Informed Trading and Maker-Taker Fees in a Low-Latency Limit Order Market[*]

Michael Brolley[†] Katya Malinova[‡]

University of Toronto

June 7, 2013

Abstract

We model a financial market where privately informed investors trade in a limit order book monitored by low-latency liquidity providers. Price competition between informed limit order submitters and low-latency market makers allows us to capture tradeoffs between informed limit and market orders in a methodologically simple way. We apply our model to study maker-taker fees — a prevalent, but controversial exchange fee system that pays a maker rebate for liquidity provision and levies a taker fee for liquidity removal. When maker-taker fees are passed through to all traders, only the total exchange fee per transaction has an economic impact, consistent with previous literature. However, when investors pay only the average exchange fee through a flat fee per transaction—as is common practice in the industry—maker-taker fees have an impact beyond that of a change in the total fee. An increase in the maker rebate lowers trading costs, increases trading volume, improves welfare, but decreases market participation by investors.

[*]Financial support from the Social Sciences and Humanities Research Council is gratefully acknowledged. We thank Andreas Park, Liyan Yang, as well as participants of the 2012 CEA meeting and 8th Annual Central Bank Workshop on the Microstructure of Financial Markets for helpful comments and discussions.

[†]Email: michael.brolley@utoronto.ca; web: http://individual.utoronto.ca/brolleym/.

[‡]Email: katya.malinova@utoronto.ca; web: http://individual.utoronto.ca/kmalinova/.

Equity trading around the world is highly automated. Exchanges maintain limit order books, where orders to trade pre-specified quantities at pre-specified prices are arranged in a queue, according to a set of priority rules.[1] A trade occurs when an arriving trader finds the terms of limit orders at the top of the queue sufficiently attractive and posts a marketable order that executes against these posted limit orders.

To improve the trading terms, or liquidity, offered in their limit order books, many exchanges incentivize traders who provide, or "make" liquidity. Trading venues pay a rebate to submitters of executed limit orders, and they finance these rebates by levying higher fees to remove, or "take" liquidity on submitters of marketable orders. This practice is referred to as "maker-taker" pricing.[2] Moreover, with the rise of algorithmic trading, exchanges have adopted technology that offers extremely high-speed, or "low-latency" market data transmission, in order to appeal to speed-sensitive participants. The rebates, along with the increased speed of trading systems, has given rise to "a new type of professional liquidity provider": proprietary trading firms that "take advantage of low-latency systems" and provide liquidity electronically.[3]

The role of maker-taker pricing and the new low-latency computerized traders remains controversial. Proponents maintain that the new trading environment benefits all market participants through increased competition. Opponents argue that the increased competition for liquidity provision makes it difficult for long-term investors to trade via limit orders and that it compels them to trade with more expensive marketable orders.[4]

To study the impact of the recent market structure developments, we must first un-

[1]Most exchanges sort limit orders first by price, and then by the time of arrival, maintaining a so-called price-time priority.

[2]IOSCO Consultation Report, International Organization of Securities Commissions (2011), page 19.

[3]SEC Concept Release on Market Structure, Securities and Exchange Commission (2010).

[4]See, e.g., GETCO's comments on maker-taker fees in options markets to SEC (available at `http://www.getcollc.com/images/uploads/getco_comment_090208.pdf`), in favor, and TD Securities' comments on IIROC 11-0225 (`www.iiroc.ca`), Alpha Trading Systems' September 2010 Newsletter (`http://www.alphatradingsystems.ca/`, against.

derstand tradeoffs between market and limit orders in current markets, namely, in limit order books where low-latency traders act as de facto market makers. It is particularly important to understand these tradeoffs in presence of private information — when some traders have a speed advantage, others arguably need an informational advantage to compete. Existing models typically either study markets where all available liquidity is provided by competitive market makers or assume that all traders strategically choose between supplying and demanding liquidity and that they have temporal market power in liquidity provision.[5] Analyzing a trader's choice between market and limit orders is methodologically challenging. When liquidity providers have market power, a limit order submitter must optimally choose the limit order price, while accounting for the impact of the price choice on the probability of the limit order execution. The resulting dynamic optimization problem is especially difficult with informed liquidity provision, as the limit order price may reveal the liquidity provider's private information.

In this paper, we build on Kaniel and Liu (2006) and provide a model of a limit order book where privately informed traders (who we refer to as "investors") trade with market and limit orders, and, when submitting a limit order, compete with uninformed low-latency market makers. Price competition in liquidity provision between informed and uninformed (but fast) traders is a key methodological insight in our paper — it allows us to circumvent the complexity of the optimization problem, because all limit orders are posted at prices that yield zero-profits to low-latency liquidity providers.

Our setup captures the low-latency liquidity providers' speed advantage in interpreting market data, such as trades and quotes. In practice, the speed advantage comes at a cost and low-latency liquidity providers are arguably at a disadvantage (relative to

[5]See, e.g., Glosten and Milgrom (1985), Kyle (1985), Easley and O'Hara (1987), or Glosten (1994) for competitive market maker models; Parlour (1998), Foucault (1999), Foucault, Kadan, and Kandel (2005), Goettler, Parlour, and Rajan (2005), Rosu (2009), Back and Baruch (2013), Baruch and Glosten (2012) for limit order books with uninformed liquidity provision, and Kaniel and Liu (2006), Goettler, Parlour, and Rajan (2009), and Rosu (2011), for informed liquidity provision. See also the survey by Parlour and Seppi (2008) for further discussion.

humans or sophisticated algorithms) when processing more complex information, such as news reports. We capture this difference in information processing skills by allowing investors an informational advantage with respect to the security's fundamental value. Additionally, investors have private valuations (e.g., liquidity needs) for the security.[6]

In equilibrium, an investor's behavior is governed by his aggregate valuation, which is the sum of his private valuation of the security and his expected value of the security. Investors with extreme aggregate valuations optimally choose to submit market orders, investors with moderate valuations submit limit orders, and investors with aggregate valuations close to the public expectation of the security's value abstain from trading.

Changes in exogenous market factors (e.g., a trading platform's fee structure) lead to changes in the marginal aggregate valuations that investors require to submit market or limit orders, and to changes in liquidity, trading volume, and market participation by investors. We apply our model to study the impact one such change: the use of maker-taker pricing.

Consistent with the previous literature (see Angel, Harris, and Spatt (2011) and Colliard and Foucault (2012)), when maker-taker fees are passed through, the split does not play an economically meaningful role in our model, because any increase in the maker rebate is passed to the takers through a narrower bid-ask spread, exactly offsetting an increase in the taker fee.[7] As a practical matter, however, many long-term investors do not pay taker fees directly and do not receive maker rebates but instead pay a flat fee per trade to their broker, while low-latency traders incur per-trade exchange fees and rebates. We investigate this variant on maker-taker pricing by applying our model to a setting where we assume that investors pay only the average maker-taker fee.

[6]Assuming that traders have liquidity needs is common practice in the literature on trading with asymmetric information, to avoid the no-trade result of Milgrom and Stokey (1982); modelling these needs as private valuations allows use to derive welfare implications.

[7]When discussing our results, we focus on the prevalent industry practice of a negative maker fee, or a rebate, but our analysis extends to the case of a positive maker fee.

Ceteris paribus, an increase in a maker rebate lowers the bid-ask spread and induces investors previously indifferent to market and limit orders to trade with market orders (since an investor's trading costs consists, loosely, of the bid-ask spread and the flat fee levied by their broker). Consequently, the probability of a market order submission increases, and so does the trading volume. This would lead to brokers paying taker fees more frequently and consequently charging investors a higher flat fee.

We support this intuition numerically and find further that the increase in the flat fee is more than offset by the decline in the bid-ask spread. For a fixed total exchange fee, investors' overall trading costs thus decline with an increase in the maker rebate. The marginal submitter of a market order then requires weaker information, and the price impact of a trade declines.

To analyze the impact of maker-taker fees on welfare, we follow Bessembinder, Hao, and Lemmon (2012) and define a social welfare measure to reflect allocative efficiency. Specifically, with each trade, the social gains from trade increase by the difference between the buyer's and the seller's private valuations, net of differences in trading fees, and we define the social welfare to be the expected social gains per period. We find numerically that, for a fixed total exchange fee, the welfare increases in the maker rebate, provided the maker rebate is not too large.[8]

This change is driven by investors switching from submitting limit orders to trading with market orders, increasing the probability that gains from trade are realized. Limit order provision by investors is inefficient for two reasons. First, an investor who submits a limit order risks non-execution of his own order. Second, this investor possibly imposes a negative externality on the previous period investor — if the earlier investor submitted a limit order on the opposite side, then that order does not execute. In the presence of low-latency liquidity providers who collect (some of) the maker rebates,

[8]When the maker rebate is sufficiently large, the spread becomes sufficiently small, and, in equilibrium, investors choose to trade exclusively with market orders. Any further increase in the maker rebate that is financed by an increase in the taker fee leads to a decline in the quoted spreads, but yields no further economically meaningful implications.

the flat fee levied on investors acts as a tax on limit orders and as a subsidy on market orders; it thus mitigates the negative externality.

Our paper is most closely related to Colliard and Foucault (2012) and Foucault, Kadan, and Kandel (2012), who theoretically analyze the impact of maker-taker fees. Colliard and Foucault (2012) study trader behavior in a model where symmetrically informed traders choose between limit and market orders. They show that, absent frictions, the split between maker and taker fees has no economic impact, and they focus on the impact of the total fee charged by an exchange. Foucault, Kadan, and Kandel (2012) argue that in the presence of a minimum tick size, limit order book prices may not adjust sufficiently to compensate traders for changes in the split between maker and taker fees. They then show that exchanges may use maker-taker pricing to balance supply and demand of liquidity, when traders exogenously act as makers or takers. Skjeltorp, Sojli, and Tham (2012) support theoretical predictions of Foucault, Kadan, and Kandel (2012) empirically, using exogenous changes in maker-taker fee structure and a technological shock for liquidity takers. Rosu (2009) finds that the prediction of Colliard and Foucault (2012) on the neutrality of the breakdown of the total fees holds in presence of asymmetric information. Our predictions on spreads, price impact, and volume, and the prediction of Colliard and Foucault (2012) are supported empirically by Malinova and Park (2011), who study the impact of the introduction of maker rebates on the Toronto Stock Exchange.

Our work is also closely linked to Degryse, Achter, and Wuyts (2012), who study the impact of the post-trade clearing and settlement fees. In their model, the clearing house may set a flat fee for all trades or impose different fees, depending on whether a trade was internalized. They find that the fee structure affects the welfare of market participants, and that the optimal structure depends on the size of the clearing fee.

The maker-taker pricing model is related to the payment for order flow model, see, e.g., Kandel and Marx (1999), Battalio and Holden (2001), or Parlour and Rajan

(2003), in the sense that both systems aim to incentivize order flow; Battalio, Shkilko, and Van Ness (2012) and Anand, McCormick, and Serban (2013) empirically compare market quality under the maker-taker pricing with that under the payment for order flow model.

Our analysis of a limit order market with competitive informed liquidity provision to the broader theoretical literature on specialist and limit order markets, see, e.g., Glosten and Milgrom (1985), Kyle (1985), Easley and O'Hara (1987), and Glosten (1994), for competitive uninformed liquidity provision; Parlour (1998), Foucault (1999), Foucault, Kadan, and Kandel (2005), Goettler, Parlour, and Rajan (2005), Rosu (2009), Back and Baruch (2013), and Baruch and Glosten (2012) for limit order books with strategic uninformed liquidity provision; Kaniel and Liu (2006), Goettler, Parlour, and Rajan (2009), and Rosu (2011), for strategic informed liquidity provision.[9] The pricing rule model is very closely related to the equilibrium pricing rule in Kaniel and Liu (2006); differently to them, all traders in our model behave strategically. We complement the theoretical literature that focuses on the trading strategies of low-latency traders, see e.g., Biais, Foucault, and Moinas (2012), Foucault, Hombert, and Rosu (2012), Hoffmann (2012), and McInish and Upson (2012).

Finally, the role of low-latency traders as competitive liquidity providers is supported empirically by, e.g., Hasbrouck and Saar (2011), Hendershott, Jones, and Menkveld (2011), Hendershott and Riordan (2012), and Jovanovic and Menkveld (2011).

1 The Model

We model a financial market where risk-neutral investors enter the market sequentially to trade a single risky security for informational and liquidity reasons (as in Glosten and Milgrom (1985)). Trading is conducted via limit order book. Investors choose between posting a limit order to trade at pre-specified prices and submitting a market order to

[9] See also the survey by Parlour and Seppi (2008) for further related papers.

trade immediately with a previously posted limit order. Additionally, we assume the presence of low-latency liquidity providers, who choose to act as market makers, and to only submit limit orders. These traders possess a speed advantage that allows them to react to changes in the limit order book faster than other market participants. We assume that they are uninformed and that they have no liquidity needs. Low-latency liquidity providers compete in the sense of Bertrand competition, are continuously present in the market, and ensure that the limit order book is always full.

Security. There is a single risky security with an unknown liquidation value. This value follows a random walk, and at each period t experiences an innovation δ_t. The fundamental value in period t is given by

$$V_t = \sum_{\tau \leq t} \delta_\tau \tag{1}$$

Innovations δ_t are identically and independently distributed, according to density function $\bar{g}$ on $[-1, 1]$, which is symmetric around zero. We focus on intraday trading, and we assume that extreme innovations to the security's fundamental value are less likely than innovations that are close to 0 (i.e., that $\bar{g}'(\cdot) \leq 0$ on $[0, 1]$).[10]

Investors. There is a continuum of risk-neutral investors. At each period t, a single investor randomly arrives at the market. Upon entering the market, the investor is endowed with liquidity needs, which we quantify by assigning the investor a private value for the security, denoted by y_t, uniformly distributed on $[-1, 1]$. Furthermore, the investor learns the period t innovation to the fundamental value, δ_t.[11]

Investor Actions. An investor can submit an order upon arrival and only then. He can buy or sell a single unit (round lot) of the risky security, or abstain from

[10] For the existence of the competitive equilibrium, we additionally require that the mean of G is not too large. At the time of writing, we do not have a closed-form condition on the primitives; we provide further details in the Theorem 1 and in the discussion of out-of-equilibrium beliefs in Appendix A.

[11] Assuming that traders have liquidity needs is common practice in the literature on trading with asymmetric information, to avoid the no-trade result of Milgrom and Stokey (1982). We also solved for an equilibrium, assuming that only a fraction of traders become informed, with qualitatively similar results.

trading.[12] If the investor chooses to buy, he either submits a market order and trades with an existing order at the previously posted ask price ask_t in period t, or he posts a limit buy order at the bid price bid_{t+1} in period t, for execution in period $t+1$. Similarly for the decision to sell. Limit orders that are submitted in period t and that do not execute in period $t+1$ are automatically cancelled. An investor may submit at most one order, and upon the order's execution or cancellation the investor leaves the market forever.

