5) at $h = 0$ gives $\partial q(0)/\partial\delta = -(\pi(q(0)) - \pi(y^{\equiv}))/\dot{\pi}(q(0)) \cdot \partial\phi/\partial\delta < 0$, because $\pi(q(0)) < 0$, $\pi(y^{\equiv}) > 0$, $\dot{\pi}(q(0)) < 0$, and ϕ strictly increases in δ. If $\bar{h} > 1$, rewrite the first-order condition (5) at $h = 0$ as $\pi(q(0)) + (1-\phi)\pi(\delta q(0))/\phi = 0$ and then, by implicit function theorem,

$$\frac{\partial q(0)}{\partial\delta} = -\frac{1-\phi}{\phi} \cdot \frac{-\lambda\pi(\delta q(0)) + \phi q(0)\dot{\pi}(\delta q(0))}{\phi\dot{\pi}(q(0)) + (1-\phi)\delta\dot{\pi}(\delta q(0))}$$

work in progress
please do not circulate

(note that $\partial(1-\phi)/\phi/\partial\delta = -\lambda\cdot(1-\phi)/\phi^2$). The denominator is negative by the optimality argument. It remains to determine the sign of the numerator. By mean value theorem, there exists some $x \in [0, \delta q(0)]$ such that $\pi(\delta q(0)) = \delta q(0)\dot{\pi}(x) \geq \delta q(0)\dot{\pi}(\delta q(0))$, where the last inequality follows assumption 3. Thus the numerator has $-\lambda\pi(\delta q(0)) + \phi q(0)\dot{\pi}(\delta q(0)) \leq (\phi - \delta\lambda)q(0)\dot{\pi}(\delta q(0)) < 0$ because $\phi = 1 - e^{-\lambda\delta} < \delta\lambda$. Therefore, $\partial q(0)/\partial\delta < 0$ also holds for $\bar{h} > 1$. □

Corollary 2

Proof. Define the moment at which the displayed book depth equates the break-even depth $y^{\equiv}$ by $t^{\equiv}(\delta) := y^{\equiv}/(q_0(\delta))$, where $q_0(\delta)$ is the notation for the equilibrium order size at $t = 0$ given latency δ. Note that $t^{\equiv}(\delta)$ has an upper bound of 1 because $q_0(\delta) > y^{\equiv}$ (liquidity oversupply). By proposition 2, $t^{\equiv}$ is strictly increasing in δ for all $\delta > 0$. In addition, $\lim_{\delta\downarrow 0} t^{\equiv}(\delta) = y^{\equiv}/(q_0(0)) > 0$. Therefore, the graph of $t^{\equiv}(\delta)$ in the first quadrant of $t^{\equiv}$-δ plane is a strictly increasing curve with $t^{\equiv}(0) > 0$ and $t^{\equiv}(\infty) < 1$, crossing each line of $t^{\equiv} = h\delta$ for all $h \in \{1, 2, ...\}$ once and only once. Denote the intersection of $t^{\equiv}(\delta) = h\delta$ by δ_h, with $\delta_0 := \infty$. Thus, for all $\delta \in [\delta_h, \delta_{h-1})$, $(h-1)\delta < t^{\equiv}(\delta) \leq h\delta$, or $h - 1 < \lceil t^{\equiv}/\delta \rceil \leq h$. Note that $\lceil t^{\equiv}/\delta \rceil$ (the integer ceiling of $t^{\equiv}/\delta$) is exactly the number of revisions that the market makers make before the "true" book depth reverts to $y^{\equiv}$. The existence of the series is validated. □

Lemma 2

Proof. Solving the lower and the upper corners of equation (6), i.e. $x_j(\lfloor\theta_j\rfloor) = 0$ and $x_j(\lceil\theta_j\rceil) = y_1$, gives the floor and the ceiling. (Note that $x_j(\theta)$ is monotone increasing in θ.) □

Proposition 3

Proof. From equation (8), the second-order derivative of $\pi_j(y)$ can be computed as

$$\ddot{\pi}_j(y) = -\rho\Big(a_j - v(\lfloor\theta_j\rfloor + \rho y)\Big) f(\lfloor\theta_j\rfloor + \rho y).$$

Suppose there exists some $y^{\equiv}$ such that $\dot{\pi}_j(y^{\equiv}) = 0$. Then $0 = \dot{\pi}_j(y^{\equiv}) = \int_{\lfloor\theta_j\rfloor+\rho y^{\equiv}}^{\infty} \Big(a_j - v(\theta)\Big) f(\theta)\mathrm{d}\theta < \Big(a_j - v(\lfloor\theta_j\rfloor + \rho y^{\equiv})\Big) \cdot \Big(1 - F(\lfloor\theta_j\rfloor + \rho y^{\equiv})\Big)$, or, $a_j - v(\lfloor\theta_j\rfloor + \rho y) = -\ddot{\pi}_j(y^{\equiv}) > 0$. The inequality follows because $v(\cdot)$ is assumed, as in Biais, Martimort, and Rochet (2000), to be increasing. By theorem M.C.4 of Mas-Colell, Whinston, and Green (1995), this implies the quasi-concavity of $\pi_j(y)$. Hence, assumption 1 holds.

