The Journal of FINANCE
The Journal of THE AMERICAN FINANCE ASSOCIATION

THE JOURNAL OF FINANCE • VOL. LXVIII, NO. 1 • FEBRUARY 2013

Liquidity Cycles and Make/Take Fees in Electronic Markets

THIERRY FOUCAULT, OHAD KADAN, and EUGENE KANDEL*

ABSTRACT

We develop a model in which the speed of reaction to trading opportunities is endogenous. Traders face a trade-off between the benefit of being first to seize a profit opportunity and the cost of attention required to be first to seize this opportunity. The model provides an explanation for maker/taker pricing, and has implications for the effects of algorithmic trading on liquidity, volume, and welfare. Liquidity suppliers' and liquidity demanders' trading intensities reinforce each other, highlighting a new form of liquidity externalities. Data on durations between trades and quotes could be used to identify these externalities.

THE SPEED AT WHICH investors respond to trading opportunities is important for asset price dynamics. For example, Duffie (2010) demonstrates that the price impact of supply and demand shocks is more persistent when new capital responds slowly to these shocks. This effect manifests itself at various time scales, including the high frequency.[1] In recent years, firms specializing in high frequency trading have made massive investments in hardware and algorithms to accelerate their response time to trading opportunities. The effects of this evolution are not yet fully understood, and high frequency trading is the subject of heated debates associated with controversial practices such as maker/taker

*Foucault is from HEC, Paris; Kadan is from the Olin Business School at Washington University in St. Louis; and Kandel is from the Hebrew University. We are grateful to an anonymous referee, an anonymous Associate Editor, and to Campbell Harvey (the Editor) for very helpful comments and suggestions. We also thank Torben Andersen, Hank Bessembinder, Bruno Biais, Lawrence Glosten, Jeff Harris, Lawrence Harris, Pete Kyle, Terrence Hendershott, Olivier Ledoit, Ernst Maug, Albert Menkveld, Erwan Morellec, Marios Panayides, Sébastien Pouget, Michel Robe, Jean-Charles Rochet, Ioanid Rosu, Gideon Saar, Elu Von Thadden, and participants at the 2009 Western Finance Association Meeting, the 2009 European Finance Association Meeting, the NYSE-Euronext Amsterdam-Tinbergen Institute workshop on volatility and liquidity, the Fédération des Banques Françaises-Institut d'Economie Industrielle conference on investment banking and financial markets in Toulouse, the CREATES symposium on market microstructure, and the Warwick Business School conference on high frequency econometrics, as well as seminar participants at Boston University, University of Calgary, the U.S. Commodity Futures Trading Commission, Humboldt University, Ecole Polytechnique Fédérale de Lausanne, University of Mannheim, University of Toronto, and University College in Dublin for their useful comments. The usual disclaimer applies.

[1] For instance, Duffie (2010, p. 1237) observes that: "The arrival of new capital to an investment opportunity can be delayed by fractions of a second in some markets, for example an electronic limit-order-book market for equities, or by months in other markets [. . .]."

pricing, latency arbitrage, and events such as the "flash crash" of May 2010.[2]

In practice, traders do not instantaneously react to a change in the state of the market because obtaining, processing, and acting upon new information takes time. High frequency traders refer to this delay as "latency" (see Hasbrouck and Saar (2010)). For human traders, reducing latency is costly as it requires attention. Algorithmic trading, that is, using computers to make trading decisions, considerably relaxes the cost of attention, but it does not eliminate this cost completely. Indeed, even computers have limited processing capacity that needs to be allocated among multiple tasks such as parallel trading in hundreds of securities within fractions of seconds. Traders therefore face a trade-off between the benefit and the cost of monitoring the market. This paper studies this trade-off, endogenizes latencies, and analyzes the effects of drastic reductions in the cost of monitoring associated with algorithmic trading. In this way, we shed light on several important issues regarding current market structures, in particular, the widespread adoption of the maker/taker pricing model and the consequences of algorithmic trading.[3]

In electronic markets, fleeting trading opportunities often arise from the trading process itself. Indeed, large market orders consume the liquidity available at the best quotes and widen the bid-ask spread. This drop in liquidity creates a profit opportunity for liquidity suppliers who react by posting new quotes, which in turn create a new trading opportunity for liquidity demanders. This process gives rise to "liquidity cycles" consisting of two phases: a "make liquidity" phase during which, after a trade, liquidity suppliers compete to provide liquidity and a "take liquidity" phase during which liquidity demanders compete to consume liquidity.[4] In this second phase, transactions deplete the market of its liquidity and ignite a new make/take cycle. The speed at which these cycles take place determines the rate at which gains from trade are realized.

Our model is designed to analyze the determinants of this rate when market monitoring is costly. It features a trading platform with two types of traders: "market makers," who post quotes, and "market takers," who hit quotes. A profit opportunity for market makers arises after each trade as the bid-ask spread becomes momentarily large, while a profit opportunity for market takers arises when a market maker posts a new offer. Each opportunity is short-lived as it disappears as soon as a trader exploits it. Thus, traders monitor the market to react faster than their competitors to profit opportunities. In

[2] See for example Jonathan Spicer and Herbert Lash "Who's afraid of high-frequency trading?", Reuters.com, December 2, 2009, available at http://www.reuters.com/article/idUSN173583920091202.

[3] Other theoretical and empirical analyses of algorithmic trading include Biais, Hombert, and Weill (2010), Broogard (2010), Chaboud et al. (2010), Foucault and Menkveld (2008), Hasbrouck and Saar (2010), Hendershott, Jones, and Menkveld (2011), Hendershott and Riordan (2009), and Hendershott and Moulton (2011).

[4] These cycles are studied empirically in Biais, Hillion, and Spatt (1995), Coopejans, Domowitz, and Madhavan (2001), Degryse et al. (2005), and Large (2007).

Table I
Make and Take Fees

This table provides trading fees (in cents per 100 shares) for limit orders (make fee) and market orders (take fee) on different U.S. trading platforms and for various categories of stocks as of August 2009. A minus sign indicates a rebate. Source: *Traders Magazine*, August 2009.

	Tape A: NYSE Stocks		Tape B: Other Stocks		Tape C: NASDAQ Stocks	
	Make Fee	Take Fee	Make Fee	Take Fee	Make Fee	Take Fee
NYSE Arca	−23	30	−22	30	−23	30
Nasdaq	−20	30	−20	30	−20	30
BATS	−24	25	−24	25	−24	25
EDGX	−25	30	−30	30	−25	30
LavaFlow	−24	27	−24	27	−24	27

choosing their monitoring intensity, market participants trade off the benefit from a higher likelihood of being first to detect an opportunity with the cost of monitoring. In this model, durations between quotes and trades depend on traders' monitoring decisions and are ultimately determined by the parameters driving these decisions, in particular, the trading fees.

As in practice, the platform can charge distinct fees to market makers and market takers. Table I gives the make/take fees charged by major U.S. equity trading platforms as of August 2009. All these platforms use the so-called "maker/taker pricing" model: when a trade takes place, they charge a take fee to market takers and rebate part of this fee to market makers. For instance, consider a trade for 100 shares on NYSE-Arca for a stock listed on the NYSE (Tape A). The trader submitting the market order triggering this transaction (the market taker) pays a fee of 30 cents, while her counterparty (the market maker), whose limit order is being executed, receives a rebate of 23 cents. The net revenue to NYSE-Arca is seven cents. This pricing policy is also used by several European exchanges, and has been recently adopted by option markets in the United States.

To our knowledge, there is no theory explaining why the breakdown of the total fee earned by a platform between makers and takers matters.[5] Yet, maker/taker pricing is very controversial as it results in significant monetary transfers between market participants. For example, the average monthly volume on NYSE-Arca during 2009 was about 32 billion shares.[6] A net fee of seven cents per round lot generates an approximate annual revenue of $270 million to the exchange. Further, compared to an equal breakdown of the fee

[5] Colliard and Foucault (2012) develop a model of limit order trading in which trading platforms can charge make and take fees. In their model, the make/take fee breakdown is neutral: Any make/take fee breakdown is optimal for the trading platform. Thus, their model does not explain why differentiating fees between makers and takers matters.

[6] We estimated the NYSE-Arca volume by combining volume data published on nyse.com and batstrading.com.

between the two sides, the maker/taker model results in an approximate annual wealth transfer of $1.0 billion from market takers to market makers on NYSE-Arca alone.[7] Hence, not surprisingly, some high frequency firms follow rebate-capture strategies and strongly support maker/taker pricing.[8] Other market participants have claimed that maker/taker pricing results in excessive fees for takers, leading the SEC to cap take fees in equity markets at 30 cents per round lot.[9]

Our model yields several new insights. First, it provides an explanation for why the make/take fee breakdown matters. We show that differentiating make and take fees is a way for the trading platform to maximize the trading rate and therefore its expected profit. Suppose that market takers' aggregate monitoring intensity is much higher than market makers' aggregate monitoring intensity. In this case, the speed at which new liquidity is supplied after a trade is smaller than the speed at which liquidity is consumed. This imbalance slows down the trading process since trades happen only when offers are available. The trading platform can then increase the trading rate, without changing its revenue per trade, by reducing its make fee while increasing its take fee by the same amount. Indeed, such a shift in the make/take fee breakdown raises the value of being first to reinject liquidity after a trade. Thus, it incentivizes market makers to monitor the market more intensively and as a result the trading rate is higher. Building on this intuition, we show that rebates to market makers are optimal for the trading platform when (i) the ratio of the number of market makers to the number of market takers, or (ii) the ratio of market takers' monitoring cost to market makers' monitoring cost are low enough.