Low-Latency Liquidity Providers. There is continuum of low-latency liquidity providers who are always present in the market. They hold a speed advantage in reacting to changes in the limit order book. These traders act as market makers and post limit orders in response to changes in the limit order book. They compete in prices in the sense of Bertrand competition. Low-latency liquidity providers are risk-neutral, they do not receive any information about the security's fundamental value, and they do not have liquidity needs.

The Limit Order Book. Trading is organized via limit order book, which is comprised of limit orders. Limit orders last for one period. Arguably, this simplifying assumption is particularly realistic in presence of low-latency traders, as slower investors may fear that their orders become stale and will be "picked off" by the low-latency traders. Low-latency liquidity providers ensure that the limit order book is always "full" by submitting a limit order when there is no standing limit order on the buy or the sell side. The limit order book thus always contains one buy limit order and one sell limit order, upon arrival of an investor in period t. A trade occurs in period t when the investor that arrives in period t chooses to submit a market order.

Exchange Fees. The limit order book is maintained by an exchange that charges time-invariant fees for executing orders. The focus of this paper is on maker-taker fees, which depend on the order type (market or limit), but do not depend on whether

[12] We will refer to investors in the male form, and we will refer to the low-latency liquidity providers in the female form.

an order is a "buy" or a "sell". To simplify the exposition, we assume that the total exchange fee per transaction is 0 and focus on the split of this fee into a taker fee, which we denote by f and a maker rebate $(= f)$.[13] For most of our discussion, we focus on the prevalent practice where the taker fee is positive. The intuition for our results extends for the reverse scenario where market order submitters receive a rebate and submitters of executed limit orders pay a positive fee.[14]

Low-latency liquidity providers receive maker rebates for executed limit orders. We study two settings. In the first, investors pay the taker fees and maker rebates on a trade-by-trade basis. In the second, "flat-fee" setting, investors only pay the average maker-taker fee, through a flat fee per transaction. Our "flat-fee" setting reflects a common practice in the industry: long-term investors typically access exchanges via brokers, who pay the exchange maker-taker fees but levy a flat fee per transaction on their customers.[15]

Public Information. Investors and low-latency liquidity providers observe the history of transactions as well as limit order submissions and cancellations. We denote the history of trades and quotes up to (but not including) period t by H_t. The structure of the model is common knowledge among all market participants, but an investor's liquidity needs and his knowledge of an innovation to the fundamental value are private.

Low-Latency Liquidity Provider Information. Low-latency liquidity providers are able to detect whether a newly posted limit order stems from an investor with liquidity and informational needs or from other low-latency liquidity providers. This assumption ensures that the model is tractable. We believe that it is consistent with reality, because low-latency traders are allegedly good at identifying, for instance, larger institutional orders. Further, within our model, low-latency liquidity providers react virtually instantaneously to changes in the limit order book, whereas investors who

[13]Colliard and Foucault (2012) study the impact of a total exchange fee.

[14]This "inverted" pricing is often referred to by industry participants as "taker-maker pricing", as it is utilized, for instance, by NASDAQ OMX BX.

[15]See, e.g., the Interactive Brokers fees: `http://www.interactivebrokers.com/en/p.php?f=commission`

trade for liquidity and informational reasons arrive at discrete time intervals — consequently, limit orders that are posted by low-latency liquidity providers are identified by the reaction time. Finally, from a technical perspective, this assumption is equivalent to assuming presence of a single low-latency liquidity provider who chooses to act competitively.

Timing of Actions. We model intraday trading. Periods are measured in discrete units (which we denote by t) with no specific beginning or end. Each period marks the arrival of an investor. At the beginning of any period t, the limit order book is full in the sense that it contains one buy limit order and one sell limit order. In each period t, an investor enters the market, observes the transaction and quote history H_t, his liquidity needs measured by his private valuation y_t, and the innovation δ_t to the security's value. This investor posts a limit or a market order, or abstains from trading.

When a market order is posted, it executes against a limit order that was posted in period $t-1$, and the investor leaves the market forever. The limit order book immediately reacts to the information contained in the period t market order and the low-latency liquidity providers post limit orders to buy and sell.

When a limit order is posted in period t, this order remains in the market until the period $t+1$ investor makes his trading choice.[16] This limit order possibly interacts with the period $t+1$ investor's market order. As with market orders, the limit order book reacts to the information contained in the period t limit order, with a low-latency liquidity provider posting a limit order on the opposite side of the book.

Investor Payoffs. The payoff to an investor who buys one unit of the security in period t is given by the difference between the security's fundamental value in period t, V_t, and the price that the investor paid for this unit; similarly for a sell decision. We normalize the payoff to a non-executed order to 0. Investors are risk neutral, and they aim to maximize their expected payoffs. The period t investor with

[16]The assumption that limit orders last for a single period is common in the literature, see, e.g., Foucault (1999), and it ensures that the model is tractable.

private valuation y_t has the following expected payoffs to submitting, respectively, a market buy order to trade immediately at the prevailing ask price ask_t and a limit buy order at price bid_{t+1}:

$$\pi^{\mathrm{MB}}_{t,\mathrm{inv}}(y_t,\delta_t) = y_t + \mathsf{E}[V_t \mid \delta_t, H_t] - \mathsf{ask}_t - \mathrm{fee}^{\mathrm{M}}_{\mathrm{inv}} \tag{2}$$

$$\begin{aligned}\pi^{\mathrm{LB}}_{t,\mathrm{inv}}(y_t,\delta_t,\mathsf{bid}_{t+1}) &= \mathsf{Pr}(\mathrm{MS}_{t+1}(\mathsf{bid}_{t+1}) \mid \delta_t, H_t) \\ &\quad \times \left(y_t + \mathsf{E}[V_{t+1} \mid \delta_t, H_t, \mathrm{MS}_{t+1}(\mathsf{bid}_{t+1})] - \mathsf{bid}_{t+1} - \mathrm{fee}^{\mathrm{L}}_{\mathrm{inv}}\right)\end{aligned} \tag{3}$$

where $\mathrm{MS}_{t+1}(\mathsf{bid}_{t+1})$ represents the period $t+1$ investor's decision to submit a market order to sell at price bid_{t+1} (this decision is further conditional on the additional information available to the period $t+1$ investor); $\mathrm{fee}^{\mathrm{M}}_{\mathrm{inv}}$ and $\mathrm{fee}^{\mathrm{L}}_{\mathrm{inv}}$ denote the exchange fees levied on investors trading with market and limit orders, respectively. An investor's payoff to submitting a limit order in period t accounts for the fact that a limit order submitted in period t either executes or is cancelled in period $t+1$. We focus on the intraday trading, and we assume no discounting. Payoffs to sell orders are analogous.

Low-Latency Liquidity Provider Payoffs. A low-latency trader observes the period t investor's action before posting her period t limit order. Moreover, she will post a limit buy order in period t only if the period t investor does not post a buy limit order.[17] A low-latency trader in period t has the following payoff to submitting a limit buy order at price bid_{t+1} is given by

$$\begin{aligned}\pi^{\mathrm{LB}}_{t,\mathrm{LLT}}(\mathsf{bid}_{t+1}) &= \mathsf{Pr}(\mathrm{MS}_{t+1}(\mathsf{bid}_{t+1}) \mid \text{investor action at } t, H_t) \\ &\quad \times \left(\mathsf{E}[V_{t+1} \mid H_t, \text{investor action at } t, \mathrm{MS}_{t+1}(\mathsf{bid}_{t+1})] - \mathsf{bid}_{t+1} - \mathrm{fee}^{\mathrm{L}}_{\mathrm{LLT}}\right)\end{aligned} \tag{4}$$

where the exchange fee $\mathrm{fee}^{\mathrm{L}}_{\mathrm{LLT}}$ incurred by a low-latency trader when her limit order is executed equals the maker rebate, i.e., $\mathrm{fee}^{\mathrm{L}}_{\mathrm{LLT}} = -f < 0$; analogously for sell orders.

[17] With unit demands of investors, a low-latency trader has no incentive to post a limit order "into a queue": a market sell order that executes against the "first in the queue" order is informative, thus the liquidity provider will want to modify her "second in the queue" order upon execution of the first.

2 Equilibrium: All Pay Maker-Taker Fees

In this section, we assume that maker rebates and taker fees are passed through to all market participants on a per-trade basis.

2.1 Pricing and Decision Rules

Equilibrium Pricing Rule. We look for an equilibrium, in which low-latency liquidity providers post competitive limit orders and make zero profits, in expectation. We denote the equilibrium bid and ask prices in period t by bid_t^* and ask_t^*, respectively, and we use MB_t^* and MS_t^* denote, respectively, the period t investor's decisions to submit a market buy order price ask_t^* and a market sell order at price bid_t^*.

The low-latency liquidity provider payoffs, given by equation (4), then imply the following competitive equilibrium pricing rules, for the maker rebate f:

$$\mathsf{bid}_t^* = \mathsf{E}[V_t \mid H_t, \mathrm{MS}_t(\mathsf{bid}_t^*)] + f \quad (5)$$

$$\mathsf{ask}_t^* = \mathsf{E}[V_t \mid H_t, \mathrm{MB}_t(\mathsf{ask}_t^*)] - f \quad (6)$$

where we used the fact that history H_{t-1} together with the period $t-1$ investor's action yield the same information about the security's value V_t as history H_t (because information about V_t is only publicly revealed through investors' actions).

Investor Actions with Competitive Liquidity Provision. We focus on investor choices to buy; sell decisions are analogous. An investor can choose to submit a market order or a limit order, and, if he chooses to submit a limit order, technically, he may also choose the limit price. We search for an equilibrium where low-latency liquidity providers ensure that bid and ask prices are set competitively and equal the expected security value, conditional on the information available to the low-latency liquidity providers. An investor's choice of the limit price is thus mute, since a limit order that is posted at a price other than the prescribed, competitive equilibrium prices either yields the submitter negative profits in expectation or does not execute, because of the presence of low-latency traders. Because an investor is always able to obtain a zero profit by abstaining from trade, we restrict attention to limit orders posted at the competitive, equilibrium prices.

Non-Competitive Limit Orders. Formally, the zero probability of execution for limit orders posted at non-competitive prices is achieved by defining appropriate beliefs of market participants, regarding the information content of a limit order that is posted at an "out-of-the-equilibrium" price (e.g., when the period t investor posts a limit order to buy at a price different from bid^*_{t+1}) — so-called out-of-equilibrium beliefs. The appropriate definition of out-of-equilibrium beliefs is frequently necessary to formally describe equilibria with asymmetric information. To see the role of these beliefs in our model, observe first that when an order is posted at the prescribed, competitive equilibrium price, market participants derive the order's information content by Bayes' Rule, using their knowledge of equilibrium strategies. The knowledge of equilibrium strategies, however, does not help market participants to assess the information content of an order that cannot occur in equilibrium — instead, traders assess such an order's information content using out-of-the-equilibrium beliefs. We describe these beliefs in Appendix A, and we focus on prices and actions that occur in equilibrium in the main text.

Investor Equilibrium Payoffs. Because innovations to the fundamental are independent across periods, all market participants interpret the transaction history in the same manner. A period t investor decision then does not reveal any additional information about innovations δ_τ, for $\tau < t$, and the equilibrium pricing conditions

(5)-(6) can be written as

$$\mathsf{bid}_t^* = \mathsf{E}[V_{t-1} \mid H_t] + \mathsf{E}[\delta_t \mid H_t, \mathrm{MS}_t(\mathsf{bid}_t^*)] + f \tag{7}$$

$$\mathsf{ask}_t^* = \mathsf{E}[V_{t-1} \mid H_t] + \mathsf{E}[\delta_t \mid H_t, \mathrm{MB}_t(\mathsf{ask}_t^*)] - f \tag{8}$$

The independence of innovations across time further allows us to decompose investors' expectations of the security's value, to better understand investor equilibrium payoffs. The period t investor's expectation of the security's value in period t is given by

$$\mathsf{E}[V_t \mid \delta_t, H_t] = \delta_t + \mathsf{E}[V_{t-1} \mid H_t] \tag{9}$$

When the period t investor submits a limit order to buy, his order will be executed in period $t+1$ (or never), and we thus need to understand this investor's expectation of the time $t+1$ value, conditional on his private and public information and on the order execution, $\mathsf{E}[V_{t+1} \mid \delta_t, H_t, \mathrm{MS}_{t+1}(\mathsf{bid}_{t+1}^*)]$. Since the decision of the period $t+1$ investor reveals no additional information regarding past innovations, we obtain

$$\mathsf{E}[V_{t+1} \mid \delta_t, H_t, \mathrm{MS}_{t+1}(\mathsf{bid}_{t+1}^*)] = \mathsf{E}[V_{t-1} \mid H_t] + \delta_t + \mathsf{E}[\delta_{t+1} \mid \delta_t, H_t, \mathrm{MS}_{t+1}(\mathsf{bid}_{t+1}^*)] \tag{10}$$

Further, the independence of innovations implies that, conditional on the period t investor submitting a limit buy order at price bid_{t+1}^*, the period t investor's private information of the innovation δ_t does not afford him an advantage in estimating the innovation δ_{t+1} or the probability of a market order to sell in period $t+1$, relative to the information H_{t+1} that will be publicly available in period $t+1$ (including the information that will be revealed by the period t investor's order). Consequently, the period t investor's expectation of the innovation δ_{t+1} coincides with the corresponding expectation of the low-latency liquidity providers, conditional on the period t investor's limit buy order at price bid_{t+1}^*.

The above insight, together with conditions (7)-(8) on the equilibrium bid and ask prices, allows us to rewrite investor payoffs, given by expressions (2)-(3) as:

$$\pi_t^{\text{MB}}(y_t, \delta_t) = y_t + \delta_t - \mathsf{E}[\delta_t \mid H_t, \text{MB}_t(\mathsf{ask}_t^*)] \tag{11}$$

$$\pi_t^{\text{LB}}(y_t, \delta_t) = \mathsf{Pr}(\text{MS}_{t+1}(\mathsf{bid}_{t+1}^*)) \mid \text{LB}_t(\mathsf{bid}_{t+1}^*), H_t) \left(y_t + \delta_t - \mathsf{E}[\delta_t \mid \text{LB}_t(\mathsf{bid}_{t+1}^*), H_t]\right) \tag{12}$$

where we used the fact that when investors pay exchange fees per-trade, an investor's fees for trading with market and limit orders, respectively, are $\text{fee}_{\text{inv}}^{\text{M}} = f$ and $\text{fee}_{\text{inv}}^{\text{L}} = -f$.

Equations (11)-(12) illustrate, in particular, that investor payoffs are independent of the exchange fees, provided the total exchange fee is 0. In the Internet Appendix, we further show that, for a non-zero exchange fee, the levels of maker the rebate and the taker fee only affect investor payoffs through the total exchange fee, consistent with Colliard and Foucault (2012).

Proposition 1 (Independence of the Maker-Taker Split) *For a fixed total exchange fee, investors' equilibrium strategies and payoffs do not depend on the split of the total fee into maker and taker fees.*

Investor Equilibrium Decision Rules. An investor submits an order to buy if, conditional on his information *and* on the submission of his order, his expected profits are non-negative. Moreover, conditional on the decision to trade, an investor chooses the order type that maximizes his expected profits. An investor abstains from trading if he expects to make negative profits from all order types.