Similarly, suppose at some $\hat{y}$ $\ddot{\pi}_j(\hat{y}) = 0$, which implies $a_j - v(\lfloor\theta_j\rfloor + \rho\hat{y}) = 0$ because Θ has a continuous support and hence the density function is strictly positive on its support. Then it can be evaluated that the third-order derivative of $\pi_j(y)$ at such a $\hat{y}$ is $\rho^2\dot{v}(\lfloor\theta_j\rfloor + \rho y)f(\lfloor\theta_j\rfloor + \rho y) > 0$ (as $\dot{v}(y) > 0$). This implies that $\dot{\pi}_j(y)$ is quasi-convex.

To prove assumption 3, suppose there exists some $y_j^{\equiv}$ such that $\dot{\pi}_j(y_j^{\equiv}) = 0$. Then by the quasi-concavity of $\pi_j(\cdot)$, $\dot{\pi}_j(y) > 0$ for all $0 \leq y < y_j^{\equiv}$. Then, $0 \leq \dot{\pi}_j(y) = \int_{\lfloor\theta_j\rfloor+\rho y}^{\infty} \Big(a_j - v(\theta)\Big) f(\theta)\mathrm{d}\theta < \Big(a_j - v(\lfloor\theta_j\rfloor + \rho y)\Big)\Big(1 - F(\lfloor\theta_j\rfloor + \rho y)\Big) = -\ddot{\pi}_j(y)f(\lfloor\theta_j\rfloor + \rho y)/(1 - F(\lfloor\theta_j\rfloor + \rho y))$, where the inequality

work in progress
please do not circulate

"<" follows the monotonicity of $v(\theta)$ and the last equality follows the expression of $\dot{\pi}_j(\cdot)$. Rearrange the above inequality to get

$$\ddot{\pi}_j(y) < -\dot{\pi}_j(y)\frac{f(\lfloor\theta_j\rfloor+\rho y)}{1-F(\lfloor\theta_j\rfloor+\rho y)} < 0 \tag{C.1}$$

because the hazard rate is always positive. This shows that $\dot{\pi}_j(y) > 0$ is sufficient for $\pi_j(y)$ to be concave. □

Proposition 4

Proof. The proof goes in several steps. Step one shows how to derive the first-order condition (10). Note that for $h \geq 1$, $\partial\pi_{j+h}(y)/\partial q_{i,j} = -\int_{\lfloor\theta_{j+h}\rfloor}^{\lfloor\theta_{j+h}\rfloor+\rho y}\left(a_{j+h}-v(\theta)\right)f(\theta)\mathrm{d}\theta = \dot{\pi}_{j+k}(y) - \dot{\pi}_{j+k}(0)$. Hence,

$$\frac{\partial}{\partial q_{i,j}}\left(\pi_{j+h}(y+q)-\pi_{j+h}(y)\right) = \dot{\pi}_{j+h}(y+q)-\dot{\pi}_{j+h}(y). \tag{C.2}$$

The Bellman equation for price level $j+1$ (see the j-th Bellman equation in equation 9) should hold in equilibrium: $u_{i,j+1} = \mathbb{E}\left[\pi_{j+1}\left(Q^{\notin}_{i,j+1}+q_{i,j+1}\right)-\pi_{j+1}\left(Q^{\notin}_{i,j+1}\right)\right]+u_{i,j+2}$. Optimality of the right-hand side requires the first order condition to hold:

$$\mathbb{E}\dot{\pi}_{j+1}\left(Q^{\notin}_{i,j+1}+q_{i,j+1}\right)+\frac{\partial u_{i,j+2}}{\partial q_{i,j+1}} = 0. \tag{C.3}$$

Also, differentiate both sides of the Bellman equation with respect to $q_{i,j}$ to get

$$\frac{\partial u_{i,j+1}}{\partial q_{i,j}} = \frac{\partial}{\partial q_{i,j}}\mathbb{E}\left[\pi_{j+1}\left(Q^{\notin}_{i,j+1}+q_{i,j+1}\right)-\pi_{j+1}\left(Q^{\notin}_{i,j+1}\right)\right]+\frac{\partial u_{i,j+2}}{\partial q_{i,j+1}} = -\mathbb{E}\dot{\pi}_{j+1}\left(Q^{\notin}_{i,j+1}\right),$$

where the second equality follows by substituting equation (C.2) with $h = 1$, $y = Q^{\notin}_{i,j+1}$, and $q = q_{i,j+1}$, and equation (C.3). Finally, note that optimality of the Bellman equation (9) gives $\mathbb{E}\dot{\pi}_j\left(Q^{\notin}_{i,j}+q_{i,j}\right)+\partial u_{i,j+1}/\partial q_{i,j}$. Substitute the above expression for $\partial u_{i,j+1}/\partial q_{i,j}$ into the optimality, and the result is the first-order condition given in equation (10).