Angel, Harris, and Spatt (2011) argue that the make/take fee breakdown is irrelevant since traders can neutralize its effects by adjusting the price at which they trade. This is not the case in our model because quotes must be expressed as multiples of a minimum monetary unit, the "tick size," as in reality. For this reason, traders cannot fully neutralize make/take fees and the make/take fee breakdown matters.

Second, the model has implications for the effects of algorithmic trading on liquidity, volume, and welfare. We analyze these effects by considering the impact of a reduction in monitoring costs for traders since algorithmic trading considerably reduces these costs. The model implies a strong positive relationship between algorithmic trading and the trading rate. For instance, consider a

[7] In 2009, trading volume on NYSE-Arca accounted for about 14% of all volume in U.S. equity markets. See batstrading.com for detailed national trading data.

[8] The liquidity rebate can constitute an important fraction of high-frequency market-makers' profit. For instance, Menkveld (2010) estimates the trading profits for a high-frequency market-maker active in Dutch stocks. He finds that the liquidity rebate accounts for about 15% of the net spread per trade earned by the market-maker (see his Table 4, Panel C).

[9] As an example of the controversies raised by these fees, see the petition for rulemaking regarding access fees by Citadel at http://www.sec.gov/rules/petitions/2008/petn4-562.pdf. In this petition, Citadel advocates a cap on access fees. For an opposite viewpoint see the comments sent by the high-frequency trading firm GETCO to the SEC at: http://www.getcollc.com/index.php/getco/commentletters/Schedule of Fees and Charges.pdf.

decrease in the monitoring cost for market takers. This decrease leads market takers to hit good prices more quickly, which in itself contributes to a higher trading rate. However, as liquidity is consumed more quickly, market makers react by supplying liquidity more quickly as well. This externality further increases the trading rate.

The model also implies that the impact of algorithmic trading on the time-weighted bid-ask spread is ambiguous. It depends on whether the reduction in the cost of monitoring mainly affects market makers or market takers. For instance, as just explained, a decrease in market takers' monitoring cost increases the speed of reaction to changes in the state of the market for both sides. But this increase is stronger for the market takers. Thus, when market takers' monitoring cost declines, the rate at which liquidity is consumed increases relative to the rate at which liquidity is supplied. As a consequence, the time-weighted bid-ask spread increases. In contrast, when market makers' monitoring cost falls, the time-weighted bid-ask spread declines. These predictions are in line with empirical findings in Hendershott and Moulton (2011) and Hendershott, Jones, and Menkveld (2011).

In our model, traders never achieve the maximum possible trading rate because they monitor the market imperfectly. A decrease in monitoring costs always leads to a higher trading rate because it alleviates this friction. For this reason, algorithmic trading increases traders' aggregate welfare in our model. Algorithmic trading, however, is not necessarily a Pareto improvement. First, a reduction in monitoring cost for one trader has a negative effect on his competitors' expected profits. Further, when the platform optimally chooses the make/take fee breakdown, a reduction in the monitoring cost for one side can make this side worse off. For instance, consider a technological improvement reducing market makers' monitoring costs. In this case, the trading platform optimally raises its make fee and reduces its take fee. Thus, the reduction in market makers' monitoring costs unambiguously improves market takers' welfare, but it can, paradoxically, render market makers worse off.

Finally, our model uncovers a new form of liquidity externality. Suppose that an exogenous shock (e.g., a decrease in monitoring costs) induces market takers to monitor a security more intensively. The immediate effect is to accelerate the speed at which takers hit offers. Thus, trading opportunities for makers are more frequent, increasing their marginal return on monitoring. Consequently, they monitor the market more frequently, and liquidity is provided more quickly. Similarly, an increase in makers' monitoring intensity exerts a positive externality on takers, and triggers an increase in the speed at which liquidity is consumed. Thus, liquidity demand begets liquidity supply and vice versa. This "cross-side" externality is new to our paper. Indeed, other theories of liquidity externalities (e.g., Admati and Pfleiderer (1988), Pagano (1989), Hendershott and Mendelson (2000), or Dow (2004)) do not distinguish between makers and takers.

Liquidity externalities are of paramount importance to our understanding of variations in trading activity across markets or over time (see Biais, Glosten, and Spatt (2005)). However, identifying such externalities empirically is

challenging (see Barclay and Hendershott (2004)). Our model suggests a new approach towards this end that is based on durations between quotes and trades. We show that exogenous shifts in the number of market makers or their monitoring costs could be used as instruments to identify these externalities empirically.

Several papers (e.g., Goettler, Parlour, and Rajan (2005), Foucault, Kadan, and Kandel (2005), Roşu (2009, 2010)) study how the limit order book replenishes after a trade, ignoring monitoring issues and trading fees. In contrast, we do not analyze how liquidity gradually builds up after a trade. Instead, we focus on the effect of monitoring decisions and trading fees on the speed at which liquidity cycles are completed. Foucault, Roëll, and Sandås (2003) and Liu (2009) provide theoretical and empirical analyses of market-making with costly monitoring. The effects in these models are driven by market makers' exposure to the risk of being picked off, which is absent from our model. Our paper also relates to the literature on payment for order flow (e.g., Kandel and Marx (1999) or Parlour and Rajan (2003)). However, this literature focuses on why rebates for liquidity takers, rather than liquidity makers, can be optimal. Finally, our model contributes to the burgeoning literature on "two-sided markets," where transaction volume depends on the allocation of fees between end-users (see Rochet and Tirole (2006) for a survey).

Section I describes the model. In Section II, we study the equilibrium when make/take fees are fixed. We derive the optimal make/take fees in Section III, and we discuss the empirical implications of the model in Section IV. In Section V, we present extensions of the baseline model. Section VI concludes. The proofs of the main results are in the Appendix. Proofs of other results as well as some additional auxiliary results are collected in an Internet Appendix.[10]

I. Model

A. Overview

Before describing the model formally, it is worth outlining the basic economic trade-offs present in this model. We study a model of trading with market makers and market takers. Market makers post quotes for an asset while market takers hit market makers' quotes. Trades occur on a trading platform, which charges a fee each time a trade happens. The platform splits this fee between market makers and market takers.

A new quote is a trading opportunity for market takers as they can trade at this quote. A transaction gives rise to a profit opportunity for market makers as it frees a slot for a new offer. In order to exploit an opportunity, traders must be first to react to this opportunity. By monitoring the market more intensively, a trader increases his likelihood of being first to grab an opportunity, but he must pay a higher monitoring cost. Ultimately, the trading rate is determined

[10] An Internet Appendix for this article is available online in the "Supplements and Datasets" section at http://www.afajof.org/supplements.asp.

by the variables affecting this trade-off, namely, the make/take fees and the costs of monitoring. As each side generates trading opportunities for the other side, the model features a cross-side externality: an increase in market makers' monitoring intensity makes market takers better-off and vice versa.

In this setting, as shown below, it is optimal for the trading platform to charge a lower fee on the side that has the lowest aggregate monitoring intensity. In this way, the platform maximizes the trading rate by optimally balancing the rates at which liquidity is consumed and supplied. For instance, subsidizing market makers is optimal when they are outnumbered by market takers or when their monitoring cost is large. Indeed, they will monitor the market more closely to capture the rebate and as a result new liquidity is supplied faster after each trade.

We now turn to formalizing these ideas.

B. Market Participants and Cycles

B.1. Market Participants

This is an infinite horizon model of trading in the market for a security, with a continuous time line. There are two types of participants in this market: M market makers and N market takers. All participants are risk neutral.

The expected payoff of the security is v_0. Market takers value the security at $v_0 + \Gamma$, with $\Gamma > 0$, while market makers value the security at v_0. Heterogeneity in traders' valuation creates gains from trade (as, for instance, in Duffie, Gârleanu, and Pedersen (2005)). As market takers have a higher valuation, they buy the security from market makers.[11]

Market makers and market takers meet on a trading platform. Market makers post the price at which market takers can trade. This price must be on a grid with a tick size equal to $\Delta > 0$, where $\ell = \frac{\Gamma}{\Delta} > 0$ is an integer. Let

$$\mathcal{P} \equiv \{v_0 + s \cdot \Delta : s \text{ is an integer}\}, \tag{1}$$

denote this grid, and assume that $\ell \geq 2$, so that the grid includes at least one price in between v_0 and $v_0 + \Gamma$. The trading price is $a = v_0 + s \cdot \Delta \in \mathcal{P}$ for some integer s. To convey the main insights of the model, we find it useful to first fix a (i.e., s) exogenously. In Section V, we endogenize a.

We view the market makers as firms that specialize in high frequency liquidity provision (e.g., Global Electronic Trading Company (GETCO), Optiver, and Timberhill). Thus, M can also be interpreted as the number of computer servers or the amount of capital devoted to high frequency market-making. The market-taking side represents buy-side institutions or their brokers who break their large orders and feed them piecemeal when the bid-ask spread is tight to minimize their trading costs. Thus, N is a measure of agency algorithmic trading.

[11] In a more complex model, market-takers could have either high or low valuations relative to market-makers, so that they could be either buyers or sellers. This adds mathematical complexity but no additional insight.

Both types of traders increasingly use highly automated algorithms to detect and exploit trading opportunities (see Hasbrouck and Saar (2010) and Hendershott, Jones, and Menkveld (2011)).[12] For instance, Hendershott, Jones, and Menkveld (2011) write (p. 1) "There are many different algorithms, used by many different types of market participants. Some hedge funds and broker-dealers supply liquidity using algorithms [...] For assets that trade on multiple venues, liquidity demanders often use smart routers to determine where to send an order [...]."