Expressions (11)-(12) illustrate that the period t investor payoffs, conditional on the order's execution, are determined by this investor's informational advantage with respect to the period t innovation to the fundamental value (relative to the information content revealed by the investor's order submission decision) and by the investor's private valuation of the security. Our model is stationary, and in what follows, we restrict attention to investor decision rules that are independent of the history but are *solely* governed by an investor's private valuation and his knowledge of the innovation to the security's value.

When the decision rules in period t are independent of the history H_t, the public expectation of the period t innovation, conditional on the period t investor's action, does not depend on the history either. Expressions (11)-(12) reveal that neither do investor equilibrium payoffs. Our setup is thus internally consistent in the sense that the assumed stationarity of the investor decision rules does not preclude investors from maximizing their payoffs.

Expected payoffs of a period t investor are affected by the realizations of his private value y_t and the innovation δ_t only through the sum of this investor's realized private value y_t and his expectation of δ_t, conditional on the period t investor's information. We thus focus on decision rules with respect to this sum, which we refer to it as the *aggregate valuation*, and we denote the period t investor's aggregate valuation by

$$z_t = y_t + \delta_t \tag{13}$$

The aggregate valuation z_t is symmetrically distributed on the interval $[-2, 2]$.

2.2 Equilibrium Characterization

We first derive properties of market and limit orders that must hold in equilibrium.

Our setup is symmetric, and we focus on decision rules that are symmetric around the zero aggregate valuation, $z_t = 0$. We focus on equilibria where investors use both limit and market orders.[18] Appendix A establishes the following result on the market's reaction to market and limit orders.

Lemma 1 (Informativeness of Trades and Quotes) *In an equilibrium where investors use both limit and market orders, both trades and investors' limit orders contain*

[18] Any equilibrium where low-latency liquidity providers are the only liquidity providers closely resembles equilibria in market maker models in the tradition of Glosten and Milgrom (1985). In such an equilibrium, trading roles are pre-defined and maker-taker fees have no economic impact.

information about the security's fundamental value; a buy order increases the expectation of the security's value and a sell order decreases it.

Lemma 1 implies that a price improvement stemming from a period t investor's limit buy order at the equilibrium price $\mathsf{bid}^*_{t+1} > \mathsf{bid}^*_t$ increases the expectation of a security's value. In our setting, such a buy order will be immediately followed by a cancellation of a sell limit order at the best period t price ask^*_t and a placement of a new sell limit order at the new ask price $\mathsf{ask}^*_{t+1} > \mathsf{ask}^*_t$ by a low-latency liquidity provider.

Lemma 2 (Equilibrium Market and Limit Order Submission) *In any equilibrium with symmetric time-invariant strategies, investors use threshold strategies: investors with the most extreme aggregate valuations submit market orders, investors with moderate aggregate valuations submit limit orders, and investors with aggregate valuations around 0 abstain from trading.*

To understand the intuition behind Lemma 2, observe first that, conditional on order execution, an investor's payoff is determined, loosely, by the advantage that his aggregate valuation provides relative to the information revealed by his order (see expressions (11)-(12)). Second, since market orders enjoy guaranteed execution, whereas limit orders do not, for limit orders to be submitted in equilibrium, the payoff to an executed limit order must exceed that of an executed market order. Consequently, the public expectation of the innovation δ_t, conditional on, say, a limit buy order in period t, must be smaller than the corresponding expectation, conditional on a market buy order in period t (in other words, the price impact of a limit buy order must be smaller than that of a market buy order). For this ranking of price impacts to occur, investors who submit limit orders must, on average, observe lower values of the innovation than investors who submit market buy orders. With symmetric distributions of both, the innovations and investor private values, we arrive at the previous lemma.

2.3 Equilibrium Existence

Utilizing Lemmas 1 and 2, we look for threshold values z^M and $z^L < z^M$ such that investors with aggregate valuations above z^M submit market buy orders, investors with aggregate valuations between z^L and z^M submit limit buy orders, investors with aggregate valuations between $-z^L$ and z^L abstain from trading. Symmetric decisions are taken for orders to sell. Investors with aggregate valuations of z^M and z^L are marginal, in the sense that the investor with the valuation z^M is indifferent between submitting a market buy order and a limit buy order, and the investor with the valuation z^L is indifferent between submitting a limit buy order and abstaining from trading. Using (11)-(12), and the definition of the aggregate valuation (13), thresholds z^M and z^L must solve the following equilibrium conditions

$$z^M - \mathsf{E}[\delta_t \mid \mathrm{MB}_t^*] = \mathsf{Pr}(\mathrm{MS}_{t+1}^*) \times \left(z^M - \mathsf{E}[\delta_t \mid \mathrm{LB}_t^*]\right) \tag{14}$$

$$z^L = \mathsf{E}[\delta_t \mid \mathrm{LB}_t^*] \tag{15}$$

where the stationarity assumption on investors' decision rules allows us to omit conditioning on the history H_t; MB_t^* denotes a market buy order in period t, which occurs when the period t investor aggregate valuation z_t is above z^M ($z_t \in [z^M, 2]$), LB_t^* denotes a limit buy order in period t ($z_t \in [z^L, z^M)$), and MS_{t+1} denotes a market order to sell in period $t+1$ ($z_{t+1} \in [-2, -z^M]$). Given thresholds z^M and z^L, these expectations and probabilities are well-defined and can be written out explicitly, as functions of z^M and z^L (and independent of the period t).

Further, when investors use thresholds z^M and z^L to determine their decision rules, the bid and ask prices that yield zero profits to low-latency liquidity providers, given by the expressions in (5)-(6), can be expressed as

$$\mathsf{bid}^*_t = p_{t-1} + \mathsf{E}[\delta_t \mid z_t \leq -z^M] + f \quad (16)$$

$$\mathsf{ask}^*_t = p_{t-1} + \mathsf{E}[\delta_t \mid z_t \geq z^M] - f \quad (17)$$

where $p_{t-1} \equiv \mathsf{E}[V_{t-1}|H_t]$. The choice of notation for the public expectation of the security's value recognizes that this expectation coincides with a transaction price in period $t-1$ (when such a transaction occurs). Since the innovations are distributed symmetrically around 0, the public expectation of the period t value of the security at the very beginning of period t, $\mathsf{E}[V_t|H_t]$, also equals p_{t-1}.

Expanding the above expressions one step further, for completeness, investors who submit limit orders to buy and sell in period t, in equilibrium, will post them at prices bid^*_{t+1} and ask^*_{t+1}, respectively, given by

$$\mathsf{bid}^*_{t+1} = p_{t-1} + \mathsf{E}[\delta_t \mid z_t \in [z^L, z^M)] + \mathsf{E}[\delta_{t+1} \mid z_{t+1} \leq -z^M] + f \quad (18)$$

$$\mathsf{ask}^*_{t+1} = p_{t-1} + \mathsf{E}[\delta_t \mid z_t \in (-z^M, -z^L]] + \mathsf{E}[\delta_{t+1} \mid z_{t+1} \geq z^M] - f \quad (19)$$

For an equilibrium to exist, we require that the bid-ask spread is positive. In the absence of fees, the bid-ask spread is positive as long as market orders are informative. When $f \neq 0$, however, this is no longer the case. Equations (16)-(17) imply that in the equilibrium where all pay maker-taker fees, $\mathsf{ask}^*_t - \mathsf{bid}^*_t > 0$ if and only if

$$f < \mathsf{E}[\delta_t \mid z_t \geq z^M] \quad (20)$$

Finally, as discussed above, the equilibrium is supported by out-of-the-equilibrium beliefs such that low-latency liquidity providers outid all non-competitive prices. When the bid-ask spread is sufficiently wide, however, a low-latency liquidity provider may not be able to outbid an investor with a sufficiently high aggregate valuation (above the highest possible value of the innovation), and the equilibrium may not exist. Specifically, an equilibrium where low-latency liquidity providers outbid all non-competitive prices exists, provided that the cum-fee spread is bounded by the highest possible value of innovation δ_t, which is 1 in our setting:[19]

$$2 \cdot \mathsf{E}[\delta_t \mid z_t \geq z^M] - 2f \leq 1 \tag{21}$$

We prove the following existence theorem in Appendix A:[20]

Theorem 1 (Equilibrium Characterization and Existence) *There exist values z^M and z^L, with $0 < z^L < z^M < 2$, that solve indifference conditions (14)-(15). These threshold values constitute an equilibrium in a setting where investors pay maker-taker fees on a per-trade basis, for any history H_t, given competitive equilibrium prices, bid_t^* and ask_t^* in (16)-(17), for the following trader decision rules, if conditions and (21) and (20) are satisfied. The investor who arrives in period t with aggregate valuation z_t*

- *places a market buy order if $z_t \geq z^M$,*
- *places a limit buy order at price bid_{t+1}^* if $z^L \leq z_t < z^M$,*
- *abstains from trading if $-z^L < z_t < z^L$.*

Investors' sell decisions are symmetric to buy decisions.

3 Equilibrium: Investors Pay Flat Fees

We now study the market where investors pay only the average exchange fee, through a flat fee per trade. Long-term investors typically trade through a broker, and the flat-fee setting reflects a common practice by brokers of levying a flat fee per trade on their clients. Since the limit order book is always full, the period t investor's market

[19]This assumption effectively restricts the aggregate amount of the information in the model; model assumption different to the current version (e.g., a restriction on the fraction of traders who become informed) may lead to a similar outcome.

[20]Appendix A further provides the out-of-the-equilibrium beliefs that support the equilibrium prices and decision rules, described in Theorem 1.

order will incur the taker fee f with certainty, and a period t investor's limit order to buy (sell) will receive the maker rebate f, provided that a market order to sell (buy) is submitted in period $t+1$. The expected fee $\bar{f}_t$ that the exchange receives from the period t investor, conditional on the execution of this order, is then given by

$$\bar{f}_t = \frac{f \cdot [\mathsf{Pr}(\mathrm{MB}_t^*) + \mathsf{Pr}(\mathrm{MS}_t^*)] + (-f) \cdot [\mathsf{Pr}(\mathrm{LB}_t^*) \cdot \mathsf{Pr}(\mathrm{MS}_{t+1}^*) + \mathsf{Pr}(\mathrm{LS}_t) \cdot \mathsf{Pr}(\mathrm{MB}_{t+1}^*)]}{\mathsf{Pr}(\mathrm{MB}_t^*) + \mathsf{Pr}(\mathrm{MS}_t^*) + \mathsf{Pr}(\mathrm{LB}_t^*) \cdot \mathsf{Pr}(\mathrm{MS}_{t+1}^*) + \mathsf{Pr}(\mathrm{LS}_t^*) \cdot \mathsf{Pr}(\mathrm{MB}_{t+1}^*)} \tag{22}$$

where LB_t^* and MB_t^* denote the period t investor's market and limit orders to buy at the equilibrium bid and ask prices; likewise for the sell orders and orders in period $t+1$.

As in Section 2, we focus on an equilibrium where investors use stationary, time-invariant threshold strategies with respect to their aggregate valuation $z_t = y_t + \delta_t$. Because innovations δ_t to the security's value and investor private valuations y_t are identically and independently distributed across time, probabilities of market and limit orders to buy and to sell are time-invariant. We continue to focus on a symmetric equilibrium, where investors decisions to buy and sell are symmetric with respect to the aggregate valuation $z_t = 0$, so that the probability of a market buy order then equals the probability of a market sell order; likewise for limit orders. Consequently, the expected per-investor fee does not depend on period t. Denoting this fee by $\bar{f}$ and writing $\mathsf{Pr}(\mathrm{LB}^*)$ for the probability of a limit (buy) order in equilibrium, we simplify (22) to

$$\bar{f} = \frac{1 - \mathsf{Pr}(\mathrm{LB}^*)}{1 + \mathsf{Pr}(\mathrm{LB}^*)} \cdot f \tag{23}$$

Since low-latency liquidity providers receive maker rebates and act competitively, limit order book prices are determined by the same conditions as in the all pay maker-taker fees setting (conditions (7)-(8)). Investor payoffs, however, are affected by the flat fee $\bar{f}$. With the decision rules being stationary, these payoffs are given by

$$\pi^{\mathrm{MB}}(y_t, \delta_t) \;=\; y_t + \delta_t - (\mathsf{E}[\delta_t \mid \mathrm{MB}_t^*] - f) - \bar{f} \tag{24}$$

$$\pi^{\mathrm{LB}}(y_t, \delta_t) \;=\; \mathsf{Pr}(\mathrm{MS}_{t+1}^* \mid \mathrm{LB}_t^*)\left(y_t + \delta_t - (\mathsf{E}[\delta_t \mid \mathrm{LB}_t^*] + f) - \bar{f}\right) \tag{25}$$

where LB_t^* and MB_t^* denote investors' limit and market buy orders at the equilibrium competitive prices; the stationarity of investor decision rules allows us to drop the dependence on the history. Substituting in the expression (23) for the flat fee, we obtain

$$\pi^{\mathrm{MB}^*}(y_t, \delta_t) \;=\; y_t + \delta_t - \mathsf{E}[\delta_t \mid \mathrm{MB}_t^*] + \frac{2\mathsf{Pr}(\mathrm{LB}_t^*)}{1 + \mathsf{Pr}(\mathrm{LB}_t^*)} \cdot f \tag{26}$$

$$\pi^{\mathrm{LB}^*}(y_t, \delta_t) \;=\; \mathsf{Pr}(\mathrm{MS}_{t+1} \mid \mathrm{LB}_t^*)\left(y_t + \delta_t - \mathsf{E}[\delta_t \mid \mathrm{LB}_t^*] - \frac{2}{1 + \mathsf{Pr}(\mathrm{LB}_t^*)} \cdot f\right) \tag{27}$$

Equations (26)-(27) illustrate, in particular, that when only investors pay a flat fee per trade, their payoffs are affected by the maker (or taker) fee beyond the effect of the total exchange fee. The split between the taker fee and the maker rebate will thus be economically relevant in this setting.

3.1 Equilibrium Characterization

Expression (23) illustrates that the flat fee coincides with the sign of the maker rebate. In particular, when the maker rebate is positive, brokers always set a positive flat fee (despite the zero total fee). The presence of low-latency liquidity providers ensures that market orders always execute, whereas limit orders only execute when another investor submits a market order. Low-latency liquidity providers must capture a fraction of the maker rebates, leaving investors to pay a positive exchange fee.