The second step derives an important properties about the expected profit function $\pi_j(y)$. The quasi-concavity of $\pi_j(y)$ and the quasi-convexity of $\dot{\pi}_j(y)$ together imply the first property that if $\dot{\pi}_j(y) \leq 0$, then $\dot{\pi}_j(y) \leq 0$ for all $y \geq 0$. To see this, note that if the opposite is true, i.e. if there exists some $\hat{y} \geq 0$ such that $\dot{\pi}_j(\hat{y}) > 0$, then by continuity $\dot{\pi}_j(y)$ must cross the horizontal axis at least once *from below to above*. The crossing point implies a local minimum of $\pi_j(y)$, contradicting the quasi-concavity of $\pi_j(y)$.

With the above results, the last step is to suppose $\dot{\pi}_j(y_j) \geq 0$, where $y_j := \sum_i q_{i,j}$, and show this leads to a contradiction. Note that $\left(Q^{\notin}_{i,j}+q_{i,j}\right) \leq y_j$ in equilibrium because there is non-zero probability for market maker i's order not to queue in the last position. Given that $\dot{\pi}_j(y_j) \geq 0$, therefore, $\mathbb{E}\ddot{\pi}_j\left(Q^{\notin}_{i,j}+q_{i,j}\right) < 0$ by the inequality (C.1). Differentiate the first-order condition (10) on both sides with respect to $q_{i,j}$: $\mathbb{E}\ddot{\pi}_j\left(Q^{\notin}_{i,j}+q_{i,j}\right)+\mathbb{E}\ddot{\pi}_{j+1}\left(Q^{\notin}_{i,j+1}\right) = 0$, or $\mathbb{E}\ddot{\pi}_{j+1}\left(Q^{\notin}_{i,j+1}\right) = -\mathbb{E}\ddot{\pi}_j\left(Q^{\notin}_{i,j}+q_{i,j}\right) > 0$,

Work in progress
Please do not circulate

where the inequality follows the above argument. Then a contradiction follows:

$$0 < \mathbb{E}\ddot{\pi}_{j+1}\left(Q^{\notin}_{i,j+1}\right) < -\mathbb{E}\left[\dot{\pi}_{j+1}\left(Q^{\notin}_{i,j+1}\right)\frac{f\left(\lfloor\theta_j\rfloor + Q^{\notin}_{i,j+1}\rho\right)}{1-F\left(\lfloor\theta_j\rfloor + Q^{\notin}_{i,j+1}\rho\right)}\right] < -\mathbb{E}\left[\dot{\pi}_{j+1}\left(Q^{\notin}_{i,j+1}\right)\right]\frac{f\left(\lfloor\theta_j\rfloor\rho\right)}{1-F\left(\lfloor\theta_j\rfloor\rho\right)} < 0$$

The second inequality follows inequality (C.1). The third inequality follows the increasing hazard rate of Θ. The last inequality follows the first-order condition (10) that $\mathbb{E}\dot{\pi}_{j+1}\left(Q^{\notin}_{i,j+1}\right) = \mathbb{E}\dot{\pi}_j\left(Q^{\notin}_{i,j} + q_{i,j}\right)$, which is positive because $\left(Q^{\notin}_{i,j} + q_{i,j}\right) \leq y_j$ and because of the quasi-concavity of $\pi_j(\cdot)$. To conclude, this contradiction refutes the assumption that $\dot{\pi}_j(y_{<j}) \geq 0$ in equilibrium. Therefore, $\dot{\pi}_j(y_{<j}) < 0$, i.e. at each price level there is liquidity oversupply. Therefore, $\dot{\pi}_j(y_j) < 0$. □

Corollary 3

Proof. Suppose the equilibrium depth at some price level a_j is $y_j > 0$. By proposition 4, $\dot{\pi}_j(y_j) = \int_{\lfloor\theta_j\rfloor+\rho y_j}^{\infty}\left(a_j - v(\theta)\right)f(\theta)\mathrm{d}\theta < 0$. Then the expected marginal profit of the first unit of limit orders at the next possible price, $a_j + \rho$, is $\int_{\lfloor\theta_j\rfloor+\rho y_j}^{\infty}\left(a_j + \rho - v(\theta)\right)f(\theta)\mathrm{d}\theta$, which is negative for sufficiently small ρ. Then by property 1) developed in the proof of proposition 4, the expected marginal profit at price $a_j + \rho$ is always negative for all depth $y \geq 0$. Hence, no market maker is willing to post any order at this price level, leaving it a hole in the book. □