In reality, the divide between the market-making side and the market-taking side is not as rigid. For instance, high frequency market makers sometimes use market orders to unwind their positions, and smart routers use a mix of market and limit orders. Yet, some specialization is evident, as high frequency market makers account for a large fraction of liquidity supply in electronic markets.[13] Our model captures this feature.

The trading platform charges make/take fees each time a trade occurs between a market maker and a market taker. The make fee per share, paid by the market maker, is denoted by c_m, whereas the take fee, paid by the market taker, is denoted by c_t. Negative fees are allowed. We normalize the cost of processing trades for the trading platform to zero so that, per transaction, the platform earns a profit of $\bar{c} \equiv c_m + c_t$. Consequently, for each transaction, the gain from trade (Γ) is split between the parties to the transaction and the trading platform as follows: the market taker obtains

$$\pi_t = v_0 + \Gamma - a - c_t, \tag{2}$$

the market maker obtains

$$\pi_m = a - v_0 - c_m, \tag{3}$$

and the platform obtains $\bar{c}$. We assume that $0 \leq \bar{c} \leq \Gamma$ since otherwise at least one side loses money on each trade and would therefore choose not to participate.

B.2. Racing to Be First

One key driver of algorithmic trading in reality is the need to be first to react to a change in the state of the market in order to exploit it. For instance, a brochure from IBM describes algorithmic trading as "The ability to reduce latency (the time it takes to react to changes in the market [...]) to an absolute minimum. Speed is an advantage [...] because usually the first mover gets the best price" (see "Tackling latency: the algorithmic arms race," IBM, 2008). In

[12] Algorithmic trading is also used to implement other trading strategies. For instance, proprietary trading desks and hedge funds use algorithms for statistical arbitrage, to anticipate the direction of future order flow, or to react to news arrivals. The analysis of these aspects of algorithmic trading is beyond the scope of our paper.

[13] For instance, Menkveld (2010) reports that the high-frequency market-maker studied in his paper executes 78% of his trades with limit orders (see Table 4 in Menkveld (2010)).

the same spirit, an article from *Traders Magazine* observes that "The reality is, that order is only there for one person. So if you react faster, you fill the order" (see "The Race to Zero," *Traders Magazine*, 2009, p. 38).

To model this race in the simplest possible way, we normalize the number of shares that can be offered at price a to one. This constraint creates competition among traders for being first to react to a trading opportunity on their side. Indeed, when there is no quote at a, there is a profit opportunity, worth π_m, for a market maker. But a market maker can exploit this opportunity only if she is first to submit an offer at a since no more than one share can be offered at price a. In a symmetric way, when there is an offer at a, there is a profit opportunity, worth π_t, for a market taker. The market taker needs to be first to react in order to grab this opportunity since the number of shares supplied at price a is limited.[14] Thus, at each point in time the market can be in one of two states:

(i) *State E (for Empty):* there is a profit opportunity for makers because no offer is posted at a.
(ii) *State F (for Full):* there is a profit opportunity for takers because an offer is posted at a.

The market moves from state F to state E when a market taker hits the best offer. The market then remains in state E until a market maker posts a new offer, at which time the market moves from state E to state F and the process starts over. We call the flow of events from the moment the market gets into state E until it returns into this state a "make/take cycle" or for brevity just a "cycle." Figure 1 illustrates the flow of events in a cycle.

C. Market Monitoring

Traders monitor the market to be the first to detect a profit opportunity. Market makers are looking for periods when liquidity is "scarce" (no offer is posted at price a) and market takers are looking for periods when liquidity is "abundant" (an offer is posted at price a). Market monitoring includes obtaining information on the state of the market, processing this information, and making decisions based on this information. We model it as follows. Each market maker $i = 1, \ldots, M$ inspects the market according to a Poisson process with parameter $\mu_i \geq 0$, characterizing her monitoring intensity. As a result, the time between two inspections by market maker i is distributed exponentially with an average interinspection time of $\frac{1}{\mu_i}$. Similarly, each market taker $j = 1, \ldots, N$ inspects the market according to a Poisson process with parameter

[14] This limit could arise endogenously from exposure to the risk of being picked off for liquidity suppliers (as in Glosten (1994) or Sandås (2001)). We take a simpler approach for tractability.

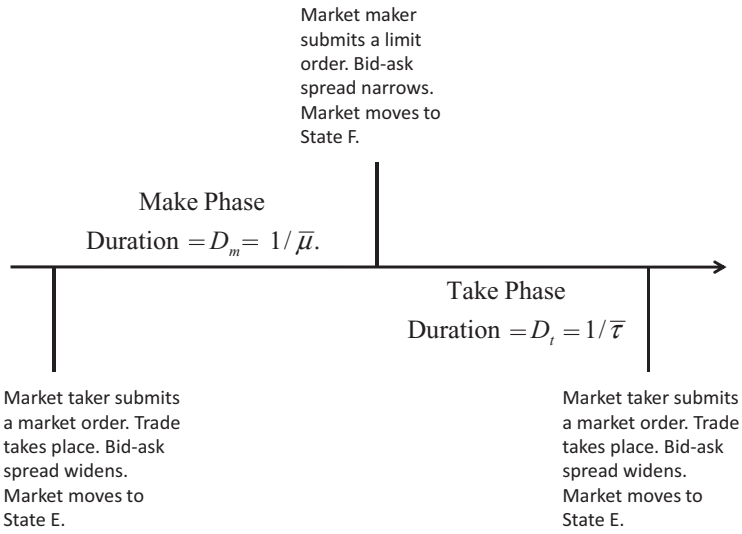

Figure 1. Flow of events in a cycle. The figure plots the timeline of events in a cycle. A cycle begins with a market taker submitting a market order, widening the spread. A cycle continues with a market maker submitting a limit order, narrowing the spread. The cycle ends when another market taker hits the limit order, widening the spread again, beginning a new cycle.

$\tau_j \geq 0$.[15] The aggregate monitoring level of the market-making side is

$$\bar{\mu} \equiv \mu_1 + \cdots + \mu_M, \tag{4}$$

and the aggregate monitoring level of the market-taking side is

$$\bar{\tau} \equiv \tau_1 + \cdots + \tau_N. \tag{5}$$

When a market maker inspects the market, she learns whether it is in state E or F. If the market is in state E, then she posts an offer at a. If instead the market is in state F, the market maker stays put until her next inspection. Similarly, a market taker submits a market order when, upon inspection, he observes an offer at price a, and stays put until the next inspection otherwise. Thus, each cycle has two phases: a "make phase" (state E to state F) and a "take phase" (state F to state E). The duration of the make phase is exponentially distributed with parameter $\bar{\mu}$, and the duration of the take phase is exponentially distributed with parameter $\bar{\tau}$. As monitoring levels are endogenous (see below), the distributions of interevent durations in the model are endogenous

[15] This approach rules out deterministic monitoring such as inspecting the market exactly once every certain number of seconds. In reality, many unforeseen events can capture the attention of a market-maker or a market-taker, be it human or a machine. For humans, the need to monitor several securities as well as perform other tasks precludes evenly spaced inspections. Computers face similar constraints as periods of high transaction volume, and unexpectedly high traffic on communication lines, prevent monitoring at exact points in time.

as well. We denote by $D_m \equiv 1/\bar{\mu}$ the expected duration from the time an offer is taken until a new offer is made, and by $\mathcal{D}_t \equiv 1/\bar{\tau}$ the expected duration from the time an offer is made until it is taken.

On average, the duration between two trades is

$$\mathcal{D}(\bar{\mu}, \bar{\tau}) \equiv \mathcal{D}_m + \mathcal{D}_t = \frac{1}{\bar{\mu}} + \frac{1}{\bar{\tau}} = \frac{\bar{\mu} + \bar{\tau}}{\bar{\mu} \cdot \bar{\tau}}, \tag{6}$$

and the trading rate, that is, the average number of transactions per unit of time, is

$$\mathcal{R}(\bar{\mu}, \bar{\tau}) \equiv \frac{1}{\mathcal{D}(\bar{\mu}, \bar{\tau})} = \frac{\bar{\mu} \cdot \bar{\tau}}{\bar{\mu} + \bar{\tau}}. \tag{7}$$

The trading rate increases when either $\bar{\mu}$ or $\bar{\tau}$ increases.

In reality, as explained in the introduction, traders do not instantaneously react to a change in the state of the market, including the execution of their own orders, because market monitoring is costly. To account for this cost, we assume that, over a time interval of length T, a market maker choosing a monitoring intensity μ_i bears a monitoring cost

$$C_m(\mu_i) \equiv \frac{1}{2}\beta\mu_i^2 T \quad \text{for } i = 1, \ldots, M. \tag{8}$$

Similarly, the cost of monitoring for market taker j over an interval of time of length T is

$$C_t(\tau_j) \equiv \frac{1}{2}\gamma\tau_j^2 T \quad \text{for } j = 1, \ldots, N. \tag{9}$$

Algorithmic trading reduces monitoring costs. We therefore analyze the effect of algorithmic trading by considering the effect of a reduction in β and γ. Parameters γ and β must remain strictly positive, but they can be very small, which may well be the case for high frequency traders. Traders' monitoring intensities and therefore their speed of reaction to changes in the state of the market can be arbitrarily high if γ and β are small enough (see below).