Lemma 3 (Flat Fee) *The average exchange fee per investor trade $\bar{f}$ is positive when the maker rebate is positive, and it is negative when the maker rebate is negative.*

Our further results on the flat fee setting are numerical. We employ the following family of distributions of the innovation parameter δ_t, for $\alpha \geq 1$.[21]

$$\bar{g}(\delta, \alpha) = \begin{cases} \frac{(1-\delta)^{(\alpha-1)}}{\alpha} & \text{if } \delta \geq 0 \\ \frac{(1+\delta)^{(\alpha-1)}}{\alpha} & \text{if } \delta \leq 0 \end{cases} \tag{28}$$

We numerically search for an equilibrium, with properties similar to those in Section 2. Specifically, we look for an equilibrium where investors use threshold rules that are symmetric and that do not depend on the history, such that investors with most extreme aggregate valuations trade with market orders, investors with moderate aggregate valuations trade with limit orders, and investors with aggregate valuations around 0 abstain from trading. The equilibrium indifference conditions are analogous to conditions (14)-(15), except that they are adjusted for the exchange fees, using (26)-(27):[22]

$$\begin{aligned} z^M - \mathsf{E}[\delta_t \mid \text{MB}_t^*] + f - \bar{f} &= \mathsf{Pr}(\text{MS}_{t+1}^*) \left(z^M - \mathsf{E}[\delta_t \mid \text{LB}_t^*] - f - \bar{f} \right) \qquad (29) \\ z^L &= \mathsf{E}[\delta_t \mid \text{LB}_t^*] + f + \bar{f} \end{aligned}$$

where the flat fee $\bar{f}$ is given by (23).

4 Impact of Maker-Taker Fees

We analyze the impact of an increase in the maker rebate (and the taker fee), measured by an increase in f, on quoted and cum-fee bid-ask spreads, trading volume, and market participation. The quoted bid-ask spread is the difference between the ask and bid prices. The cum-fee spread additionally accounts for the fee paid by a submitter of a market order; this fee is the taker fee in the all pay maker-taker fees setting and the

[21] Density $2\bar{g}$ is a Beta-distribution on [0,1].

[22] Numerically, the solution is always unique. If it were not unique, we would focus on the one that delivers the smallest bid-ask spread in equilibrium.

flat fee $\bar{f}$ in the flat fee setting. We measure market participation by the probability that an investor does not abstain from submitting an order, and we measure trading volume by the probability that an investor submits a market order (since market orders always execute in our setting).

Theorem 1 implies the following result for the setting where all market participants pay maker-taker fees per-trade.

Corollary 1 (Impact of Maker-Taker Fees: All Pay Maker Taker Fees) *In an equilibrium of the symmetric fee setting, thresholds z^M and z^L, market participation, trading volume, and cum-fee bid-ask spreads are independent of f. Quoted bid ask-spreads decline in f.*

Trading Volume and Market Participation. Equations (24)-(25), which define investor payoffs in the flat fee setting, illustrate that, ceteris paribus, an increase in the maker rebate provides investors with incentives to switch from limit to market orders. All else equal, such an increase will decrease the spread, thus increasing the payoff to market orders and simultaneously reducing the payoff to limit orders. In contrast to the all pay maker-taker fees setting, however, changes in the bid-ask spread are not offset by the changes in investor fees — because the flat fee charged by brokers does not depend on the order type. Since trade occurs in our model when a market order is submitted, an increase in the probability of a market order implies an increase in trading volume.

The impact on investors who were previously indifferent between submitting a limit order and abstaining from trading is more complex. On the one hand, ceteris paribus, as traders increase their usage of market orders, limit orders are submitted by less informed traders, the price impact of a limit order declines, and limit orders become more attractive. On the other hand, an increase in the maker rebate leads to a decline in the bid-ask spread, making limit order prices less attractive to investors who do not receive the rebate. Numerical simulations reveal that the latter effect dominates in our setting; that is, market participation declines.

What happens when the maker rebate is very large? As the taker fee and the maker rebate increase, threshold z^M decreases and threshold z^L increases. When the maker rebate is sufficiently high (relative to the spread), a limit order yields negative profits to investors in expectation, because they do not receive maker rebates. When this happens, low-latency liquidity providers become the only submitters of limit orders, while investors trade exclusively with market orders. As a consequence, the flat fee equals the taker fee. The marginal submitter of a market order is then exactly indifferent between submitting a market order and abstaining from trading, and he earns zero expected profits. We denote the aggregate valuation of such a marginal submitter by z_0, and the value of f that yields $z^M = z^L = z_0$ in equilibrium by f_0. Using investor payoffs, given by expressions (24)-(25), together with $\bar{f} = f$, we find that z_0 solves

$$z_0 - \mathsf{E}[\delta_t | z_t \geq z_0] = 0 \tag{30}$$

A further increase in the maker rebate (above f_0) then leads to a further decline in the quoted spread but does not have an effect on investors payoffs, because a decline in the quoted spread is exactly offset by an increase in the average fee, which equals the taker fee. As with the all pay maker-taker fees setting, an equilibrium fails to exist when the maker rebate is so large that the bid-ask spread becomes nonpositive. Similarly to condition (21) for the setting where all market participants pay maker-taker fees per-trade, the bid-ask spread remains positive for fees f that are below value f_1 that solves

$$f_1 = \mathsf{E}[\delta_t \mid z_t \geq z_0^M] \tag{31}$$

What happens when the maker rebate is negative? When $f < 0$, i.e. limit order submitters pay a positive fee for executed orders, whereas market order submitters receive a positive taker rebate, liquidity providers offer less than the expected value of the security when buying and they demand more than the expected value when selling the security. Consequently, as the maker fee $(-f)$ increases from 0, quoted spreads widen. Investors pay a flat fee (in this case, the fee is negative, so they receive a flat positive rebate), therefore market orders become less attractive to them and limit orders become more attractive.

Intuitively, when the maker fee is positive and high (f is low and negative), the bid ask spread becomes too wide, market orders earn negative profits for all investors (even after accounting for the positive flat rebate that investors receive on each transaction), and trade does not occur. The equilibrium of the model relies on the ability of low-latency liquidity providers to compete with investors. In the present setting, the spread will widen to the point where a low-latency liquidity provider is unable to outbid investors with sufficiently high valuations without making a loss. In the present version of the paper, we thus restrict attention to a positive maker rebate.

Figure 2 illustrates the following observation on order submission decisions.

Numerical Observation 1 (Fee Thresholds and Equilibrium Actions: Flat Fee) *There exist f_0, f_1, with $0 < f_0 < f_1$, such that in the flat fee setting*

(i) investors submit both market and limit orders in equilibrium with $f < f_0$;

t (ii) investors submit only market orders in equilibrium when $f_0 \leq f < f_1$;

(iii) a stationary equilibrium with trade does not exist when $f \geq f_1$.

Threshold f_0 is the value of f that yields solutions $z^M = z^L = z_0^M$ to equations (29), and threshold f_1 solves (31).

Figure 3 illustrates the following observation on probabilities of order submissions and the implications for trading volume and market participation.

Numerical Observation 2 (Volume and Market Participation: Flat Fee) *As the maker rebate f increases, for $0 \leq f \leq f_0$, the probability that an investor* *(i)* *submits a market order increases (trading volume increases);*

(ii) *submits a limit order decreases;*

(iii) *abstains from trading (weakly) increases (market participation declines).*

These probabilities do not depend on f when $f_0 < f < f_1$ (i.e., when the maker rebate is sufficiently large and investors trade only with market orders).

Quoted Bid-Ask Spread. As the maker rebate increases (f increases), more investors submit market orders, that is they submit aggressive orders for lower values of the innovations δ_t. Furthermore, as f increases, the bid-ask spread declines because low-latency liquidity providers compete the benefits of the increased rebate away. Both of these effects lead to a decline in the quoted bid-ask spread.

Cum-Fee Bid-Ask Spread. The cum-fee spread accounts for the fee that an investor pays to his broker:

$$\textit{cum-fee spread} = \mathsf{ask}_t^* - \mathsf{bid}_t^* + 2\bar{f} \tag{32}$$

where the factor 2 accounts for the fact that the bid-ask spread is a cost of a round-trip transaction, so that the fee is paid twice. As the maker rebate increases (f increases), the probability of a limit order declines, and expression (23) reveals that $\bar{f}$ increases as long as $f < f_0$. Numerically, this increase is more than offset by the decline in the quoted spread, so that the cum-fee spread declines. Figure 4 illustrates the following observation

Numerical Observation 3 (Quoted and Cum-Fee Spreads: Flat Fee) *As the taker fee and the maker rebate increase (f increases), for $0 < f < f_1$,*

(i) *the quoted bid-ask spread declines;*

(ii) *the broker flat fee $\bar{f}$ increases;*

(iii) *the cum-fee spread declines for $f < f_0$, and it is independent of f for $f \geq f_0$ (when investors trade only with market orders).*

Price Impact. The price impact of a trade measures the change in the public expectation following the execution of a trade. In our model, this change is determined, loosely, by the information content of market orders about the time-t innovation δ_t. Specifically, the price impact of a buyer-initiated transaction is given by:

$$\text{price impact}_{\text{buy},t} = \mathsf{E}[V_t \mid \text{MB}_t] - p_{t-1} = \mathsf{E}[\delta_t \mid \text{MB}_t] \tag{33}$$

Using expression (17) for the equilibrium ask price ask_t^*, we find that for a positive maker rebate ($f > 0$) the price impact of a trade is higher than indicated by a transaction price:

$$\text{price impact}_{\text{buy},t} = \mathsf{E}[\delta_t \mid \text{MB}_t] = \mathsf{ask}_t^* + f - p_{t-1} > \mathsf{ask}_t^* - p_{t-1} \tag{34}$$

Figure 5 illustrates the above relation between the quoted half-spread, $\mathsf{ask}_t^* - p_{t-1}$.

Numerical Observation 2 illustrated, in particular, that as the taker fee f increases, the marginal submitter of a market order requires a lower aggregate valuation. Market orders are then submitted for lower absolute values of realizations of the innovations δ_t. This insight explains the following numerical observation, illustrated by Figure 5.

Numerical Observation 4 (Price Impact: Flat Fee) *The price impact of a trade is decreasing in the level of the maker rebate f on $[0, f_0]$, and constant on $(f_0, f_1]$ (when investors trade only with market orders).*

Numerical Observation 4 is supported empirically by Malinova and Park (2011).

Welfare. Each investor in our setting has a private valuation for the security, and we follow Bessembinder, Hao, and Lemmon (2012) to define a social welfare measure that reflects allocative efficiency. Specifically, we define welfare as the expected gain from trade in the market for a given period t. If a transaction occurs in period t, then the welfare gain is given by the private valuation of a buyer, net of the exchange fee paid by the buyer, minus the private valuation of a seller, net of the exchange fee paid by the seller.

A transaction in period t occurs when the period t investor submits a market buy or a market sell order. Focusing on a submitter of a buy market order: this investor trades with the period $t-1$ investor if the period $t-1$ investor submitted a limit sell order and he trades with a low-latency liquidity provider otherwise. With a flat fee set to equal the average fee paid by an investor, the expected aggregate fee on each transaction is zero. Accounting for the fact that a low-latency liquidity supplier has a zero private valuation, by symmetry, we obtain the following expression for the welfare:

$$W_t = 2 \cdot \mathsf{Pr}(MB_t) \left(\mathsf{E}[y_t \mid H_t, \mathrm{MB}_t] - \mathsf{Pr}(\mathrm{LS}_{t-1}) \cdot \mathsf{E}[y_{t-1} \mid H_{t-1}, \mathrm{MB}_t, \mathrm{LS}_{t-1}] \right) \tag{35}$$

Theorem 1 implies the following result for the case where investors pay maker-taker fees on a trade-by-trade basis.

Corollary 2 (Social Welfare: All Pay Maker-Taker Fees) *In a setting where investors pay maker-taker fees on a per-trade basis, expected total welfare W_t is not affected by the split of the total exchange fee into a maker rebate and a taker fee.*

When the maker fee increases (and the taker fee increases by the same amount), quoted spread narrows, and two changes happen. First, some investors switch from submitting limit orders to trading with market orders, increasing the execution probability of their own order (to certainty), and also increasing the execution probability of a limit order, the so-called fill rate, for the remainder of the limit order submitters. Second, some investors switch from submitting limit orders to abstaining from trade, failing to realize any potential gains from trade. Figure 6 and Numerical Observation 5 illustrate that the benefit of an increased fill rate to investors who remain in the market exceeds the loss of potential gains from trade to investors who choose to leave the market.

Numerical Observation 5 (Social Welfare: Flat Fee) *In a setting where investors pay a flat fee per trade, expected total welfare* W_t *is increasing in the level of the maker rebate* f *on* $[0, f_0]$*, and it is constant in* f *on* $(f_0, f_1]$ *(when investors trade only with market orders).*

In a world where exchange maker-taker fees are only passed through on average, our results suggest that positive maker rebates have a positive effect on social welfare. Allocative efficiency is highest when investors only trade with market orders (or not at all). An implication of our result on social welfare is that it is socially beneficial for investors and low-latency liquidity providers to specialize: investors submitting market orders, and; low-latency liquidity providers providing liquidity.

The intuition for the welfare increase stems from the increased frequency of investors realizing their gains from trade. Low-latency liquidity providers do not have any private values and thus do not receive any gains from trade. Since these traders ensure that the limit order book is always full, liquidity provision by investors is socially inefficient. An investor who submits a limit order faces the risk of non-execution for his own limit order and also possibly imposes a negative externality on the previous period investor if that investor's limit order did not get executed. As discussed above, in the presence of low-latency liquidity providers, the flat fee is positive but below the taker fee. Under the flat fee structure, market orders incur the taker fee but their submitters pay the lower flat fee, whereas limit orders deliver the rebate, yet their submitters pay the fee. Flat fee thus effectively works as a tax that redistributes the exchange fee among investors with heterogeneous valuations and mitigates the aforementioned externality.

5 Conclusion

We provide a model to analyze a financial market where investors trade for informational and liquidity reasons in a limit order book that is monitored by low-latency liquidity providers. Methodologically, price competition between informed investors and uninformed low-latency market makers is a key feature of our model; it allows us study the impact of exogenous market factors on the tradeoffs between market and limit orders.

We apply our model to study the impact of maker-taker fees. When all traders pay the maker-taker fees, investor behavior is affected only through the total fee charged by the exchange (the taker fee minus the maker rebate), consistent with Colliard and Foucault (2012). When, however, we study the most common implementation of the maker-taker fees through a flat fee levied on investors (e.g., Fidelity cites a flat fee of $7.95 per trade on their website), the split of the total exchange fee into the maker fee and the taker fee also plays a meaningful role (even when the maker-taker fees are passed through to investors on average).

Our empirical predictions support the industry's opinions on the impact of maker-taker pricing on long-term investors. Indeed, we predict that if a positive maker rebate is introduced (financed by an increase in the taker fee), investors trade on the liquidity demanding side more frequently, submit fewer limit orders, and choose to abstain from trading more often. As investors realize their gains from trade more frequently, allocative efficiency improves. Our model also predicts an increase in the average exchange fee that a broker incurs when executing client orders, consistent with industry concerns. Contrary to industry opinions, we find that trading costs for liquidity demanders decrease, because a decline in the quoted spreads more than offsets the increase in the average exchange fee. One key contributor to the decline in trading costs for liquidity demanders is the decrease in price impact of trades — as more less-informed investors trade aggressively, using market orders, trades become less informative. Malinova and Park (2011) find empirical support for our predictions.