Corollary 4

Proof. In this proof, the superscript "*" indicates "with queuing uncertainty", while the superscript "≡" indicates "under the break-even condition of Glosten (1994)". Consider the expected marginal profit functions at price level j: $\dot{\pi}^*_j(y) = \int_{\lfloor\theta_j\rfloor^*+\rho y}^{\infty}\left(a^*_j - v(\theta)\right)f(\theta)\mathrm{d}\theta$ and $\dot{\pi}^{\equiv}_j(y) = \int_{\lfloor\theta_j\rfloor^{\equiv}+\rho y}^{\infty}\left(a^{\equiv}_j - v(\theta)\right)f(\theta)\mathrm{d}\theta$. In particular, $\lfloor\theta_j\rfloor^* = a^*_j - v_0 + \rho y^*_{<j}$ and $\lfloor\theta_j\rfloor^{\equiv} = a^{\equiv}_j - v_0 + \rho y^{\equiv}_{<j}$. Decompose $\dot{\pi}^*_j(y)$ as

$$\dot{\pi}^*_j(y) = \underbrace{\int_{\lfloor\theta_j\rfloor^*+\rho y}^{\infty}\left(a^{\equiv}_j - v(\theta)\right)f(\theta)\mathrm{d}\theta}_{=\dot{\pi}^{\equiv}_j\left(y+(a^*_j-a^{\equiv}_j)/\rho+(y^*_{<j}-y^{\equiv}_{<j})\right)} + \underbrace{\left(a^*_j - a^{\equiv}_j\right)\left(1-F(\lfloor\theta_j\rfloor^* + \rho y)\right)}_{>0}.$$

When there is no holes in the order book, $a^*_j = a^{\equiv}_j$, and the following inequality follows at equilibrium:

$$\dot{\pi}^{\equiv}_j\left(y^*_j + (a^*_j - a^{\equiv}_j)/\rho + (y^*_{<j} - y^{\equiv}_{<j})\right) < \dot{\pi}^*_j\left(y^*_j\right) < 0 = \dot{\pi}^{\equiv}_j\left(y^{\equiv}_j\right).$$

By quasi-concavity of $\pi_j(\cdot)$, therefore, $y^*_j + (y^*_{<j} - y^{\equiv}_{<j}) > y^{\equiv}_j$, which simplifies to $y^*_{\leq j} > y^{\equiv}_{\leq j}$. □

Lemma 3

Proof. Fix a limit order vector $\boldsymbol{q}$ and a possible queue realization $\boldsymbol{k}$. The expected profit of market maker i, given the queue, is $\pi\left(Q^{\notin}_i(\boldsymbol{k}) + q_i\right) - \pi\left(Q^{\notin}_i(\boldsymbol{k})\right)$. Sum over all market makers' expected

working in progress
please do not circulate

profit to get

$$\sum_i \left[\pi\left(Q_i^{\not q}(\boldsymbol{k}) + q_i\right) - \pi\left(Q_i^{\not q}(\boldsymbol{k})\right)\right] = \left[\pi\left(q_{i(1)}\right) - \pi(0)\right] + \left[\pi\left(q_{i(1)} + q_{i(2)}\right) - \pi\left(q_{i(1)}\right)\right] + \cdots + \left[\pi\left(\sum_{k=1}^{n} q_{i(k)}\right) - \pi\left(\sum_{k=1}^{n-1} q_{i(k-1)}\right)\right] = \pi\left(\sum_i q_i\right).$$

In the equation, subscripts on q means that $i(k)$ is the index of the agent at the k-th position of the queue. The last equality follows because all the in-between $\pi(\cdot)$ terms cancel. (Recall also that $\pi(0) = 0$.) The aggregate expected profit is queue-irrelevant. □

Lemma 4 and proposition 5

Proof. That $w(y) > \pi(y)$ is trivial because by definition welfare accounts for both the certainty equivalent of the investor and the expected profit of market makers: $w(y) = \pi(y) + \mathrm{ce}(y)$. Evaluate the difference between $\dot{w}(y)$ and $\dot{\pi}(y)$ to get $\dot{w}(y) - \dot{\pi}(y) = \int_{\lfloor\theta\rfloor+\rho y}^{\infty} (v_0 + \theta - \rho y - a)\, f(\theta)\mathrm{d}\theta > 0$. The integration is positive because $\theta > \lfloor\theta\rfloor + \rho y = a - v_0 + \rho y$. Therefore, for all $y \geq 0$, $\dot{w}(y) > \dot{\pi}(y)$. By proposition 3, $\pi(y)$ exhibits quasi-concavity and hence, for all $0 \leq y \leq y^{\equiv}$, $\dot{w}(y) > \dot{\pi}(y) \geq 0$. Similarly, after some simplification, $\ddot{w}(y) - \ddot{\pi}(y) = -\rho \cdot (1 - F(\lfloor\theta\rfloor + \rho y)) < 0$ for all y. By proposition 3, $\ddot{w}(y) < \ddot{\pi}(y) < 0$ on $y \in [0, y^{\equiv}]$. Hence, $w(y)$ is concavely increasing on $[0, y^{\equiv}]$. The above proves lemma 4. Proposition 5 then immediately follows: the welfare maximizing $y^* > y^{\equiv}$ because $\dot{w}(y^{\equiv}) > 0$. □