D. Objective Functions, Timing, and Externalities

D.1. Objective Functions

Each player chooses her action to maximize her steady-state payoff per unit of time. Consider a market maker first. Each time a make/take cycle is completed a transaction occurs. The probability that market maker i is active in this transaction is the probability that she is first to post an offer at price a after the market entered state E. Given our assumptions on the monitoring process, this probability is $\frac{\mu_i}{\mu_1+\cdots+\mu_M} = \frac{\mu_i}{\bar{\mu}}$. Thus, in each cycle, the expected profit gross of monitoring costs for market maker i is $\frac{\mu_i}{\bar{\mu}} \cdot \pi_m$. Using this remark, we show in the Appendix that market maker i's expected profit per unit of time, net of

monitoring costs, is

$$\Pi_{im} = \frac{\mu_i}{\bar{\mu}} \cdot \pi_m \cdot \mathcal{R}(\bar{\mu}, \bar{\tau}) - \frac{1}{2}\beta\mu_i^2. \tag{10}$$

This is intuitive: the expected profit of a market maker per unit of time is her expected profit per cycle ($\frac{\mu_i}{\bar{\mu}} \cdot \pi_m$) times the number of cycles per unit of time, less the monitoring cost. Similarly, the expected profit per unit of time of market taker j is

$$\Pi_{jt} = \frac{\tau_j}{\bar{\tau}} \cdot \pi_t \cdot \mathcal{R}(\bar{\mu}, \bar{\tau}) - \frac{1}{2}\gamma\tau_j^2, \tag{11}$$

while the expected profit of the trading platform per unit of time is:

$$\Pi_e \equiv \bar{c} \cdot \mathcal{R}(\bar{\mu}, \bar{\tau}) = (c_m + c_t) \cdot \mathcal{R}(\bar{\mu}, \bar{\tau}), \tag{12}$$

since in each cycle, it earns a fee $\bar{c}$.

D.2. Timing and Equilibrium

The trading game unfolds in three stages as follows:

Stage 1: The trading platform chooses its make/take fees c_m and c_t.
Stage 2: Market makers and market takers simultaneously choose their monitoring intensities μ_i and τ_j ($i = 1, \ldots, M$ and $j = 1, \ldots, N$).
Stage 3: From this point onward, the game is played on a continuous time line indefinitely, with the monitoring intensities and fees determined in Stages 1 and 2.

We solve the model backwards. First, for given fees, we look for Nash equilibria in monitoring intensities in Stage 2. A Nash equilibrium in this stage is a vector of monitoring intensities $(\mu_1^*, \ldots, \mu_M^*, \tau_1^*, \ldots, \tau_N^*)$ such that for all $i = 1, \ldots, M$, μ_i^* maximizes market maker i's expected profit per unit of time (given by (10)), and for all $j = 1, \ldots, N$, τ_j^* maximizes market taker j's expected profit per unit of time (given by (11)), taking the monitoring intensities of all other traders as given. Second, given a Nash equilibrium in the monitoring intensities, we solve for the make/take fees (c_m^*, c_t^*) that maximize the trading platform's expected profit (given by (12)).

D.3. Liquidity Externalities and Cross-Side Complementarities

An increase in one trader's monitoring level hurts the traders who are on his side. That is, $\frac{\partial \Pi_{im}}{\partial \mu_j} < 0$ and $\frac{\partial \Pi_{it}}{\partial \tau_j} < 0$ (for $j \neq i$). This effect captures the horse race to be first to detect a trading opportunity in our model. In contrast, an increase in the aggregate monitoring level of one side exerts a positive externality on the other side since $\frac{\partial \Pi_{im}}{\partial \bar{\tau}} > 0$ and $\frac{\partial \Pi_{jt}}{\partial \bar{\mu}} > 0$. For instance, an increase in market makers' aggregate monitoring increases the likelihood that market takers will find a trading opportunity, which makes the latter better off. Further, the

marginal benefit of monitoring for traders on one side increases in the aggregate monitoring level of traders on the other side since $\frac{\partial^2 \Pi_{im}}{\partial \bar{\tau} \partial \mu_i} > 0$ and $\frac{\partial^2 \Pi_{jt}}{\partial \bar{\mu} \partial \tau_j} > 0$.[16] For this reason, market makers (resp., market takers) check the state of the market more frequently when they expect market takers (resp., market makers) to check the state of the market more frequently. Thus, there is a cross-side complementarity in monitoring decisions: the monitoring intensities of traders on different sides reinforce each other.

E. Discussion

Our model is clearly stylized. First, we assume that orders are for one share and that market makers cannot queue at or behind the best price. As explained previously, these constraints create competition among traders for being first to react to a trading opportunity. They considerably simplify the analysis as allowing market makers to queue at a given price in our model is very difficult. Our results, however, should be robust in more complex environments as long as there is a benefit to being first to react to a trading opportunity. In reality, market makers naturally benefit from being first because early limit orders have time priority. Hence, a limit order at the front of the queue at a given price has a greater expected profit than other limit orders in the queue since its execution probability is higher. In addition, market takers benefit from being first because the number of shares available at the best quote is limited.

Second, we assume that, after a trade, market takers immediately receive a new buy order for one share to execute. That is, the intensity at which market takers receive a new buy order after each trade is infinite. In Section V, we consider the less extreme case in which this intensity is finite. In this case, the model becomes intractable unless $N = M = 1$. Analysis of this case, however, shows that the insights of the baseline model are robust even when market takers receive new buy orders at a finite rate. Indeed, the rate at which market takers receive new buy orders sets the maximum possible trading rate for the security, but it does not affect imperfect monitoring, that is, the friction preventing traders from achieving this maximum in our model.

For tractability, we also assume that the value of the asset (v_0) is fixed and there is no arrival of information regarding this value. Thus, there is no role for news monitoring in our model. In reality, traders monitor both changes in the state of the market and the flow of information, either to get protection against the risk of being picked off or to pick off stale quotes (as in Foucault, Roëll, and Sandås (2003), for instance). The trade-offs present in our model would still play a role in a more general model with both news and market monitoring. An increase in the risk of being picked off reduces market makers' expected profits, other things being equal. Hence, the logic of our model suggests that the make fee should decrease when this risk increases so as to strengthen market makers' incentive to monitor the market.

[16] For instance, $\frac{\partial^2 \Pi_{im}}{\partial \bar{\tau} \partial \mu_i} = \frac{\bar{\mu} \sum_{j \neq i} \mu_j + \bar{\tau}(\bar{\mu} + \mu_i)}{(\bar{\mu} + \bar{\tau})^3} > 0$.

II. Equilibria with Fixed Fees

In this section we study the equilibrium monitoring intensities for given fees (c_m, c_t). For all parameter values, the model has two equilibria: one equilibrium with no trade and one equilibrium with trade.

Consider first how the no-trade equilibrium arises. If market makers believe that market takers will not monitor the trading platform, then they optimally choose not to monitor as well since monitoring is costly. Symmetrically, if market takers expect market makers to pay no attention to the trading platform, then they optimally choose to be inactive. Thus, traders' beliefs that the other side will not be active are self-fulfilling.

PROPOSITION 1: *There exists an equilibrium in which traders do not monitor:* $\mu_i^* = \tau_j^* = 0$ *for all* $i \in \{1, \ldots, M\}$ *and* $j \in \{1, \ldots, N\}$. *The trading rate in this equilibrium is zero.*

The second equilibrium does involve monitoring and trade. To describe it let $r \equiv \gamma/\beta$ and

$$z \equiv \frac{\pi_m}{\pi_t}\frac{\gamma}{\beta} = \frac{\pi_m}{\pi_t} \cdot r. \tag{13}$$

When $z > 1$ (resp., $z < 1$), the ratio of profits to costs per cycle is higher for market makers (resp., market takers).

PROPOSITION 2: *There exists a unique equilibrium with trade. In this equilibrium, traders' monitoring intensities are given by*

$$\mu_i^* = \frac{M + (M-1)\mathcal{V}^*}{(1+\mathcal{V}^*)^2} \cdot \frac{\pi_m}{M\beta} \quad i = 1, \ldots, M \tag{14}$$

$$\tau_j^* = \frac{\mathcal{V}^*\left((1+\mathcal{V}^*)N - 1\right)}{(1+\mathcal{V}^*)^2} \cdot \frac{\pi_t}{N\gamma} \quad j = 1, \ldots, N, \tag{15}$$

where $\mathcal{V}^*$ *is the unique positive solution to the cubic equation*

$$\mathcal{V}^3 N + (N-1)\mathcal{V}^2 - (M-1)z\mathcal{V} - Mz = 0. \tag{16}$$

In addition, in equilibrium, $\frac{\bar{\mu}^*}{\bar{\tau}^*} = \mathcal{V}^*$.