Our results have several policy implications. First, we find that in markets where brokers charge investors a flat fee per trade, the levels of maker and taker fees have an economic effect beyond that of the total exchange fee. Our results show, in particular, that when the fee is passed through only on average, through a flat commission, investors' trading incentives are different to the situation where investors pay taker fees and receive maker rebates for each executed trade. The flat fee improves welfare by acting as a tax on the socially inefficient liquidity provision by investors and as a subsidy for investors who trade with market orders.

Second, we reiterate the importance of accounting for the exchange trading fees (see, e.g., Angel, Harris, and Spatt (2011), Colliard and Foucault (2012), or Battalio, Shkilko, and Van Ness (2012).) A lower quoted spread need not imply lower trading costs for investors, and consequently, routing orders to the trading venue that is quoting the best price need not guarantee the best execution.

Third, we caution that the causal relations among trading volume, trading costs, and competition for liquidity providers are more complex than the taken-at-face-value intuition would suggest. An increase in volume in our setting is driven by changes in investor trading behavior. These changes necessitate a higher rate of participation by low-latency liquidity providers, which may manifest itself empirically as an increase in competition among low-latency liquidity providers.[23] Hence, an empirically observed increase in competition need not be the driving force of changes in trading volume and trading costs.

[23]In our model, low-latency liquidity providers compete in prices; empirical assessments typically measure competition in quantities.

A Appendix

A.1 Proofs of Lemmas 1 and 2

Proof. In the main text, we present the two lemmas separately, for the ease of exposition. Here we establish the two results simultaneously. We restrict attention to an equilibrium where investors use symmetric, time-invariant strategies and trade with both, market and limit orders. Since we search for an equilibrium with competitive pricing, an investor's equilibrium action does not affect the price that he pays or the probability of his limit order execution. We show, in 5 steps, that in any such equilibrium investors must use decision rules that lead to Lemmas 1 and 2.

Step 1: In any equilibrium, an investor with the aggregate valuation z_t prefers a market (limit) buy order to a market (limit) sell order if and only if $z_t \geq 0$.

Proof: Using (11), an investor's payoff to a market buy order is $z_t - \mathsf{E}[\delta_t \mid H_t, MB_t(\mathsf{ask}_t^*)]$. When innovations δ_t are independent across time and investors' equilibrium strategies are time-invariant functions of z_t, the expectation $\mathsf{E}[\delta_t \mid H_t, MB_t(\mathsf{ask}_t^*)]$ does not depend on the history H_t or on the ask price ask_t^*. With symmetric decision rules, $\mathsf{E}[\delta_t \mid MB_t] = -\mathsf{E}[\delta_t \mid \mathrm{MS}_t]$; investor payoff (11) and an analogous payoff for sell orders then yield Step 1 for market orders. Similarly, symmetry, expression (12) and an analogous expression for limit sell orders yield the result for limit orders.

Step 2: In any equilibrium, there must exist $z^* \in (0, 2)$ such that an investor with aggregate valuation z_t prefers a market buy order to a limit buy order if and only if $z_t \geq z^*$, with indifference if and only if $z_t = z^*$.

Proof: Comparing investor equilibrium payoffs (11) and (12), an investor with valuation z_t prefers a market buy order to a limit buy order if and only if

$$z_t \geq \frac{\mathsf{E}[\delta_t \mid MB_t] - \mathsf{Pr}(\mathrm{MS}_t)\mathsf{E}[\delta_t \mid \mathrm{LB}_t]}{1 - \mathsf{Pr}(\mathrm{MS}_t)} \equiv z^*. \tag{36}$$

The fraction in (36) is well-defined in an equilibrium where investors submit both market and limit orders, since $0 < \mathsf{Pr}(\mathrm{MS}_t) < 1$. Next, for investors to submit limit orders with positive probability, there must exist z such that for the investor with the aggregate valuation $z_t = z$, the payoff to a limit buy order (i) exceeds that to the market buy order and (ii) is non-negative. For this z, we then have

$$z - \mathsf{E}[\delta_t \mid MB_t] \leq \mathsf{Pr}(\mathrm{MS}_t)(z - \mathsf{E}[\delta_t \mid \mathrm{LB}_t]) \leq z - \mathsf{E}[\delta_t \mid \mathrm{LB}_t] \tag{37}$$

Hence, $\mathsf{E}[\delta_t \mid MB_t] \geq \mathsf{E}[\delta_t \mid \mathrm{LB}_t]$. Since $0 < \mathsf{Pr}(\mathrm{MS}_t) < 1$, the following inequalities are strict: $\mathsf{E}[\delta_t \mid MB_t] > \mathsf{Pr}(\mathrm{MS}_t)\mathsf{E}[\delta_t \mid \mathrm{LB}_t]$ and $z^* > 0$.

Step 3: In any equilibrium, submitting the market buy order is strictly optimal for an investor with aggregate valuation $z_t > z^*$.

Proof: By Steps 1 and 2, an investor with valuation z_t such that $z_t > z^* > 0$ strictly prefers a market buy order to a market sell order and to a limit buy order (and, consequently, by Step 1, to a limit sell order). Finally, an investor with valuation $z_t > z^*$ strictly prefers submitting a market order to abstaining from trade, as:

$$z_t - \mathsf{E}[\delta_t \mid MB_t] > \frac{\mathsf{E}[\delta_t \mid MB_t] - \mathsf{Pr}(\mathrm{MS}_t)\mathsf{E}[\delta_t \mid \mathrm{LB}_t]}{1 - \mathsf{Pr}(\mathrm{MS}_t)} - \mathsf{E}[\delta_t \mid MB_t] \geq 0,$$

where the last inequality follows since $\mathsf{E}[\delta_t \mid \mathrm{LB}_t] \leq \mathsf{E}[\delta_t \mid MB_t]$ by Step 2.

Step 4: In any equilibrium, an optimal action for an investor with aggregate valuation $z_t \in (0, z^*)$ must be either a limit buy order or a no trade.

Proof: This investor prefers a limit buy order to a market buy order by Step 2, and The investor prefers a limit buy order to a limit sell order by Step 1, which in turn is preferred by a market sell order by symmetry and Step 2.

Step 5: There exists $z^{**} \in (0, z^*)$ such that an investor with the valuation $z_t = z^{**}$ is indifferent between submitting a limit buy order and abstaining from trade; it is strictly optimal for an investor with valuation $z_t \in (z^{**}, z^*)$ to submit a limit buy order, and it is strictly optimal for an investor with valuation $z_t \in [0, z^{**})$ to abstain from trading.

Proof: In an equilibrium where investors submit both market and limit orders the probability of a limit order is strictly positive, consequently, the limit buy order is preferred to abstaining from trade if and only if an investor's valuation $z_t > \mathsf{E}[\delta_t \mid \mathrm{LB}_t]$ (and, by Step 4, the limit buy order is then the optimal action for this investor, and abstaining from trade is optimal for an investor with $z_t < \mathsf{E}[\delta_t \mid \mathrm{LB}_t]$). For investors to submit both market and limit orders with non-zero probability, in equilibrium we must have $\mathsf{E}[\delta_t \mid \mathrm{LB}_t] < z^*$ (otherwise, by Step 3, any investor, except for the zero-probability case of $z_t = z^*$ that prefers the limit order to abstaining from trade also strictly prefers the market buy order to the limit buy order). We are looking for a stationary equilibrium and the distribution of δ_t does not depend on t, hence $\mathsf{E}[\delta_t \mid \mathrm{LB}_t]$ does not depend on t and we can thus set $z^{**} = \mathsf{E}[\delta_t \mid \mathrm{LB}_t]$.

What remains to be shown is that $\mathsf{E}[\delta_t \mid \mathrm{LB}_t] > 0$. We proceed by contradiction. Suppose not and $\mathsf{E}[\delta_t \mid \mathrm{LB}_t] \leq 0$. Then, by Steps 1-4, in a symmetric equilibrium, the limit buy is strictly optimal for an investor with $z \in (0, z^*)$; it is strictly optimal for an investor with $z > z^*$ to submit the market buy order; it is strictly optimal for an investor with valuation $z_t < 0$ to submit either the market or the limit sell orders; finally, investors with $z_t = 0$ and $z_t = z^*$ are indifferent between the limit buy and a different action (the limit sell and the market buy, respectively) and they occur with zero probability. This implies that limit buy orders are only submitted by investors whose aggregate valuations are (weakly or strictly) in the interval of $[0, z^*]$ and only by these investors. But then $\mathsf{E}[\delta_t \mid \mathrm{LB}_t] = \mathsf{E}[\delta_t \mid z_t \in (0, z^*)] > 0$, a contradiction.[24]

[24]The inequality follows because $z = y_t + \delta_t$, where y_t and δ_t are independent and symmetrically distributed on $[-1, 1]$; the explicit derivation of this expectation is in the Internet Appendix.

Steps 1-5 show that threshold rules are optimal in any symmetric, time-invariant equilibrium where traders submit both market and limit orders, and that investors with the more extreme valuations submit market orders, investors with moderate valuations submit limit orders, and investors with valuations close to zero abstain from trade. Given threshold rules described in these steps, (investors') quotes are informative because $\mathsf{E}[\delta_t \mid \mathrm{LB}_t] = E[\delta_t \mid z \in (z^{**}, z^*)] > 0$ and trades are informative because $\mathsf{E}[\delta_t \mid MB_t] = \mathsf{E}[\delta_t \mid z \in (z^*, 2)] > 0$. Furthermore, by the proof of Step 2, a trade has a higher price impact than a quote. ■

A.2 Proof of Theorem 1

A.2.1 Preliminary Notation and Results

Equilibrium thresholds solve equations ((14)-(15)). An informed investor submits a market buy over a limit buy as long as $z_t \geq z^M$, submits a limit buy if $z^M > z_t \geq z^L$, and abstains from trading otherwise. To show existence of a threshold equilibrium, we need to show existence of thresholds z^M and z^L and prove the optimality of investor strategies. Given symmetric threshold decision rules and stationarity of the equilibrium, equilibrium conditions ((14)-(15)) can be rewritten as:

$$z^M - \mathsf{E}[\delta_t \mid z_t \in [z^M, 2]] = \mathsf{Pr}(z_t \in (z^M, 2])(z^M - \mathsf{E}[\delta_t \mid z_t \in [z^L, z^M)]), \tag{38}$$

$$z^L = \mathsf{E}[\delta_t \mid z_t \in [z^L, z^M)]. \tag{39}$$

In what follows, we omit subscript t and use the following short-hand notion for the expectations of the innovation δ_t and the probability of a limit order execution:

$$\mathsf{E}^M\delta := \mathsf{E}[\delta_t \mid z_t \in [z^M, 2]]; \tag{40}$$

$$\mathsf{pr}^M := \mathsf{Pr}(z_t \in (z^M, 2]); \tag{41}$$

$$\mathsf{E}^L\delta := \mathsf{E}[\delta_t \mid z_t \in [z^L, z^M)]). \tag{42}$$

With slight abuse of notation, we view $\mathsf{E}^M\delta, \mathsf{E}^L\delta$, and pr^M as functions of z^L and z^M.

In the main text, we assume that the innovation δ_t is distributed on [-1,1], symmetrically around 0, according to the density function $\bar{g}$. By symmetry, we can define g as the density function on $[0,1]$ such that $\bar{g}(\cdot) = g(\cdot)/2$ on $[0,1]$ and g is decreasing; denote the corresponding distribution function by G. Since g is decreasing, we obtain the following lemma by direct computation (details will be available in the Internet Appendix):

Lemma 4 *Function g satisfies:* $(1 - G(\delta))(1 - \delta) < g(\delta) < G(\delta)/\delta$, *for all* $\delta \in (0, 1)$.

A.2.2 Proof Outline

We proceed in 4 steps. In step 1, we show that for any given $z^M \in [0, 3/4]$ there exists the unique z^L that solves (39).[25] We denote this solution by $z^L_*(z^M)$ and show, in Step 2, that $z^L_*(z^M)$ is increasing in z^M. In Step 3, we show that there exists z^M that solves

$$z^M - \mathsf{E}^M\delta = \mathsf{pr}^M(z^M - z^L_*(z^M)). \tag{43}$$

Finally, in Step 4, we show the optimality of the strategies and discuss out-of-equilibrium beliefs that support these strategies in a perfect Bayesian equilibrium.

A.2.3 Step 1: Existence and Uniqueness of $z^L_*(z^M)$

We first derive the expression for $\mathsf{E}^L\delta$ in terms of the model primitives:

$$\mathsf{E}^L\delta = \frac{\int_{-1}^{1} d\delta \int_{-1}^{1} dy(\delta \cdot h^L(\delta, y|\mathrm{LB}))}{\int_{-1}^{1} d\delta \int_{-1}^{1} dy(h^L(\delta, y|\mathrm{LB}))}, \tag{44}$$

[25] Threshold $3/4$ may seem arbitrary, but we can also show, for decreasing g, that there does not exist $z^M > 3/4$ that solves (38) for $\mathsf{E}^L\delta \geq 0$ (which, by the proof of Lemmas 1 and 2 must hold in any symmetric stationary equilibrium where investors use both market and limit orders). The proof will be provided in the Internet Appendix. The idea is that when z^M is high, the price impact of a market order is The bound is derived using the uniform density;

where function $h^L(\delta, y|\text{LB})$ is defined as follows:

$$h^L(\delta, y|\text{LB}) = \begin{cases} \frac{1}{2} \cdot \bar{g}(\delta), \text{ if } \delta \in [z^L - 1, 1] \text{ and } y \in [z^L - \delta, z^M - \delta] \\ 0, \text{ otherwise.} \end{cases} \quad (45)$$

The denominator of (44) equals the probability of a limit buy order submission pr^L, and we use $\mathsf{num}(\mathsf{E}^L\delta)$ to denote the numerator. Using this notation, we then have $\mathsf{E}^L\delta = \mathsf{num}(\mathsf{E}^L\delta)/\mathsf{pr}^L$. Substituting h^I in, putting in appropriate integral bounds and expressing f as a function of g, we express the probability of a limit buy as follows:

$$\mathsf{pr}^L = \gamma^L \cdot (z^M - z^L) - \frac{1}{4} \int_{1-z^M}^{1-z^L} (\delta - (1 - z^M))g(\delta)d\delta,$$

where $\gamma^L := (1 + G(1 - z^L))/4$. Probability pr^L can also be expressed as

$$\mathsf{pr}^L = \gamma^M \cdot (z^M - z^L) + \frac{1}{4} \int_{1-z^M}^{1-z^L} (1 - z^L - \delta)g(\delta)d\delta, \quad (46)$$

where $\gamma^M := 1/4(1+G(1-z^M))$. The numerator of the $\mathsf{E}^L\delta$ function can be expressed as

$$\mathsf{num}(\mathsf{E}^L\delta) = -\frac{1}{4} \int_{1-z^M}^{1-z^L} \delta(1 - z^L - \delta)g(\delta)d\delta + \frac{1}{4}(z^M - z^L) \int_{1-z^M}^{1} \delta g(\delta)d\delta, \quad (47)$$

where we used identity $z^M - z^L = (1 - z^L - \delta) + (z^M - 1 + \delta)$. Define β^L and β^M as

$$\beta^L = -\frac{\partial \mathsf{num}(\mathsf{E}^L\delta)}{\partial z^L} = \frac{1}{4} \int_{1-z^L}^{1} \delta g(\delta)d\delta \text{ and } \beta^M = \frac{\partial \mathsf{num}(\mathsf{E}^L\delta)}{\partial z^M} = \frac{1}{4} \int_{1-z^M}^{1} \delta g(\delta)d\delta \quad (48)$$

Lemma 5 (Bounds on $\mathsf{E}^L\delta$) *Expectation* $\mathsf{E}^L\delta$ *satisfies* $\mathsf{E}^L\delta < z^M/2$.