Proposition 6

Proof. Define $y := q_1 + q_2$ as the aggregate book depth. In equilibrium, the first-order condition system (12) holds (the interior solution is guaranteed by lemma 1) and by chain-rule,

$$\frac{\partial q_1}{\partial \alpha} = \frac{\dot{\pi}(y)}{\Delta}\left[\frac{1-\alpha}{\alpha}\ddot{\pi}(q_2) + 2\ddot{\pi}(y)\right] > 0$$
$$\frac{\partial q_2}{\partial \alpha} = -\frac{\dot{\pi}(y)}{\Delta}\left[\frac{\alpha}{1-\alpha}\ddot{\pi}(q_1) + 2\ddot{\pi}(y)\right] < 0$$

where $\Delta = \alpha \cdot (1-\alpha)\ddot{\pi}(q_1)\ddot{\pi}(q_2) + (1-\alpha)^2\ddot{\pi}(q_2)\ddot{\pi}(y) + \alpha^2\ddot{\pi}(q_1)\ddot{\pi}(y) > 0$ is the determinant of the Jacobian matrix. To see why the partial derivatives (and Δ) are so signed, note that by lemma 1, $0 \leq y = q_1 + q_2 \leq 2y^{\equiv}$ in equilibrium. Then by condition (13) $\ddot{\pi}(y) < 0$ because $\dot{\pi}(y)$ is quasi-convex and $\ddot{\pi}(0) < 0$ as implied by proposition 3. (Recall that $\dot{\pi}(y) < 0$ in equilibrium because of liquidity oversupply.) The effect on aggregate book depth, y, is

$$\frac{\partial y}{\partial \alpha} = \frac{\partial q_1}{\partial \alpha} + \frac{\partial q_2}{\partial \alpha} = -\frac{\dot{\pi}(y)}{\Delta}\left[\frac{\alpha}{1-\alpha}\ddot{\pi}(q_1) - \frac{1-\alpha}{\alpha}\ddot{\pi}(q_2)\right].$$

By symmetry, $\partial y/\partial\alpha = 0$ at $\alpha = 1/2$. It remains to show that $\partial^2 y/\partial\alpha^2 < 0$ at $\alpha = 1/2$, making a local maximum: $\frac{\partial y}{\partial \alpha}\big|_{\alpha=1/2} = -4\dot{\pi}(y)\ddot{\pi}(q_1)/\Delta < 0$ (note that $q_1 = q_2$ at $\alpha = 1/2$ by symmetry). □

Working in progress
please do not circulate

Corollary 5

Proof. Rewrite $\partial y/\partial\alpha$ as

$$\frac{\partial y}{\partial \alpha} = -\frac{\dot{\pi}(y)}{\Delta}\dot{\pi}(q_1)\frac{\alpha}{1-\alpha}\left(\frac{\ddot{\pi}(q_1)}{\dot{\pi}(q_1)} - \frac{\ddot{\pi}(q_2)}{\dot{\pi}(q_2)}\right)$$

by noting $\dot{\pi}(q_1)/\dot{\pi}(q_2) = (1-\alpha)^2/\alpha^2$. Because $\partial q_1/\partial\alpha > 0$ and $\partial q_2/\partial\alpha < 0$ (see the proof of proposition 6), for $\alpha \geq 1/2$, $q_1 \geq q_2$ (that is, the faster market maker supplies more). Therefore, it suffices to sign the derivative of $\ddot{\pi}(x)/\dot{\pi}(x)$ on $x \in [0, y^{\equiv}]$: If $\ddot{\pi}(x)/\dot{\pi}(x)$ is decreasing, then $\partial y/\partial\alpha \leq 0$ and y reaches its global maximum at $\alpha = 1/2$. Note that

$$\frac{\partial}{\partial x}\left(\frac{\ddot{\pi}(x)}{\dot{\pi}(x)}\right) = \frac{1}{\dot{\pi}(x)^2}\left(\dddot{\pi}(x)\dot{\pi}(x) - \ddot{\pi}(x)^2\right)$$

is negative if $\dddot{\pi}(x) \leq 0$ ($\dot{\pi}(x) > 0$ for $x \in [0, y^{\equiv}]$ by quasi-concavity of $\pi(\cdot)$). That is, concavity of $\dot{\pi}(x)$ on $[0, y^{\equiv}]$ is a sufficient condition for the result.