The ratio $\mathcal{V}^* = \frac{\bar{\mu}^*}{\bar{\tau}^*} = \frac{\mathcal{D}t}{\mathcal{D}_m}$ measures the speed of reaction of the market-making side $(\frac{1}{\mathcal{D}_m})$ relative to the market-taking side $(\frac{1}{\mathcal{D}_t})$ in equilibrium. We call it the *velocity ratio.*

As an illustration of Propositions 1 and 2, consider the case $M = N = 1$. Figure 2 plots traders' best response functions, denoted by $\rho_m(\tau_1)$ and $\rho_t(\mu_1)$, when $\pi_m = \pi_t = 0.5$, $\beta = \gamma = 0.5$. For instance, $\rho_m(\tau_1)$ is the optimal monitoring level of the market maker given that the market taker's monitoring level is τ_1. The two best response functions meet at $(0, 0)$ and $(0.25, 0.25)$, which are the

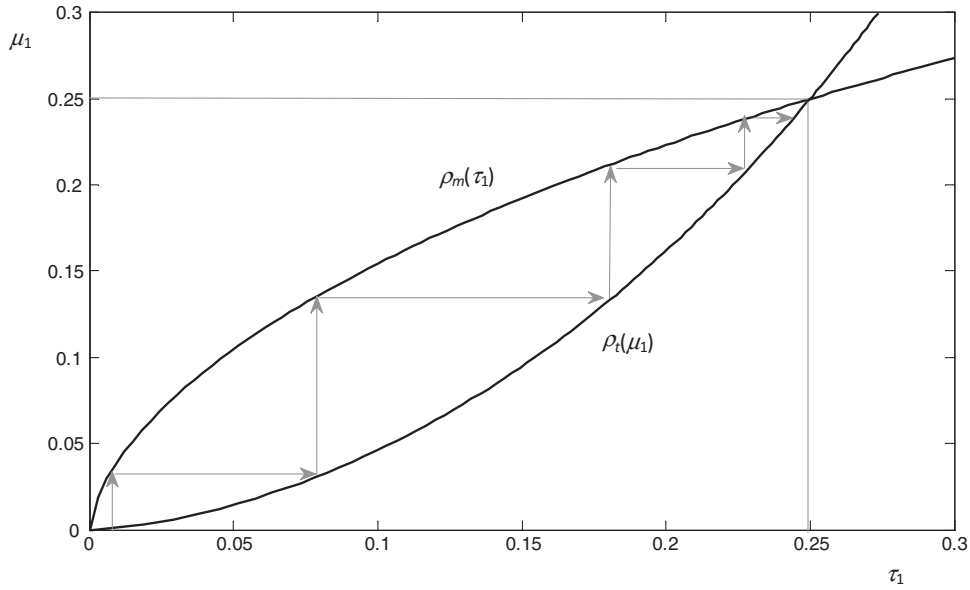

Figure 2. Equilibria with low and high trading rates. The market maker's best-response function, $\rho_m(\cdot)$, is plotted as a function of the market taker's monitoring intensity (τ_1), whereas the market taker's best-response function, $\rho_t(\cdot)$, is plotted as a function of the market makers monitoring intensity (μ_1).

two equilibria corresponding to Propositions 1 and 2. Due to the cross-side complementarities, the slope of both reaction functions is positive: an increase in the monitoring intensity of one side triggers an increase in the monitoring intensity of the other side. Further, it can be checked that these slopes are infinite at zero. Thus, the no-trade equilibrium is unstable: in this situation, an infinitesimal increase in, say, the market maker's monitoring intensity, μ_1, triggers a relatively large increase in the market taker's monitoring intensity, τ_1, which in turn triggers even more attention by the market maker and so on. Along this off-equilibrium path, illustrated by the arrows in the figure, the trading rate gets higher and higher since it increases with monitoring levels on either side. This process ends when monitoring intensities reach their equilibrium level in Proposition 2.

From now on we focus our attention on the equilibrium with trade. An implication of the cross-side complementarity is that a change in the cost and benefit of monitoring for one side triggers a change in the monitoring intensities on *both* sides in equilibrium, as highlighted by the next corollary (proved in the Internet Appendix).

COROLLARY 1: *In the unique equilibrium with trade*,

(i) *The aggregate monitoring level of both sides increases in the number of participants on either side, and decreases in monitoring costs and in the fee charged on either side.*

(ii) *The trading rate decreases in the monitoring costs and trading fees, and increases in the number of participants on either side.*

To understand the first part of the corollary, consider an increase in market makers' monitoring cost. This increase reduces their individual monitoring levels, other things being equal. The marginal benefit of monitoring for market takers is then smaller as they are less likely to find a good price when they inspect the market. Consequently, market takers monitor the market less intensively, even though their own monitoring cost has not changed. The same reasoning applies for an increase in the trading fee or the number of participants on one side.[17] The second part of the corollary follows from the first part, since any change in the aggregate monitoring intensities $\bar{\mu}^*$ and $\bar{\tau}^*$ translates into a change in the same direction for the trading rate.

COROLLARY 2: *In equilibrium, for fixed fees, the market-making side monitors the market more intensively than the market-taking side* ($\bar{\mu}^* > \bar{\tau}^*$) *if and only if* $\frac{z(2M-1)}{2N-1} \geq 1$.

This corollary implies that the velocity ratio, $\mathcal{V}^* = \frac{\bar{\mu}^*}{\bar{\tau}^*}$, is different from one in equilibrium, unless $\frac{z(2M-1)}{(2N-1)} = 1$. A velocity ratio greater (less) than one means that liquidity is consumed by market takers relatively less (more) quickly than it is supplied by market makers. For instance, if $M = N$ and $z > 1$, the market-making side reacts more quickly than the market-taking side because market makers' cost of missing a trading opportunity is relatively higher and, as a result, $\mathcal{V}^* > 1$. A situation in which the velocity ratio differs too much from one is suboptimal for the platform. Indeed, it means that one side is very quick in taking advantage of trading opportunities, but this velocity does not translate into a high trading rate since the other side is relatively slow. In this situation it is optimal for the platform to adjust its fees so as to reduce the imbalance between the speed at which liquidity is consumed and the speed at which it is supplied (see the next section). The next corollary shows how the velocity ratio changes when trading fees or other parameters change.

COROLLARY 3: *The velocity ratio increases in the take fee,* c_t, *and decreases in the make fee,* c_m. *In addition, it increases with the relative size of the market-making side,* $q \equiv \frac{M}{N}$, *and the relative monitoring cost for the market-taking side,* $r \equiv \frac{\gamma}{\beta}$.

Thus, the platform controls the velocity ratio with its fees. For instance, the platform can reduce the velocity ratio without changing its revenue per trade by increasing the make fee while reducing the take fee.

[17] When the number of participants on one side increases, the individual monitoring levels on this side may decrease since the likelihood of being first to grab a profit opportunity declines. This is in contrast to the aggregate level of monitoring, which goes up.

In general we do not have an explicit solution for traders' monitoring levels because we cannot solve for $\mathcal{V}^*$ (the unique positive root of equation (16)) in closed form. However, there are a few cases in which a closed-form solution can be obtained. One insightful case is when the number of participants on both sides becomes very large but the size of the market-making side relative to the size of the market-taking side, $q \equiv \frac{M}{N}$, remains fixed.[18] We refer to this as "the thick market case." In this case, we have[19]

$$\mathcal{V}^\infty \equiv \lim_{M\to\infty} \mathcal{V}^* = (zq)^{\frac{1}{2}}. \tag{17}$$

Using this observation and Proposition 2, the next corollary provides closed-form expressions for traders' monitoring levels when the market is thick.

COROLLARY 4: *Fix $q > 0$ and assume $N = \frac{M}{q}$. Then, the monitoring levels when the market is thick are*

$$\mu_i^\infty \equiv \lim_{M\to\infty} \mu_i^* = \frac{1}{1+(zq)^{\frac{1}{2}}} \cdot \frac{\pi_m}{\beta} \quad i = 1, 2, 3, \ldots \tag{18}$$

$$\tau_j^\infty \equiv \lim_{M\to\infty} \tau_j^* = \frac{1}{1+(zq)^{-\frac{1}{2}}} \cdot \frac{\pi_t}{\gamma} \quad j = 1, 2, 3, \ldots \tag{19}$$

III. Optimal Make/Take Fees

We now study the fees set by the trading platform given the monitoring strategies derived in the previous section.

A. *Determinants of Make/Take Fees*

The platform's optimization problem can be decomposed into two steps: (i) choose the optimal make/take fees for a given $\bar{c}$; and (ii) choose the optimal $\bar{c}$. As we are interested in the optimal breakdown of the total fee between market makers and market takers, we focus on the first step. That is, we take the total fee, $\bar{c}$, as given throughout. In the Internet Appendix, we solve for the optimal total fee when the market is thick and show that our conclusions are not affected when this fee is endogenous.[20] We refer to the differential between the take fee and the make fee, $c_t - c_m$, as the take/make spread.

[18] Another case is when $M = N = 1$. See the Internet Appendix for a detailed analysis of this case.

[19] To see this, note that equation (16) implies $z = \frac{\mathcal{V}^{*3}\frac{M}{q} + (\frac{M}{q}-1)\mathcal{V}^{*2}}{(M-1)\mathcal{V}^* + M}$. Equation (17) follows by taking the limit as $M \to \infty$.

[20] As shown below, the optimal make/take fees, (c_m^*, c_t^*) increase in $\bar{c}$. Consequently, in choosing its total fee, the trading platform faces the standard price-quantity trade-off: by raising $\bar{c}$, the trading platform gets a larger revenue per trade but it decreases the rate at which trades occur since the trading rate decreases in both the make fee and the take fee (Corollary 1).

The optimal fees depend on the price at which market makers and market takers agree to trade. We thus denote by (c^*_{ms}, c^*_{ts}) the optimal make and take fees for the platform when the makers' spread, $a - v_0$, equals $s \cdot \Delta$. For a given total fee $\bar{c}$, the optimal make/take fees are the solution of

$$\begin{aligned} \max_{c_m, c_t} \ & \Pi_e = (c_m + c_t)\,\mathcal{R}(\bar{\mu}^*, \bar{\tau}^*) \\ s.t: \quad & c_m + c_t = \bar{c}, \end{aligned} \tag{20}$$

where $\bar{\mu}^* = M\mu^*_i$, $\bar{\tau}^* = N\tau^*_j$, and μ^*_i and τ^*_j are given by Proposition 2. Trading fees affect traders' monitoring decisions and thereby the trading rate (Corollary 1). The first-order conditions for this problem impose that

$$\frac{\partial \mathcal{R}(\bar{\mu}^*, \bar{\tau}^*)}{\partial c_m} = \frac{\partial \mathcal{R}(\bar{\mu}^*, \bar{\tau}^*)}{\partial c_t}. \tag{21}$$

That is, the trading platform chooses its fee structure so as to equalize the marginal negative impact of an increase in each fee on the trading rate.