Proof sketch: We use the following bounds on expressions for $\mathsf{num}(\mathsf{E}^L\delta)$ and pr^L:

$$\mathsf{num}(\mathsf{E}^L\delta) \leq \frac{1}{4}(z^M - z^L) \int_{1-z^M}^{1} \delta g(\delta) d\delta \text{ and } \mathsf{pr}^L \geq \gamma^M \cdot (z^M - z^L)$$

to obtain

$$\mathsf{E}^L\delta \leq \frac{(1 - G(1 - z^M)\mathsf{E}[\delta|\delta \geq 1 - z^M]}{1 + G(1 - z^M)} < \frac{z^M}{2}.$$

Further details are in the proof of the existence of z^L; specifically, see equation (51).

Lemma 6 (Monotonicity of $\mathsf{E}^L\delta$) *Function $\mathsf{E}^L\delta$ increases in z^L and in z^M: (i) $\partial\mathsf{E}^L\delta/\partial z^L > 0$ and (ii) $\partial\mathsf{E}^L\delta/\partial z^M > 0$.*

Proof of (i): Differentiating $\mathsf{E}^L\delta$ with respect to z^L, we obtain

$$\frac{\partial\mathsf{E}^L\delta}{\partial z^L} = \frac{1}{\mathsf{pr}^L}\left[\frac{\partial\mathsf{num}(\mathsf{E}^L\delta)}{\partial z^L} - \mathsf{E}^L\delta\frac{\partial\mathsf{pr}^L}{\partial z^L}\right] = \frac{1}{\mathsf{pr}^L}\left(\gamma^L\mathsf{E}^L\delta - \beta^L\right).$$

Since $\mathsf{pr}^L \geq 0$ (with equality only at $z^L = z^M$), it suffices to show that $\mathsf{pr}^L(\gamma^L\mathsf{E}^L\delta - \beta^L) > 0$ for all $z^L \in [0, z^M)$. We will show that $\mathsf{pr}^L(\gamma^L\mathsf{E}^L\delta - \beta^L)$ is strictly decreasing in z^L on $[0, z^M 0$. The desired inequality then follows because $\mathsf{pr}^L(\gamma^L\mathsf{E}^L\delta - \beta^L) = 0$ at $z^L = z^M$. Differentiating $\mathsf{pr}^L(\gamma^L\mathsf{E}^L\delta - \beta^L)$, we obtain

$$\frac{\partial(\mathsf{pr}^L(\gamma^L\mathsf{E}^L\delta - \beta^L))}{\partial z^L} = -\frac{1}{4}(1 - z^L + \mathsf{E}^L\delta)g(1 - z^L) < 0 \qquad (49)$$

Proof of (ii): Differentiating $\mathsf{E}^L\delta$ with respect to z^M, we obtain

$$\frac{\partial\mathsf{E}^L\delta}{\partial z^M} = \frac{1}{\mathsf{pr}^L}\left[\frac{\partial\mathsf{num}(\mathsf{E}^L\delta)}{\partial z^M} - \mathsf{E}^L\delta\frac{\partial\mathsf{pr}^L}{\partial z^M}\right] = \frac{1}{\mathsf{pr}^L}\left(\beta^M - \gamma^M\mathsf{E}^L\delta\right).$$

The derivative is positive if $\beta^M \mathsf{pr}^L - \gamma^M \mathsf{num}(\mathsf{E}^L\delta) > 0$. Expanding this,

$$\beta^M \mathsf{pr}^L - \gamma^M \mathsf{num}(\mathsf{E}^L\delta) \;=\; \frac{1}{4}\int_{1-z^M}^{1-z^L} \delta(1-z^L-\delta)g(\delta)d\delta \cdot \left(\gamma^M + \beta^M\right) > 0. \quad (50)$$

Lemma 7 (MLRP Results) *For a family of densities $g(\delta|\theta)$ that obeys MLRP in θ, i.e. for $\theta_1 > \theta_2$, $g(\delta|\theta_1)/g(\delta|\theta_2)$ increases in δ, (i) probability of a limit buy pr^L decreases in θ, and (ii) expectation $\mathsf{E}^L\delta$ decreases in θ.*

The proof is obtained by direct computation, using the definition of the monotone likelihood ratio property. Details will be made available in the Internet Appendix.

Lemma 7 implies that solution z^L is largest for the uniform distribution of innovations δ_t. If this solution is below $z^M/3$ then z^L is below $z^M/3$ for any distribution $\bar{g}$. Results for the uniform distribution can be obtained by direct (numerical) computation.

Existence of $z^L_*(z^M)$. First, we establish that for any given z^M there exists $z^L_*(z^M) \in [0, z^M]$ that solves the indifference condition for the marginal limit order buyer (39). To see this, observe that (i) at $z^L = 0$, we have $\mathsf{E}^L\delta > 0 = z^L$, since

$$\mathsf{num}(\mathsf{E}^L\delta) = \frac{1}{4}\int_{1-z^M}^{1} \delta(\delta - (1 - z^M - 1))g(\delta)d\delta > 0;$$

(ii) at $z^L = z^M$, we have $\mathsf{E}^L\delta > 0 = z^L$; to see this, note that both, pr^L and $\mathsf{num}(\mathsf{E}^L\delta)$ are 0 at $z^L = z^M$. Hence,

$$\begin{aligned}
\mathsf{E}^L\delta|_{z^L=ym} &= \frac{\partial \mathsf{num}(\mathsf{E}^L\delta)/\partial z^L \mid_{z^L=z^M}}{\partial \mathsf{pr}^L/\partial z^L \mid_{z^L=z^M}} = \frac{(1/4)\cdot \int_{1-z^M}^{1} \delta g(\delta)d\delta}{(1/4) + (1/4)\cdot G(1-z^M)} \\
&= \frac{(1-G(1-z^M)\mathsf{E}[\delta \mid \delta \geq 1-z^M]}{1+G(1-z^M)} \leq \frac{(1-G(1-z^M)(2-z^M)/2}{1+G(1-z^M)} \leq \frac{z^M}{2},
\end{aligned}$$

where the inequalities hold because the uniform distribution FOSD G (hence, $\mathsf{E}[\delta \mid \delta \geq 1-z^M] \leq (2-z^M)/2$ and $G(1-z^M) \geq 1-z^M$. Existence follows by continuity of $\mathsf{E}^L\delta$.

Lemma 8 (Bounds on $z^L_*(z^M)$) *Any z^L that solves $z^L = \mathsf{E}^L\delta$ for a given z^M must be below $z^M/2$.*

Proof: The lemma follows since (i) $\mathsf{E}^L\delta$ is increasing in z^L and (ii) at $z^L = z^M$ we have $\mathsf{E}^L\delta < z^M/2$.

Uniqueness of $z^L_*(z^M)$. To show uniqueness, we will show that for a fixed z^M function $z(z^L) := \mathsf{E}^L\delta - z^L$ only crosses 0 once on $[0, z^M]$. Note that $z(0) > 0 > z(z^M)$. Since $z(\cdot)$ is continuous, it suffices to show that at z^L such that $z(z^L) = 0$, we have $\partial z/\partial z^L < 0$. (That is $z(\cdot)$ must cross 0 from above and cannot touch the x-axis). We need to show that at z^L such that $z(z^L) = 0$ (in what follows "at solution"), we have $\partial \mathsf{E}^L\delta/\partial z^L < 1$.

At solution, $\partial \mathsf{E}^L\delta/\partial z^L < 1 \Leftrightarrow \mathsf{pr}^L > \gamma^L\mathsf{E}^L\delta - \beta^L \Leftrightarrow \mathsf{pr}^L > \gamma^L z^L - \beta^L$. Since $\gamma^L z^L - \beta^L = \frac{1}{2}z^L - \frac{1}{4}(1 - G(1 - z^L))(\mathsf{E}[\delta \mid \delta > 1 - z^L] + z^L)$, it suffices to show that

$$\mathsf{pr}^L > \frac{1}{2}z^L - \frac{1}{4}(1 - G(1 - z^L))(\mathsf{E}[\delta \mid \delta > 1 - z^L] + z^L). \tag{51}$$

At at z^L such that $z(z^L) = 0$, $z^L \cdot \mathsf{pr}^L = \mathsf{num}(\mathsf{E}^L\delta)$; writing out the probability pr^L explicitly and we can rewrite inequality (51) as:

$$\left(\frac{1}{4} + G(1 - z^L)\right)(z^M - z^L) + \frac{1}{4}\int_{1-z^M}^{1-z^L} \frac{\delta - z^M}{z^M - z^L} \cdot (\delta - (1 - z^M)) \cdot g(\delta)d\delta > 0. \tag{52}$$

The first term is always positive. The second term is positive for $z^M \leq {}^1\!/_2$ (since $\delta > 1 - z^M \geq z^M$). $\Rightarrow$ remains to prove the above inequality for $z^M > {}^1\!/_2$. Denote the left-hand side of the above inequality by Δ^L. Observe that $\Delta^L = 0$ at $z^L = z^M$ (the first term is 0, and the second is 0 by l'Hôpital's rule). It thus suffices to show that

Δ^L decreases in z^L on $[0, z^M]$. Compute the appropriate derivative:

$$\frac{\partial \Delta^L}{\partial z^L} = -\frac{1}{2} + \frac{1}{4}\left(-g(1-z^L)(1-z^L-z^M) + \frac{1}{(z^M-z^L)^2}\int_{1-z^M}^{1-z^L} (\delta^2 - \delta + z^M(1-z^M))g(\delta)d\delta \right.$$
$$\left. +(1-G(1-z^L)) - (z^M - z^L)g(1-z^L)\right).$$

Since $\delta^2 - \delta$ is minimized at $\delta = 1/2$, the upper bound on the integral term depends on z^L. There are three possibilities (for $z^L < z^M$ and $z^M < 1/2$):

(i) For $z^L < 1 - z^M < 1/2 < z^M$, we have, for $\delta \in [1-z^M, 1-z^L]$, $\delta^2 - \delta < (1-z^L)^2 - (1-z^L)$, and further, $1 - z^L - z^M > 0$. Thus

$$\int_{1-z^M}^{1-z^L} (\delta^2 - \delta + z^M(1-z^M))g(\delta)d\delta < (1-z^L-z^M)g(1-z^L).$$

Consequently, since $1 - G(1-z^L) < 1$,

$$\frac{\partial \Delta^L}{\partial z^L} < -\frac{1}{2} + \frac{1}{4}(1-G(1-z^L)) - \frac{1}{4}(z^M - z^L)g(1-z^L) < 0.$$

(ii) For $1 - z^M < z^L < 1/2 < z^M$, we have, for $\delta \in [1-z^M, 1-z^L]$, $\delta^2 - \delta < (1-z^M)^2 - (1-z^M)$, and further, $1 - z^L - z^M < 0$ and $2z^L - 1 < 0$. The integral term is then negative. Consequently, since $1 - G(1-z^L) < 1$,

$$\frac{\partial \Delta^L}{\partial z^L} < -\frac{1}{2} + \frac{1}{4}(1-G(1-z^L)) + \frac{1}{4}(2z^L - 1)g(1-z^L) < 0.$$

(iii) For $1 - z^M < 1/2 < z^L < z^M$, we have, for $\delta \in [1-z^M, 1-z^L]$, $\delta^2 - \delta < (1-z^M)^2 - (1-z^M)$, and further, $1 - z^L - z^M < 0$ and $2z^L - 1 > 0$. The integral term is then negative, and we have

$$\frac{\partial \Delta^L}{\partial z^L} < -\frac{1}{4} - \frac{1}{4}G(1-z^L) + \frac{1}{4}(2z^L - 1)g(1-z^L).$$

Using the upper bound on g from Lemma (4), it remains to show that

$$-1 - G(1-z^L) + (2z^L - 1)\frac{G(1-z^L)}{1-z^L} < 0.$$

The above inequality is true for all $z^L < 3/4$, since:

$$-(1-z^L) + (-1 + z^L + 2z^L - 1)G(1-z^L) < 4z^L - 3 < 0.$$

We have thus shown that function $z(z^L) = \mathsf{E}^L\delta - z^L$ only crosses 0 once on $[0, z^M]$, for any fixed z^M.

We have thus shown existence and uniqueness of z^L that solves the indifference equation for the limit order buyer, for all $z^M \in [0, 3/4]$.

A.2.4 Step 2: Monotonicity of $z_*^L(z^M)$

To show that z^L increases in z^M, it suffices to show that the partial derivative of $z(z^L, z^M)$ in z^M is positive (the positive partial derivative implies that z, viewed as a function of z^L, will then necessarily cross 0 further to the right, since it crosses from above). This is equivalent to showing that $\partial\mathsf{E}^L\delta/\partial z^M > 0$, which in turn follows from Lemma 6.

A.2.5 Step 3: Existence of z^M

Similarly to $\mathsf{E}^L\delta$, we derive the expressions for pr^M and $\mathsf{num}(\mathsf{E}^M\delta) = \mathsf{E}^M\delta/\mathsf{pr}^M$, for $z^M \in [0, 3/4]$:

$$\begin{aligned}
\mathsf{pr}^M &= \frac{1}{4}(1-z^M) + \frac{1}{4}(1-z^M)\ G(1-z^M) + \frac{1}{4}\int_{1-z^M}^{1} \delta g(\delta)d\delta \\
\mathsf{num}(\mathsf{E}^M\delta) &= \frac{1}{4}\int_{0}^{1-z^M} 2\delta^2 g(\delta)d\delta + \frac{1}{4}\int_{1-z^M}^{1} \delta(1-z^M+\delta)g(\delta)d\delta
\end{aligned}$$

Existence of z^M that solves equation (43), i.e. z^M that solves $(1 - \mathsf{pr}^M)z^M - \mathsf{E}^M\delta + \mathsf{pr}^M z_*^L(z^M)) = 0$, follows by continuity. At $z^M = 0$, the LHS $= -\mathsf{E}^M\delta < 0$ (the inequality is strict, because $\mathsf{num}(\mathsf{E}^M\delta) > 0$ and $\mathsf{pr}^M > 0$). At $z^M = {}^3\!/_4$, we have

$$(1 - \mathsf{pr}^M)z^M - \mathsf{E}^M\delta + \mathsf{pr}^M z_*^L(z^M)) > (1 - \mathsf{pr}^M)z^M - \mathsf{E}^M\delta$$
$$> (1 - \mathsf{pr}^M)z^M - \mathsf{E}^M\delta \,|_{\text{for uniform distribution } g} > 0.$$

A.2.6 Step 4: Optimality of the Threshold Strategies

The intuition for the optimality of the threshold strategies stems from competitive pricing and stationarity of investor decisions. An investor's deviation from one equilibrium action to another equilibrium action will not affect equilibrium bid and ask prices or probabilities of the future order submissions. Consequently, it is possible to show that the difference between a payoff to a market order and a payoff to a limit order at the equilibrium price to an investor with an aggregate valuation above z^M is strictly greater than 0. (The formal argument is to be typeset).