Consider next the case where $\dot{\pi}(x)$ is not concave but only the first three order effects matter:

$$\dot{\pi}(x) = c_1 \cdot (x - y^{\equiv}) + c_2 \cdot (x - y^{\equiv})^2 + O((x - y^{\equiv})^3) \approx c_1 \cdot (x - y^{\equiv}) + c_2 \cdot (x - y^{\equiv})^2.$$

Then it can be derived immediately that

$$\frac{\partial}{\partial x}\left(\frac{\ddot{\pi}(x)}{\dot{\pi}(x)}\right) = \frac{1}{\dot{\pi}(x)^2}\left(\dddot{\pi}(x)\dot{\pi}(x) - \ddot{\pi}(x)^2\right) = -\frac{1}{\dot{\pi}(x)^2}\left(c_2^2 \cdot (x - y^{\equiv})^2 + (c_1 + c_2 \cdot (x - y^{\equiv}))^2\right) < 0.$$

Hence, when the above approximation holds, $\partial y/\partial\alpha$ is negative on $\alpha \geq 1/2$ and, by symmetry, positive on $\alpha \leq 1/2$. At $\alpha = 1/2$, y has the unique maximum. □

Corollary 6

Proof. The proof directly follows corollary 5, which implies the equilibrium book depth y is a one-to-one mapping from the effective queue distribution parameter, α_R, which in turn is a one-to-one mapping from β, which governs the queue randomizer. Therefore, by choosing β (hence also $\mathbf{R}$), the equilibrium depth can be adjusted to maximize welfare, which is a function in y: $w(y) = w(y(\alpha_R(\beta)))$. (Note that the optimal level β might be cornered.) □

Lemma 5

Proof. The result follows equation (A.1): $\mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{K}) = \prod_{j=1}^{n}[\mu_{i(j)}^{-1}/\sum_{h=j}^{n}\mu_{i(h)}^{-1}] = \prod_{j=1}^{n}[c\mu_{i(j)}^{-1}/\sum_{h=j}^{n}c\mu_{i(h)}^{-1}] = \mathbb{P}_{c\boldsymbol{\mu}}(\boldsymbol{K})$, which holds for all $c > 0$. □

work in progress
please do not circulate

Lemma 6

Proof. By the closed-form solution of equation (A.1), the log-likelihood of a queue, $\boldsymbol{k}$, is

$$\ln \mathbb{P}(\boldsymbol{K} = \boldsymbol{k}) = \ln \mathbb{P}\left(L_{i(1)} < \cdots < L_{i(n)}\right) = \sum_{j=1}^{n} \ln \mu_{i(j)}^{-1} - \sum_{j=1}^{n} \ln\left(\sum_{h=j}^{n} \mu_{i(h)}^{-1}\right).$$

Consider agent i with position p in the queue. Denote the sub-queue of all other agents by $\boldsymbol{K}_{-i}$. Then fixing $p \in \{1, \ldots, n\}$ and a sub-queue realization $\boldsymbol{k}_{-i}$, the log-likelihood of such a queue is

$$\ln \mathbb{P}(K_i = p, \boldsymbol{k}_{-i}) = \ln \mu_i^{-1} + \sum_{j \neq p}^{n} \ln \mu_{i(j)}^{-1} - \sum_{j=1}^{p} \ln\left(\mu_i^{-1} + \sum_{h \geq j, h \neq p} \mu_{i(h)}^{-1}\right) - \sum_{j=p+1}^{n} \ln\left(\sum_{h=j}^{n} \mu_{i(h)}^{-1}\right).$$

The marginal effect of an increase in μ_i^{-1} on the log-likelihood is

$$\frac{\partial \ln \mathbb{P}(K_i = p, \boldsymbol{k}_{-i})}{\partial(\mu_i^{-1})} = \mu_i - \sum_{j=1}^{p} \left[\mu_i^{-1} + \sum_{h \geq j, h \neq p} \mu_{i(h)}^{-1}\right]^{-1}.$$