To gain insight on the optimal make/take fee breakdown, we first solve for the optimal make/take fees when the market is thick (see Corollary 4). As the market becomes thick, traders' aggregate monitoring levels and the trading rate explode. Yet the fee structure that maximizes the trading rate converges to a well defined limit, as shown in our next proposition.

PROPOSITION 3: *Suppose that $a = v_0 + s \cdot \Delta$. In the thick market case, the trading platform optimally allocates its fee $\bar{c}$ between the market-making side and the market-taking side as follows*

$$c^*_{ms} = s \cdot \Delta - \frac{\Gamma - \bar{c}}{1 + (qr)^{\frac{1}{3}}} \quad \textit{and} \quad c^*_{ts} = \bar{c} - c^*_{ms}. \tag{22}$$

Thus, in general, it is optimal for the trading platform to differentiate its make and take fees. To understand this point, it is useful to first analyze how the optimal make and take fees depend on the exogenous parameters, q and r.

COROLLARY 5: *In the thick market case, the take fee decreases and the make fee increases with the relative size of the market-making side, q, and the relative monitoring cost for the market-taking side, r. Thus, the take/make spread decreases with these parameters.*

Evidently, the optimal pricing policy follows a simple principle: when a change in parameters raises market makers' aggregate monitoring intensity, the platform allocates a greater fraction of the total fee to the market makers. For instance, an increase in the relative size of the market-making side, q, or a decrease in its relative monitoring cost, r, results in a higher aggregate monitoring intensity for market makers relative to market takers, other things equal (Corollary 1). As a result, a new offer is posted very quickly

after a trade but market takers are relatively slow to hit this offer. Therefore, it is optimal for the platform to lower its fee on market takers to accelerate their response to market makers' offers, and compensate the loss in revenue by increasing its fee on market makers since they are relatively fast anyway. Hence, fee differentiation is a way for the trading platform to increase the trading rate by better balancing the speeds at which liquidity is consumed and supplied.

Corollary 5 suggests that the optimal take/make spread depends on stock characteristics as q is likely to vary across stocks. Proposition 3 also shows that the trading platform is more likely to subsidize market makers ($c^*_{ms} < 0$) when the makers' spread, $a - v_0$, is small relative to the gains from trade, $\Gamma - \bar{c}$. Indeed, in this case market makers' aggregate monitoring intensity is low since they obtain a small fraction of gains from trade. A rebate is thus a way to incentivize market makers to monitor the market more intensively. By symmetry, if the makers' spread is large relative to the gains from trade, then the trading platform is more likely to subsidize the market takers. We show in Section V.C that this reasoning remains valid even when the price at which investors choose to trade is endogenous.

Proposition 3 and Corollary 5 hold when the market is thick. In the other polar case, when $M = N = 1$, the expressions for the optimal fees are very similar and Corollary 5 is still valid (see the Internet Appendix). For intermediate values of M and N, we cannot obtain a closed-form solution for the optimal fees. However, we can numerically solve for these fees using the characterization of traders' monitoring levels in Proposition 2. Applying this approach, we check through extensive numerical simulations that Corollary 5 is robust.[21]

B. Example: Uniform, Optimal, and Capped Make/Take Fees

Achieving the optimal make/take fee breakdown can have a first-order effect on a platform's revenue. To see this point, consider the following numerical example. Assume $M = 10$, $N = 20$, $\Gamma = 20$, $v_0 = 250$, and $\bar{c} = 1/10$ (all monetary amounts in cents).[22] We arbitrarily set the time unit as one second, and we choose $\beta = 0.4$ and $\gamma = 0.1$ so that the trading rate per second in our example does not exceed that for NYSE stocks.[23] Finally, the tick size is set at one cent,

[21] Matlab code for the simulations is provided in the Internet Appendix.

[22] In our example, the total gains from trade represent 8% (20/250) of the total value of the asset. This specification is based on Hollifield et al. (2006). Using data from the Vancouver Stock Exchange, they estimate that gains from trade in a limit order market vary between 6% and 9% of the common value of the asset (see their Table X, on p. 2790).

[23] Chordia, Roll, and Subrahmanyam (2011) report that the average value-weighted daily number of trades on the NYSE in 2008 was about 90,000, or about 230 trades per minute (assuming 250 trading days and 6.5 trading hours per day). They also report that the average trade size in 2008 was about $10,000. Given an average share price of $87 in January 2008 (from CRSP), this implies an average trade size of 115 shares. Thus, the average number of shares traded per minute on NYSE during 2008 was about $230 \times 115 = 26{,}450$, which is equivalent to about 440 shares per second.

$\Delta = 1$, as in U.S. equity markets, and we consider three possible values for the makers' spread, $a - v_0$, namely, 5, 10, or 15 ticks.

For each value of the makers' spread, Table II reports the optimal make and take fees for the platform and its annualized revenue, assuming that 2,000 stocks trade on the platform (slightly below the number of stocks listed on the NYSE in 2008) and that there are 250 trading days of 6.5 hours per year. Further, to show the importance of optimally differentiating make and take fees for the platform, we compare its revenue when make and take fees are set optimally to its revenue under two benchmark scenarios: (i) "Uniform Fees," where the platform follows a "naive" pricing strategy, dividing its fee equally between makers and takers; and (ii) "Capped Fees," where the platform chooses its fees optimally under the constraint that the take fee does not exceed 0.3 cents per share, the maximum allowed by Regulation NMS in the United States.

Consider first the case in which the makers' spread is equal to five ticks. In this case, with the uniform pricing scheme, market makers' aggregate monitoring intensity is relatively small because (i) they obtain a relatively small fraction (about 25%) of the gains from trade since the spread is tight, (ii) they have relatively large monitoring costs ($\beta > \gamma$), and (iii) they are fewer than market takers ($M < N$). Hence, as implied by Proposition 3, the trading platform can increase its trading rate by reducing its make fee while increasing its take fee. In fact, the optimal pricing policy in this case requires that market makers be subsidized. This results in a trading rate of about 143 trades per second, which is much higher than the trading rate achieved by the platform with uniform pricing (84 trades per second). For this reason, the difference in annual revenue when the platform chooses the optimal make/take fees and when it does not is large ($621 million).

As market makers' spread increases from 5 to 10 or 15 ticks, they obtain a higher fraction of gains from trade, the make/take fees being fixed. Hence, their incentive to monitor the market increases. Accordingly, the trading platform optimally charges a higher make fee and reduces its take fee. In fact, when the makers' spread is equal to 15 ticks, it is optimal for the platform to subsidize market takers rather than market makers.

When the take fee is capped at 0.3 cents, the platform cannot offer a rebate greater than 0.2 cents to the makers since its total trading fee is fixed at 0.1 cents per share. When the makers' spread is relatively small (5 or 10 ticks), this constraint is binding. Thus, the trading rate when the take fee is capped is smaller than when the platform can choose its fees freely, but greater than in the uniform pricing case.

Table II also shows that the optimal make and take fees can become large when they are unconstrained. In reality, one would expect that market makers' offers would neutralize large variations in the make/take fees. For example, if market makers receive a large subsidy, they may bid more aggressively. We address this issue in Section V, where we show that, even when market makers' offer prices are endogenous, differentiating make/take fees remains optimal for the platform, as long as the tick size is positive.

Table II

The Effect of Make/Take Fees on the Trading Rate and Fee Revenue

This table reports the trading fees charged by the platform (make, take, and total), the trading rate on the trading platform, and the annualized revenues of the platform for various bid-ask spreads and various scenarios for make/take fees: (i) "uniform": the total fee is equally split between makers and takers, (ii) "optimal": the make and take fees are optimally chosen by the platform, and (iii) "capped": make and take fees are optimally chosen by the platform under the constraint that the take fee does not exceed 0.3 cents per share. Annualized revenues are computed assuming that there are 2,000 stocks traded on the platform, with 250 trading days per year, each consisting of 6.5 trading hours. Parameter values are $M = 10$, $N = 20$, $\Gamma = 20$, $v_0 = 250$, $\beta = 0.4$, $\gamma = 0.1$, $\Delta = 1$, and $\overline{c} = 0.1$.