Out-Of-The-Equilibrium-Beliefs. A more complex scenario arises when an investor deviates from his equilibrium strategy by submitting an limit order at a price different to the prescribed competitive equilibrium price. Whether or not this investor expects to benefit from such a deviation depends on the reaction to this deviation by the low-latency liquidity providers and investors in the next period. For instance, can an investor increase the execution probability of his limit buy order by posting a price that is above the equilibrium bid price?

We employ a perfect Bayesian equilibrium concept. This concept prescribes that investors and low-latency liquidity providers update their beliefs by Bayes rule, whenever possible, but it does not place any restrictions on the beliefs of market participants when they encounter an out-of-equilibrium action.

To support competitive prices in equilibrium we assume that if a limit buy order is posted at a price different to the competitive equilibrium bid price bid^*_{t+1}, then market participants hold the following beliefs regarding this investor's knowledge of the period t innovation δ_t.

If a limit buy order is posted at a price $\widehat{\mathsf{bid}} < \mathsf{bid}^*_{t+1}$, then market participants assume that this investor followed the equilibrium threshold strategy, but "made a mistake" when pricing his orders. A low-latency liquidity provider then updates his expectation about δ_t to the equilibrium value and posts a buy limit order at bid^*_{t+1}. The original investor's limit order then executes with zero probability.

If a limit buy order is posted at a price $\widehat{\mathsf{bid}} > \mathsf{bid}^*_{t+1}$, then market participants believe the this order stems from an investor from a sufficiently high aggregate valuation (e.g., $z_t = 2$) and update their expectations about δ_t to $\mathsf{E}[\delta_t \mid \widehat{\mathsf{bid}}]$ accordingly (to $\mathsf{E}[\delta_t \mid \widehat{\mathsf{bid}}] = 1$ if the belief on z_t is $z_t = 2$). The new posterior expectation of V_t equals to $p_{t-1} + \mathsf{E}[\delta_t \mid \widehat{\mathsf{bid}}]$. A low-latency liquidity provider is then willing to post a bid price $\mathsf{bid}^{**}_{t+1} \leq p_{t-1} + \mathsf{E}[\delta_t \mid \widehat{\mathsf{bid}}] + \mathsf{E}[\delta_{t+1} \mid \mathrm{MS}_{t+1}]$. With the out-of-the-equilibrium belief of $\delta_t = 1$ and with the bid-ask spread< 1, a limit order with the new price bid^{**}_{t+1} outbids any limit buy order that yields investors positive expected profits.

The beliefs upon an out-of-equilibrium sell order are symmetric. The above out-of-equilibrium beliefs ensure that no investor deviates from his equilibrium strategy.

We want to emphasize that these beliefs and actions do *not* materialize in equilibrium. Instead, they can be loosely thought of as a "threat" to ensure that investors do not deviate from their prescribed equilibrium strategies.

References

Anand, Amber, Tim McCormick, and Laura Serban, 2013, Incentives for liquidity provision: Is the make-take structure the answer?, Discussion paper, Syracuse University.

Angel, James, Lawrence Harris, and Chester Spatt, 2011, Equity trading in the 21st century, *The Quarterly Journal of Finance* 1, 1–53.

Back, Kerry, and Shmuel Baruch, 2013, Strategic liquidity provision in limit order markets, *Econometrica* 81, 363–392.

Baruch, Shmuel, and Larry Glosten, 2012, Flickering quotes, Working paper series University of Utah.

Battalio, Robert, and Craig Holden, 2001, A simple model of payment for order flow, internalization, and total trading cost, *Journal of Financial Markets* 4, 33–71.

Battalio, Robert H., Andriy Shkilko, and Robert A. Van Ness, 2012, To pay or be paid? The impact of taker fees and order flow inducements on trading costs in U.S. options markets, *SSRN eLibrary.*

Bessembinder, Hendrik, Jia Hao, and Michael L. Lemmon, 2012, Why designate market makers? Affirmative obligations and market quality, *SSRN eLibrary.*

Biais, Bruno, Thierry Foucault, and Sophie Moinas, 2012, Equilibrium high-frequency trading, *SSRN eLibrary.*

Colliard, Jean-Edouard, and Thierry Foucault, 2012, Trading fees and efficiency in limit order markets, *Review of Financial Studies, Forthcoming.*

Degryse, Hans, Mark Van Achter, and Gunther Wuyts, 2012, Tinternalization, clearing and settlement, and liquidity, *SSRN eLibrary.*

Easley, David, and Maureen O'Hara, 1987, Price, trade size, and information in securities markets, *Journal of Financial Economics* 19, 69–90.

Foucault, T., 1999, Order flow composition and trading costs in a dynamic limit order market, *Journal of Financial Markets* 2, 99–134.

Foucault, Thierry, Johan Hombert, and Ioanid Rosu, 2012, To pay or be paid? The impact of taker fees and order flow inducements on trading costs in U.S. options markets, *SSRN eLibrary.*

Foucault, T., O. Kadan, and E. Kandel, 2005, Limit order book as a market for liquidity, *Review of Financial Studies* 18, 1171–1217.

Foucault, Thierry, Ohad Kadan, and Eugene Kandel, 2012, Liquidity cycles and make/take fees in electronic market, *Journal of Finance, Forthcoming.*

Glosten, L., 1994, Is the electronic open limit order book inevitable?, *The Journal of Finance* 49, 1127–1161.

———, and P. Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogenously informed traders, *Journal of Financial Economics* 14, 71–100.

Goettler, R., C. Parlour, and U. Rajan, 2005, Equilibrium in a dynamic limit order market, *Journal of Finance* 60, 2149–2192.

———, 2009, Informed traders and limit order markets, *Journal of Financial Economics* 93, 67–87.

Hasbrouck, Joel, and Gideon Saar, 2011, Low-latency trading, *SSRN eLibrary.*

Hendershott, T., C. Jones, and A. Menkveld, 2011, Does algorithmic trading improve liquidity?, *Journal of Finance* 66, 1–33.

Hendershott, Terrence J., and Ryan Riordan, 2012, Does algorithmic trading improve liquidity?, *Journal of Financial and Quantitative Analysis, Forthcoming.*

Hoffmann, Peter, 2012, A dynamic limit order market with fast and slow traders, *SSRN eLibrary.*

International Organization of Securities Commissions, 2011, Consultation report: Regulatory issues raised by the impact of technological changes on market integrity and efficiency, Discussion paper, International Organization of Securities Commissions http://marketsnexchanges.com/assets/media/dl/whitepaper/IOSCO.pdf.

Jovanovic, Boyan, and Albert J. Menkveld, 2011, Middlemen in limit-order markets, *SSRN eLibrary.*

Kandel, Eugene, and Leslie M. Marx, 1999, Payments for order flow on Nasdaq, *Journal of Finance* 49, 35–66.

Kaniel, R., and H. Liu, 2006, So what orders so informed traders use?, *Journal of Business* 79, 1867–1913.

Kyle, Albert S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–1336.

Malinova, Katya, and Andreas Park, 2011, Subsidizing liquidity: The impact of make/take fees on market quality, *SSRN eLibrary*.

McInish, Thomas H., and James Upson, 2012, Strategic liquidity supply in a market with fast and slow traders, *SSRN eLibrary*.

Milgrom, P., and N. Stokey, 1982, Information, trade and common knowledge, *Journal of Economic Theory* 26, 17–27.

Parlour, C., 1998, Price dynamics in limit order markets, *Review of Financial Studies* 11, 789–816.

Parlour, Christine, and Uday Rajan, 2003, Payment for order flow, *Journal of Financial Economics* 68, 379–411.

Parlour, C., and D. Seppi, 2008, Limit order markets: A survey, in A. W. A. Boot, and A. V. Thakor, ed.: *Handbook of Financial Intermediation and Banking* (Elsevier Science).

Rosu, I., 2009, A dynamic model of the limit order book, *Review of Financial Studies* 22, 4601–4641.

———, 2011, Liquidity and information in order driven markets, Discussion paper, HEC.

Securities and Exchange Commission, 2010, Concept release on market structure, Release No. 34-61358, File No. S7-02-10, Discussion paper, Securities and Exchange Commission http://www.sec.gov/rules/concept/2010/34-61358.pdf.

Skjeltorp, Johannes A., Elvira Sojli, and Wing Wah Tham, 2012, Identifying cross-sided liquidity externalities, *SSRN eLibrary*.

Figure 1: Entry and Order Submission Timeline

This figure illustrates the timing of events upon the arrival of an investor at an arbitrary period, t, until their departure from the market. Value y_t is the private valuation of the period t investor and δ_t is the innovation to the security's fundamental value in period t.

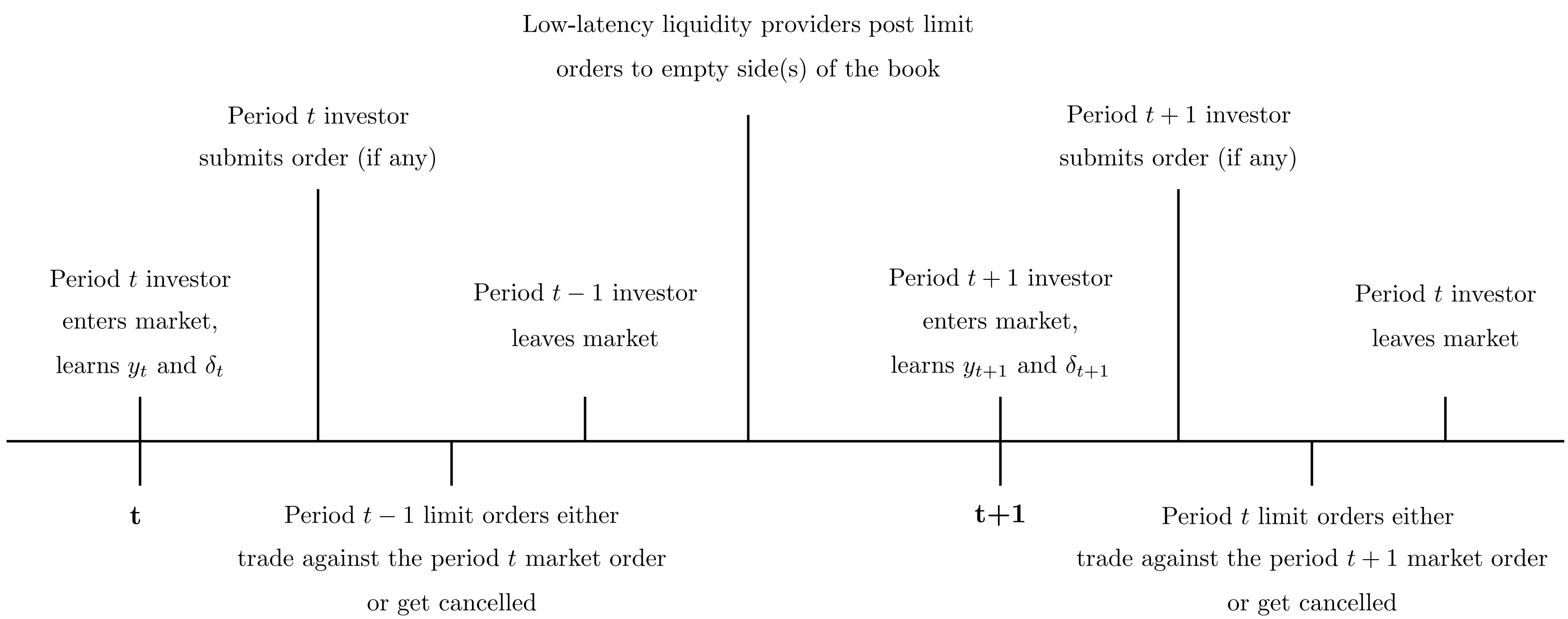

Figure 2: Equilibrium Thresholds and Payoffs to the Marginal Market and Limit Orders: Flat Fee

The left panel depicts the equilibrium aggregate valuations z^M (red line) and z^L (blue line) for the marginal market and limit order submitters, respectively. The right panel depicts the expected payoff that the investors with an aggregate valuation of z^M and z^L receive in equilibrium, as functions of the taker fee f. Both panels are for the setting where investors pay a flat, average fee per trade. An investor submits a market buy order when his aggregate valuation z_t is above z^M, a limit buy order when $z^L \leq z_t < z^M$, and abstains from trading when $|z_t| < z^L$; sell decision are symmetric to buy decisions. The plot illustrates that as f increases, investors submit more market orders and fewer limit orders. There exist the level of the taker fee $f_0 > 0$, at which the investor with aggregate valuation z^M receives zero profit from submitting a market order. The plot illustrates that investors do not submit limit orders for values of $f \geq f_0$. Parameter α in the distribution of innovations is set to $\alpha = 1.5$; results for other values of α are qualitatively similar.