In particular,

$$\frac{\partial \ln \mathbb{P}(K_i = 1, \boldsymbol{k}_{-i})}{\partial(\mu_i^{-1})} = \frac{\mu_i - 1}{\mu_i^{-1} + \sum_{j \neq i} \mu_j^{-1}} > 0;$$

$$\frac{\partial \ln \mathbb{P}(K_i = n, \boldsymbol{k}_{-i})}{\partial(\mu_i^{-1})} = -\sum_{j=1}^{n-1} \left[\mu_i^{-1} + \sum_{h=j}^{n-1} \mu_{i(h)}^{-1}\right]^{-1} < 0;$$

and for $1 \leq p < n$, the difference $\ln \mathbb{P}(K_i = p, \boldsymbol{k}_{-i}) - \ln \mathbb{P}(K_i = p+1, \boldsymbol{k}_{-i}) = 1/\left(\sum_{h=p+1}^{n} \mu_{i(h)}^{-1}\right) > 0$. That is, the marginal effect is monotone decreasing, from positive to negative, along the queue position, p, of agent i. This monotonicity holds for any sub-queue realization $\boldsymbol{k}_{-1}$. Therefore, for $c > 1$, there exists a $p^* \in \{1, \ldots, n-1\}$ such that $\mathbb{P}_{\boldsymbol{\mu}}(K_i \leq k) = \sum_{p \leq k} \mathbb{P}_{\boldsymbol{\mu}}(K_i = p) < \sum_{p \leq k} \mathbb{P}_{C_i\boldsymbol{\mu}}(K_i = p) = \mathbb{P}_{C_i\boldsymbol{\mu}}(K_i \leq k)$ if and only if $p \leq p^*$, i.e. $\mathbb{P}_{C_i\boldsymbol{\mu}}(K_i \leq k)$ is first-order stochastically dominated by $\mathbb{P}_{\boldsymbol{\mu}}(K_i \leq k)$. The proof for the other half of the lemma is similar. □

References

Anand, Amber and Kumar Venkataraman. 2013. "Should Exchanges impose Market Maker obligations?" Manuscript. 8

Back, Kerry and Shmuel Baruch. 2013. "Strategic Liquidity Provision in Limit Order Markets." *Econometrica* 81 (1):363–392. 2, 7, 32, 41

Baruch, Shmuel and Laurence R. Glosten. 2013. "Fleeting Orders and the Competitive Equilibrium." Manuscript. 6, 24, 30

work in progress please do not circulate

Biais, Bruno, Thierry Foucault, and Sophie Moinas. 2013. "Equilibrium Fast Trading." Manuscript. 7

Biais, Bruno, Pierre Hillion, and Chester Spatt. 1995. "An Empirical Analysis of the Limit Order Book and the Order Flow in the Paris Bourse." *The Journal of Finance* 50:1655–1689. 1, 31

Biais, Bruno, David Martimort, and Jean-Charles Rochet. 2000. "Competing Mechanisms in a Common Value Environment." *Econometrica* 68 (4):799–837. 25, 26, 28, 30, 32, 49

———. 2013. "Corrigendum to 'Competing Mechanisms in a Common Value Environment'." *Econometrica* 81 (1):393–406. 25

Brogaard, Jonathan, Björn Hagströmer, Lars L. Nordén, and Ryan Riordan. 2013. "Trading Fast and Slow: Colocation and Market Quality." Manuscript. 18

Brogaard, Jonathan A. 2011. "Minimum quote life and maximum order message-to-trade ratio." Techreport, Foresight, Government Officie for Science, U.K. 24

Bulow, Jeremy I., John D. Geanakoplos, and Paul D. Klemperer. 1985. "Multimarket Oligopoly: Strategic Substitutes and Complements." *Journal of political Economy* 93 (3):488–511. 13

Easley, David and Maureen O'Hara. 1992. "Time and the Process of Security Price Adjustment." *The Journal of Finance* 47 (2):576–605. 32

Egginton, Jared, Bonnie F. van Ness, and Robert A. van Ness. 2012. "Quote stuffing." Manuscript. 23

Fisher, Franklin M. 1961. "The Stability of the Cournot Oligopoly Solution: The Effects of Speeds of Adjustment and Increasing Marginal Costs." *The Review of Economic Studies* 28 (2):125–135. 16

Foucault, Thierry. 1999. "Order Flow Composition and Trading Costs in a Dynamic Limit Order Market." *Journal of Financial Markets* 2:99–134. 2

Foucault, Thierry, Ohad Kadan, and Eugene Kandel. 2005. "Limit Order Book as a Market for Liquidity." *Review of Financial Studies* 18(4):1171–1217. 2

Foucault, Thierry and Albert J. Menkveld. 2008. "Competition for Order Flow and Smart Order Routing Systems." *The Journal of Finance* 63:119–158. 2, 33

Friederich, Sylvain and Richard Payne. 2012. "Order to trade ratios and their impact on Italian stock market quality." Techreport, Foresight, Government Officie for Science, U.K. 7, 24

Gai, Jiading, Chen Yao, and Mao Ye. 2013. "The Externalities of High-Frequency Trading." Manuscript. 1, 7, 19