	$a - v_0 = 5$			$a - v_0 = 10$			$a - v_0 = 15$		
Spread (in ticks)	Uniform Fees	Optimal Fees	Capped Fees	Uniform Fees	Optimal Fees	Capped Fees	Uniform Fees	Optimal Fees	Capped Fees
Make Fee (cents/share)	0.05	−8.19	−0.2	0.05	−3.19	−0.2	0.05	1.8	1.8
Take Fee (cents/share)	0.05	8.29	0.3	0.05	3.29	0.3	0.05	−1.7	−1.7
Total Fee (cents/share)	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Trading Rate (shares/sec)	83.6	142.5	86.8	131.8	142.5	133.3	138.3	142.5	142.5
Estimated Annual Platform Fee Revenue ($ in millions)	979	1,667	1,016	1,524	1,667	1,559	1,618	1,667	1,667

C. Welfare, Algorithmic Trading, and Make/Take Fees

Algorithmic trading is often portrayed as socially useless.[24] In contrast, our model captures one social benefit of algorithmic trading: by reducing monitoring costs, it increases the speed at which traders are matched and gains from trade are realized. To see this, let W be the aggregate welfare of all market participants, including the platform. Using equations (10), (11), and (12), we have

$$W(\gamma, \beta, c_m, c_t, M, N) = \mathcal{R}\left(\bar{\mu}^*, \bar{\tau}^*\right) \cdot \Gamma - M \cdot C_m(\mu_1^*) - N \cdot C_t(\tau_1^*). \tag{23}$$

Other things being equal, aggregate welfare increases in the trading rate. Corollary 1 shows that a decrease in traders' monitoring costs on either side results in a higher trading rate. For this reason, for fixed trading fees, a decrease in monitoring costs for one side is a Pareto improvement: it makes all participants better off and therefore also increases total welfare. We provide a formal proof of this intuitive result in the Internet Appendix.

More surprisingly, this conclusion does not necessarily hold when we account for the effects of the reduction in monitoring costs on the optimal make/take fee breakdown. Indeed, in this case, the side with declining monitoring costs can sometimes be worse off. To see this point, suppose, for example, that market makers' monitoring cost, β, decreases. As discussed previously, the trading platform optimally reacts to this decline by increasing the make fee and decreasing the take fee. As the total fee is unchanged and the take fee is smaller, market takers are clearly better off. In contrast, the change in market makers' welfare is ambiguous. On the one hand they incur lower monitoring costs but on the other hand they pay higher fees. The net effect on their expected profit can be either positive or negative.

As an example, suppose that $M = N = 1$. In this case, in equilibrium, the market maker's expected profit per unit of time when fees are set optimally is[25]

$$\Pi_{im}(\beta, \gamma) = \beta^{\frac{1}{4}} \times \frac{(\Gamma - \bar{c})^2\left(\beta^{\frac{1}{4}} + 2\gamma^{\frac{1}{4}}\right)}{4\left(\beta^{\frac{1}{4}} + \gamma^{\frac{1}{4}}\right)^6}. \tag{24}$$

The market maker's expected profit decreases in γ. That is, a reduction in the market taker's monitoring cost always benefits the market maker. In contrast, the market maker's expected profit vanishes as her monitoring cost goes to zero since the platform reacts to this decline by charging a higher make fee. This example underscores the importance of accounting for changes in trading platforms' pricing policies in analyzing welfare effects of algorithmic trading.

[24] See, for instance, Paul Krugman, "Rewarding bad actors," *New York Times*, August 2, 2009.

[25] The expression for market-makers' expected profit does not depend on market-makers' spread because the division of gains from trade does not depend on this spread when make/take fees are set optimally by the platform.

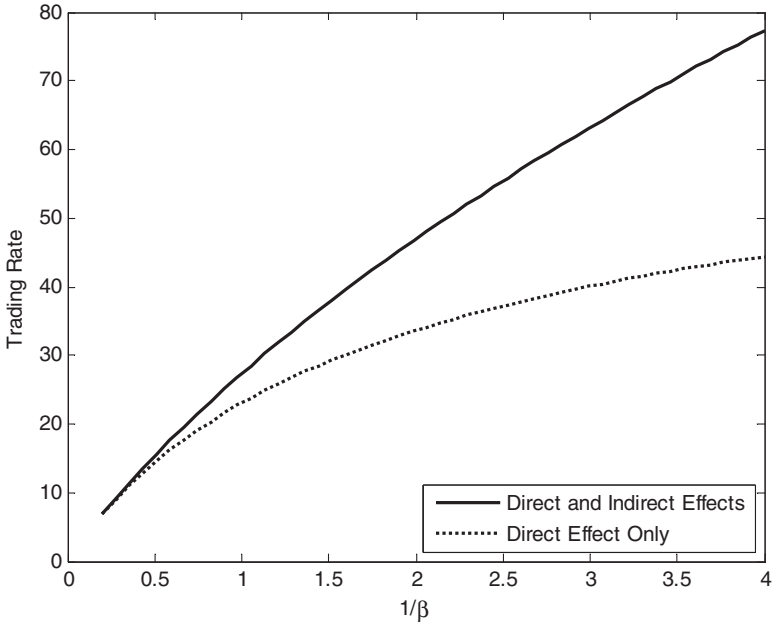

Figure 3. Algorithmic trading and the trading rate. The figure plots the equilibrium trading rate as a function of the reciprocal of market makers' monitoring costs, $1/\beta$. Parameter values are $v_0 = 300$ (expected payoff), $a = 304$ (ask price), $M = 10$ (number of market makers), $N = 20$ (number of market takers), $\Gamma = 25$ (gains from trade), $\Delta = 1$ (tick size), $\gamma = 0.5$ (market takers' monitoring costs), $c_m = -0.2$ (make fee), and $c_t = 0.3$ (take fee). All monetary values are in cents. The solid curve depicts the equilibrium trading rate. The dotted curve shows the trading rate for a fixed aggregate monitoring level of the market takers (fixed at its level when $1/\beta = 0.25$).

IV. Empirical Implications

A. Algorithmic Trading, Volume, and Liquidity

Corollary 1 shows that a decrease in the monitoring cost for market makers or market takers triggers an increase in the trading rate. Thus, algorithmic trading should be associated with an increase in the trading rate. This association can be particularly strong because it is amplified by the complementarity in monitoring decisions between market makers and market takers. That is, algorithmic trading causes an increase in the trading rate via a direct channel, the reduction in monitoring costs, and an indirect channel, liquidity externalities.

To see this point, consider Figure 3. This figure illustrates how a reduction in monitoring costs for market makers (β) affects the trading rate $\mathcal{R}$ for fixed values of the other parameters: $v_0 = 300$, $a = 304$, $M = 10$, $N = 20$, $\Gamma = 25$, $\Delta = 1$, $\gamma = 0.5$, $c_m = -0.2$, and $c_t = 0.3$ (all monetary values are in cents). The solid curve depicts the equilibrium trading rate when $1/\beta$ increases from 0.2 to four (β decreases from 5 to 0.25). It accounts for both the direct and the indirect

channels. That is, a decrease in market makers' monitoring costs leads to more monitoring by market makers, which prompts market takers to monitor more and therefore amplifies the initial effect through a chain reaction. In contrast, the dotted curve shows the evolution of the trading rate when market makers' monitoring cost decreases for a fixed aggregate monitoring level of the market takers (fixed at its level when $\beta = 4$). Hence, the dotted curve shows the evolution of the trading rate when liquidity externalities are turned off.

In both cases, the trading rate increases when market makers' monitoring cost decreases. However, in equilibrium (solid curve), the trading rate increases at a much faster rate because of the complementarity in market makers' and market takers' monitoring decisions.[26] This could explain why the rate of increase in trading volume has been so steep in recent years. For instance, from 2005 to 2007, the number of shares traded on the NYSE rose by 111% despite the fact that NYSE market share has declined over the same period. This evolution is mainly driven by an increase in the trading rate since the size of trades has steadily declined in recent years (see Chordia, Roll, and Subrahmanyam (2011)).

Brogaard (2010) finds cross-sectional variations in the proportion of time high frequency traders set the inside quote on NASDAQ. Our model can shed light on the determinants of this proportion. To see this, let $\hat{a} \geq v_0 + \Gamma$ be the offer posted in the market when there is no offer posted at price a. Offers at $\hat{a}$ can be seen as being posted by non-high-frequency market makers (e.g., human traders) for whom the cost of providing liquidity is higher than for market makers in our model.[27] Thus, they post offers that are not attractive for market takers.

Between trades, the offer price is a for an average duration of $\mathcal{D}_t$ and $\hat{a}$ for an average duration of $\mathcal{D}_m$. The fraction of time during which market makers set the inside quote is therefore

$$\phi \equiv \frac{\mathcal{D}_t}{\mathcal{D}_m + \mathcal{D}_t} = \frac{\mathcal{V}^*}{1 + \mathcal{V}^*}. \tag{25}$$

Clearly, ϕ increases in the velocity ratio, $\mathcal{V}^* = \frac{\mathcal{D}_t}{\mathcal{D}_m} = \frac{\bar{\mu}^*}{\bar{\tau}^*}$, and can therefore serve as a proxy for this ratio. Corollary 3 implies that, for fixed fees, the velocity ratio should increase in $q = M/N$ and $r = \gamma/\beta$. Hence, a decrease in market takers' relative monitoring cost, $r = \gamma/\beta$, reduces the fraction of time market makers set the inside quote in equilibrium. Similarly, a decrease in q implies that the market spends less time in this state. Thus, other things being equal, high frequency market makers should set the inside spread less frequently

[26] In this discussion, we take the make/take fees as fixed. The effect of algorithmic trading on the trading rate is even stronger when fees are set optimally. The reason is that the trading platform adjusts its fees so as to maximize the trading rate, contributing even further to the surge in volume when monitoring costs decline.

[27] For NYSE stocks, Hendershott, Jones, and Menkveld (2011) find evidence consistent with the fact that algorithmic liquidity suppliers have a competitive cost advantage over more traditional liquidity suppliers (like the specialist).

in stocks in which their number is small relative to liquidity demand.[28] This observation implies that the effect of algorithmic trading on the time-weighted average bid-ask spread is ambiguous. Indeed, the time-weighted average half bid-ask spread (denoted ES) is

$$ES = \phi a + (1-\phi)\hat{a} - v_0 = \hat{a} - v_0 - \phi(\hat{a} - a), \tag{26}$$

which decreases with ϕ. As we just explained, a reduction in market makers' monitoring cost increases ϕ while a reduction in market takers' monitoring cost reduces ϕ. Thus, investments in algorithmic trading technologies by the market-making side lower the time-weighted bid-ask spread. In contrast, investments in algorithmic trading by the market-taking side increase the time-weighted bid-ask spread.