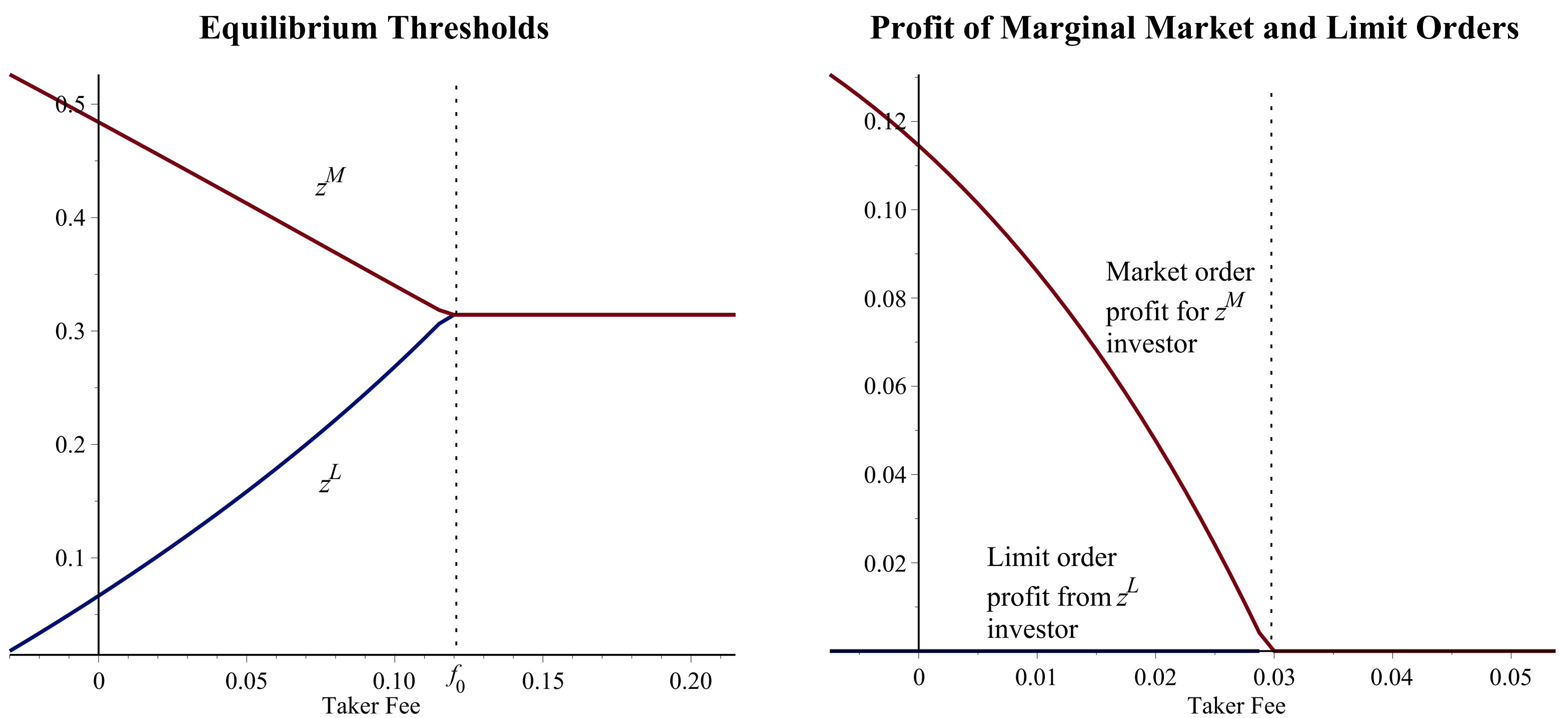

Figure 3: Trading Volume and Market Participation: Flat Fee

The left panel plots trading volume, measured as Pr(market order), as a function of the taker fee f, for the setting where investors pay a flat fee per trade. The right panel plots the level of market participation, measured as Pr(market order) + Pr(limit order), as a function of the taker fee level f. The value f_0 represents the taker fee level at which the equilibrium threshold values z^M and z^L coincide, and the marginal market order submitter z^M earns zero profits in expectation. Parameter α in the distribution of innovations is set to $\alpha = 1.5$; results for other values of α are qualitatively similar.

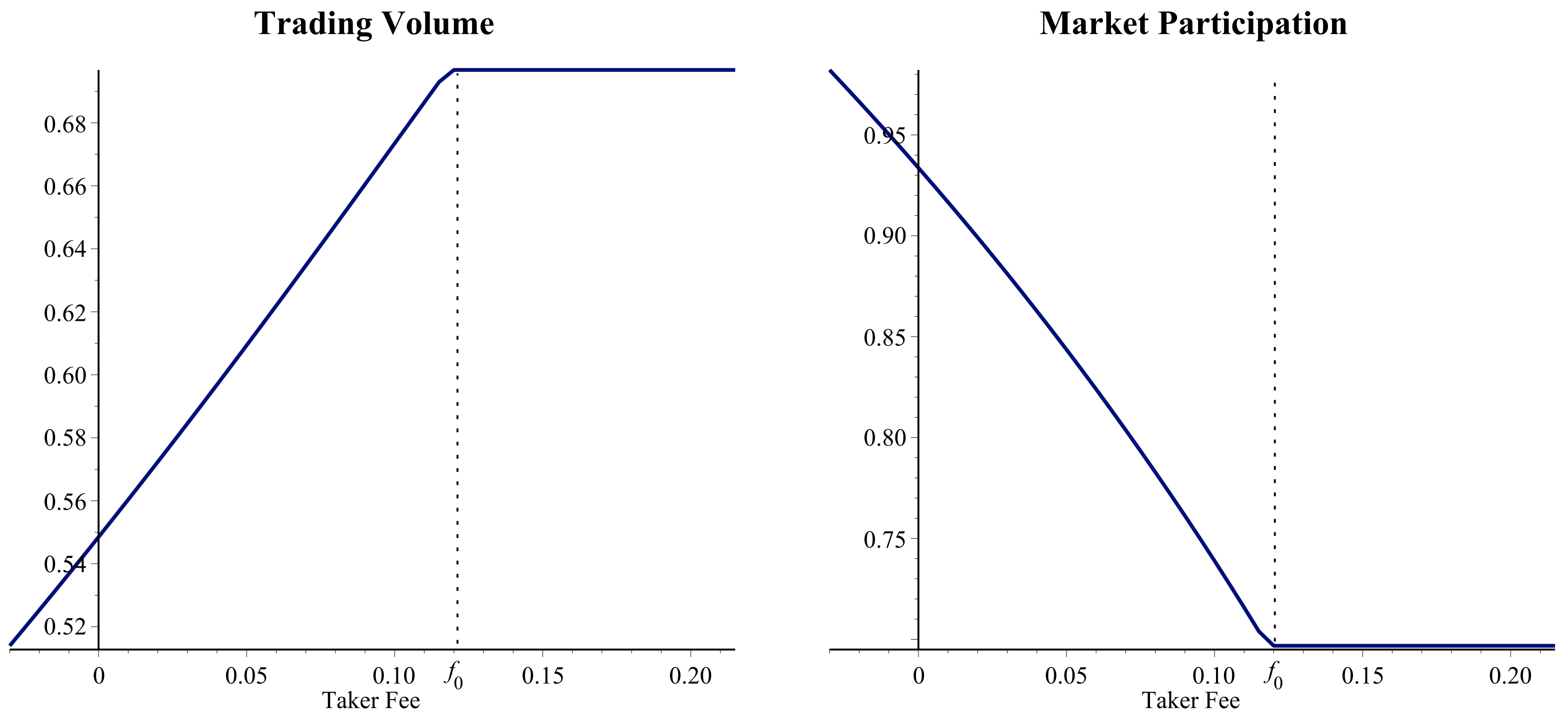

Figure 4: Quoted and Cum-Fee Spreads

The left panel plots the quoted spread (the inner, blue lines) and the cum-fee spread (the outer, red lines) as a function of the taker fee f, for the setting where all traders pay maker-taker fees. The right panel plots the quoted spread (the inner, blue lines) and the cum-fee spread (the outer, red lines) as a function of the taker fee f, for the setting in which the investor pays a flat fee per trade. The value f_0 represents the taker fee level at which the equilibrium threshold values z^M and z^L coincide, and the marginal market order submitter z^M earns zero profits in expectation. Parameter α in the distribution of innovations is set to $\alpha = 1.5$; results for other values of α are qualitatively similar.

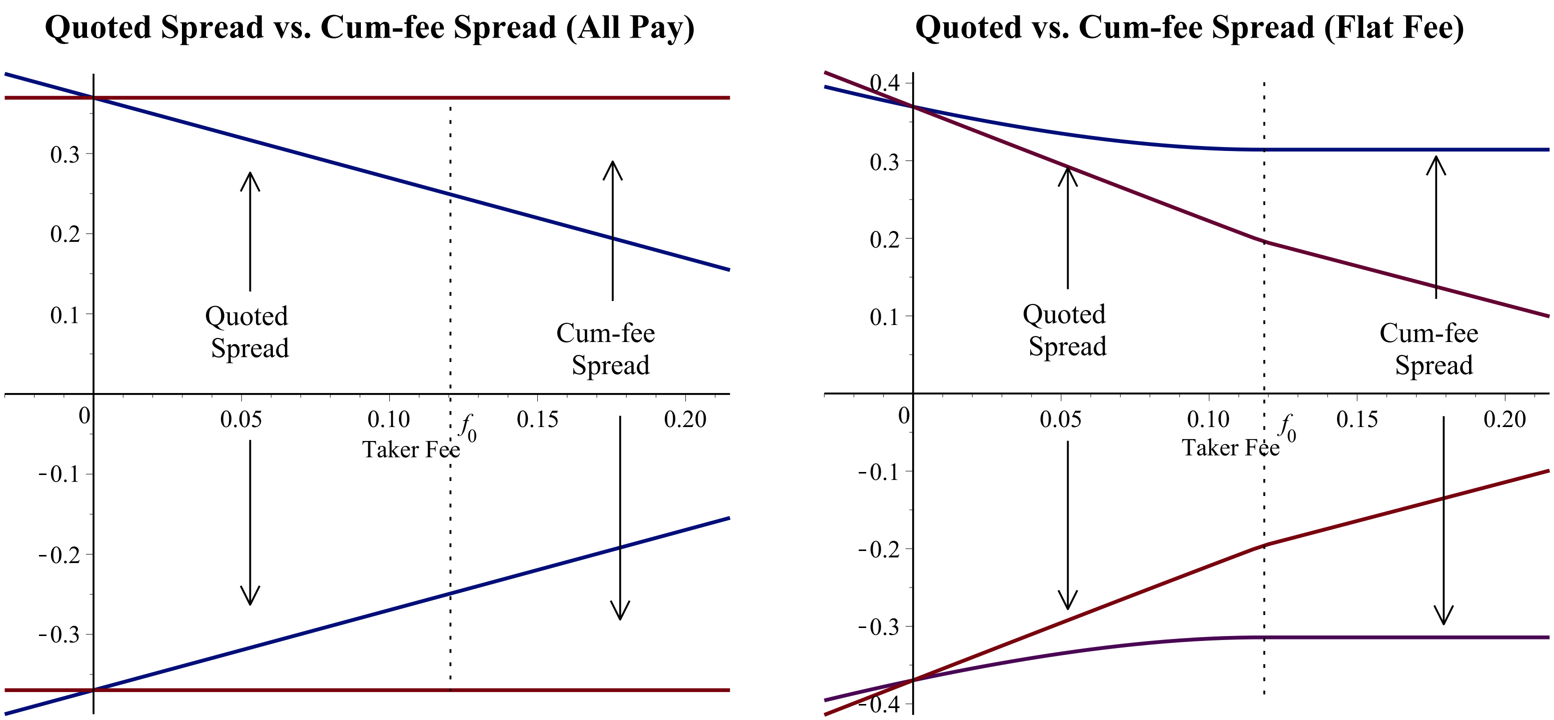

Figure 5: Price Impact

The left panel plots price impact, quoted, and cum-fee half-spreads as functions of the taker fee f for the setting where all traders pay maker-taker fees per trade. The right panel plots price impact, quoted, and cum-fee half-spreads as functions of the taker fee f for the setting in which investors pay a flat fee per trade. The value f_0 represents the taker fee level at which the equilibrium threshold values z^M and z^L coincide, and the marginal market order submitter z^M earns zero profits in expectation. Parameter α in the distribution of innovations is set to $\alpha = 1.5$; results for other values of α are qualitatively similar.

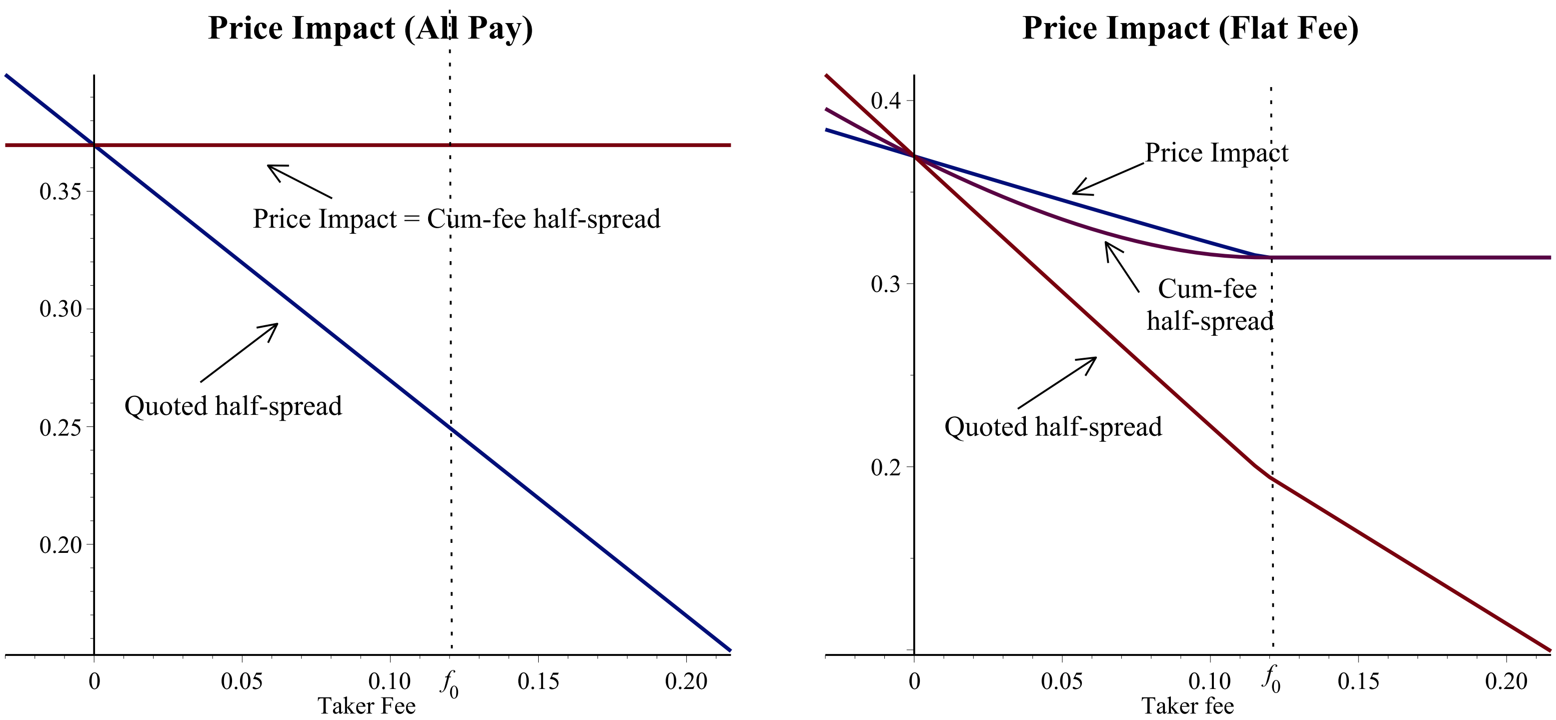

Figure 6: Social Welfare: Flat Fee

The figure plots total expected social welfare, as defined in Section 4, as a function of the taker fee f, for the setting where investors pay a flat fee per trade. The value f_0 represents the taker fee level at which the equilibrium threshold values z^M and z^L coincide, and the marginal market order submitter z^M earns zero profits in expectation. Parameter α in the distribution of innovations is set to $\alpha = 1.5$; results for other values of α are qualitatively similar.