Glosten, Laurence R. 1994. "Is the Electronic Limit Order Book Inevitable?" *The Journal of Finance* 49:1127–1161. 2, 3, 4, 7, 8, 12, 17, 20, 23, 31, 32, 41, 51

work in progress
please do not circulate

Glosten, Lawrence R. and Paul R. Milgrom. 1985. "Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Agents." *Journal of Financial Economics* 14:71–100. 32

Goettler, Ronald L., Christine A. Parlour, and Uday Rajan. 2005. "Equilibrium in a Dynamic Limit Order Market." *The Journal of Finance* 60(5):2149–2192. 2, 26

———. 2009. "Informed Traders in Limit Order Markets." *Journal of Financial Economics* 93:67–87. 2, 26

Hahn, Frank H. 1962. "The stability of the Cournot oligopoly solution." *The Review of Economic Studies* 29 (4):329–331. 16

Hasbrouck, Joel. 2013. "High Frequency Quoting: Short-Term Volatility in Bids and Offers." Manuscript. 6

Hasbrouck, Joel and Gideon Saar. 2009. "Technology and Liquidity Provision: The Blurring of Traditional Definitions." *Journal of Financial Markets* 12:143–172. 3

———. 2013. "Low-Latency Trading." Manuscript, Cornell University. 1, 4

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld. 2011. "Does Algorithmic Trading Improve Liquidity?" *The Journal of Finance* 66:1–33. 4

Hoffmann, Peter. 2013. "A Dynamic Limit Order Market with Fast and Slow Traders." Manuscript. 7

Hollifield, Burton, Robert A. Miller, and Patrick Sandås. 2004. "Empirical Analysis of Limit Order Markets." *Review of Economic Studies* 71:1027–1063. 3

Kervel, Vincent Van. 2013. "Liquidity: What you see is what you get?" Manuscript. 33

Kreps, David M. and José A. Scheinkman. 1983. "Quantity precommitment and Bertrand competition yield Cournot outcomes." *The Bell Journal of Economics* :326–337. 38

Kyle, A. S. 1985. "Continuous auctions and insider trading." *Econometrica* 53:1315–36. 32

Liu, Wai-Man. 2009. "Monitoring and limit order submission risks." *Journal of Financial Markets* 12:107–141. 3

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*. New York: Oxford university press. 49

Menkveld, Albert J. and Marius A. Zoican. 2013. "Need for speed? Low latency trading and adverse selection." Manuscript. 7

Milgrom, Paul and Chris Shannon. 1994. "Monotone Comparative Statics." *Econometrica* 62 (1):157–180. 16

work in progress
please do not circulate

Pagnotta, Emiliano and Thomas Philippon. 2012. "Competing on Speed." Manuscript. 7

Parlour, Christine A. 1998. "Price dynamics in limit order markets." *The Review of Financial Studies* 11:789–816. 2

Parlour, Christine A. and Duane J. Seppi. 2008. "Limit Order Markets: A Survey." In *Handbook of Financial Intermediation and Banking*, edited by A.W.A. Boot and A.V. Thakor. Amsterdam, Netherlands: Elsevier Publishing. 2

Raman, Vikas and Pradeep Yadav. 2013. "The Who, Why, and How Well of Order Revisions: An Analysis of Limit Order Trading." Manuscript. 3

Riordan, Ryan and Andreas Storkenmaier. 2012. "Latency, liquidity and price discovery." *Journal of Financial Markets* 15:416–437. 7

Roşu, Ioanid. 2009. "A Dynamic Model of the Limit Order Book." *The Review of Financial Studies* 22 (11):4601–4641. 2

Samuelson, Paul A. 1941. "The Stability of Equilibrium: Comparative Statics and Dynamics." *Econometrica* 9 (2):97–120. 16, 38

———. 1942. "The Stability of Equilibrium: Linear and Nonlinear Systems." *Econometrica* 10 (1):1–25. 16, 38

Sandås, Patrick. 2001. "Adverse Selection and Competitive Market Making: Empirical Evidence from a Limit Order Market." *Review of Financial Studies* 14 (3):705–734. 2, 7, 33

Seppi, Duane J. 1997. "Liquidity provision with limit orders and a strategic specialist." *The Review of Financial Studies* 10:103–150. 2

Vayanos, Dimitri and Jiang Wang. 2012. "Liquidity and Asset Returns Under Asymmetric Information and Imperfect Competition." *The Review of Financial Studies* 25 (5):1339–1365. 27

Zhu, Haoxiang. 2012. "Finding a Good Price in Opaque Over-the-Counter Markets." *The Review of Financial Studies* 25 (4):1255–1285. 6

List of Figures

1 Game structure comparison 2

2 Equilibrium illustration with $n = 2$ market makers 12

3 Order book depth dynamics 20

4 The investor's optimal market order size 28

work in progress
please do not circulate

5 Welfare as a function of liquidity supply 36
6 Queuing uncertainty and welfare 42

work in progress
please do not circulate