Hendershott, Jones, and Menkveld (2011) study a change in NYSE market structure ("Autoquote") that reduces the cost of monitoring for algorithmic liquidity suppliers. They write (p. 13): "Autoquote allowed algorithmic liquidity suppliers to, say, quickly notice an abnormally wide inside quote and provide liquidity accordingly via a limit order." This change is similar to a reduction in the cost of monitoring for market makers, β, in our model. Hendershott, Jones, and Menkveld (2011) find empirically that this change reduced the average bid-ask spread, as predicted by our model. In contrast, Hendershott and Moulton (2011) study a change in the organization of the NYSE that increases the speed at which liquidity demanders can react to new quotes. They find an increase in the average bid-ask spread following this event, which again is in line with the logic of the model.

B. Liquidity Externalities and Time

B.1. Identifying Liquidity Externalities

As explained previously, our model features a new type of liquidity externality: the monitoring decisions of the market-making side and the market-taking side reinforce each other. As stressed by Barclay and Hendershott (2004), the empirical identification of liquidity externalities is nontrivial. In our framework, the challenge is to establish that a greater monitoring intensity by one side has a positive causal effect on the monitoring intensity of the other side. To identify this relation one needs to find exogenous shocks that directly affect the monitoring intensity of one side without affecting the other side.

Monitoring intensities are not directly observable. However, the average durations of make and take phases can serve as proxies for these intensities, since $\mathcal{D}_m = \frac{1}{\bar{\mu}^*}$ and $\mathcal{D}_t = \frac{1}{\bar{\tau}^*}$. That is, one can test for the presence of cross-side externalities by checking whether an increase in the *average* duration of the make phase, $\mathcal{D}_m$, has a positive causal effect on the average duration of the take phase, $\mathcal{D}_t$, and vice versa.[29] Our model suggests several variables that

[28] This implication also holds when fees are set at their optimal level.

[29] Coopejans, Domowitz, and Madhavan (2001) capture cross-side complementarities using a

could serve as instruments for this test. To see this, it is useful to express the average duration of the make phase as a function of the average duration of the take phase and vice versa.

PROPOSITION 4: *When traders choose their monitoring levels optimally, there exist two functions* $f(\cdot)$ *and* $g(\cdot)$ *such that we can write* $\mathcal{D}_m = f(\mathcal{D}_t; \beta, M, c_m)$ *and* $\mathcal{D}_t = g(\mathcal{D}_m; \gamma, N, c_t)$. *Furthermore*,

(i) *The function* $f(\cdot; \beta, M, c_m)$ *is increasing in* $\mathcal{D}_t$ *and the function* $g(\cdot; \gamma, N, c_t)$ *is increasing in* $\mathcal{D}_m$.
(ii) *The function* $f(\mathcal{D}_t; \cdot)$ *increases in market makers' monitoring cost and the make fee while it decreases in the number of market makers. The function* $g(\mathcal{D}_m; \cdot)$ *increases in market takers' monitoring cost and the take fee while it decreases in the number of market takers.*

The second part of the proposition shows that parameters c_m, β, and M can be used as instruments to identify the effect of the duration of the make phase on the duration of the take phase: changes in these variables directly affect $\mathcal{D}_m$ without directly affecting $\mathcal{D}_t$. In a symmetric way, the parameters c_t, γ, and N can be used as instruments to identify the effect of the duration of the take phase, $\mathcal{D}_t$, on the duration of the make phase, $\mathcal{D}_m$.

The number of market-making firms in a stock, especially high frequency market makers, is a natural proxy for M in our model. Exogenous shocks to this number could be used to test whether the speed of reaction of the market-making side ($\mathcal{D}_m$) has a positive effect on the speed of reaction of the market-taking side ($\mathcal{D}_t$). For instance, Chaboud et al. (2010) study algorithmic trading on EBS (an electronic market for currencies) and exploit observable exogenous changes in the number of algorithmic market makers on this platform to measure the effects of algorithmic trading on volatility. Their data could also be used to perform the test we just described.

Another approach involves using technological changes that reduce the cost of monitoring for market makers or market takers. As mentioned previously, Hendershott, Jones, and Menkveld (2011) study the liquidity effects of a technological change on the NYSE that is similar to a reduction in the cost of monitoring for market makers, β, in our model. Hence, this change could also be used to test whether an increase in market makers' speed of reaction triggers an increase in market takers' speed of reaction. Similarly, improvements in smart routing technologies (a reduction in γ) could be used to identify the impact of an increase in market takers' speed of reaction on market makers' speed of reaction.

In reality, fluctuations in interevent durations are also driven by other factors, such as changes in asset values. This source of fluctuation is not captured by our model since we take the asset value as given.

VAR model for depth on both sides of the limit order book. Here, we propose to quantify cross-side complementarities using interevent durations.

B.2. Explaining Duration Clustering

Modeling time between events is very important for the empirical analysis of high frequency data (see Engle (2000)). As a result, this topic has generated a voluminous literature in financial econometrics (see Pacurar (2006) for a survey). One challenge for research in this area is to explain the clustering in durations that is pervasive in trade and quote data (see Engle and Russell (1998)). That is, long (short) durations between orders and trades tend to be followed by long (short) durations between the same events.

Duration clustering is usually interpreted through the lens of information-based models, in particular, Easley and O'Hara (1992). In this model, the trading rate is higher when there is an information event. One implication is that short durations between trades signal the presence of informed investors, leading to an inverse relationship between price impacts and durations between trades. However, Engle and Russell (1998) find that short durations have no effect on price impacts when the bid-ask spread is small. They conclude (p. 1158): "this suggests that both liquidity- and information-based clustering of transaction rates occur."

In line with this finding, our model suggests a liquidity-based explanation for duration clustering. To see this, consider a positive shock on the "demand for trading" from the market taking side (an increase in N). The direct effect of this shock is to increase market takers' aggregate monitoring level, which reduces the average duration from a quote to a trade, $\mathcal{D}_t$. In turn, market makers monitor the market more intensively since they expect market takers to hit their quotes more quickly. Thus, there is also a decline in the duration from a trade to a quote, $\mathcal{D}_m$. More generally, any change in a factor that directly affects the monitoring intensity of one side without directly affecting the other side induces a change in interevent durations ($\mathcal{D}_t$ and $\mathcal{D}_m$) in the same direction (see Proposition 4). Thus, time-series fluctuations in these factors (e.g., the number of market takers) induce a positive correlation between interevent durations and therefore a clustering in durations between trades.

V. Robustness and Extensions

In this section we check whether the main results of the baseline model are robust to relaxing some of the simplifying assumptions.[30]

A. Finite Arrival Rate for Market Takers' Trading Needs

In the baseline model we assume that, after a trade, market takers immediately receive a new buy order to execute. In this section we relax this assumption. That is, we assume that, after a trade, a market taker receives a new buy order for one share after a waiting time, which is exponentially distributed with intensity $\kappa > 0$. The baseline model is then the special case

[30] We are grateful to the anonymous referee for motivating much of the analysis in this section.

in which $\kappa = \infty$. Unfortunately, when $\kappa < \infty$, the model becomes intractable when either $M > 1$ or $N > 1$.[31] Hence, in this section, we must restrict our attention to the case in which there is one market maker and one market taker ($M = N = 1$).

Suppose that a trade just took place. It takes on average $\frac{1}{\kappa}$ units of time before the market taker has a new need to buy the security and $\frac{1}{\mu_1} + \frac{1}{\tau_1}$ units of time for this order to execute. Therefore, the average duration of a cycle is now

$$\mathcal{D}(\mu_1, \tau_1, \kappa) = \frac{1}{\mu_1} + \frac{1}{\tau_1} + \frac{1}{\kappa}, \tag{27}$$

and the trading rate is

$$\mathcal{R}(\mu_1, \tau_1, \kappa) = \frac{1}{\frac{1}{\mu_1} + \frac{1}{\tau_1} + \frac{1}{\kappa}}. \tag{28}$$

For fixed monitoring levels, the trading rate increases in κ and is bounded by κ: market makers and market takers cannot be matched faster than the rate at which trading needs occur. In general, they are matched at a smaller rate because traders' monitoring levels are finite. This is the main source of inefficiency in the model, and it is precisely this inefficiency that make/take fees help to alleviate since the platform chooses its make/take fee breakdown to maximize the trading rate. As expected, when κ goes to infinity, the trading rate converges to its value in the baseline model (equation (7)).

The objective functions of the market maker and the market taker are as given in the baseline model, except that the expression for the trading rate is now given by equation (28). In contrast to the baseline case, we cannot solve for traders' monitoring levels in closed form.[32] However, we show analytically in the Internet Appendix that our comparative statics regarding traders' monitoring levels (Corollary 1) are valid for all values of κ. The next proposition (proved in the Internet Appendix) characterizes the optimal pricing policy of the platform for all values of κ and shows that this policy has the same properties as those obtained in the baseline case.

PROPOSITION 5: *Assume $M = N = 1$. The optimal make/take fee breakdown does not depend on κ. In addition, when $a - v_0 = s \cdot \Delta$, the optimal make and*

[31] To see this, recall that a market taker's expected profit per cycle depends on his probability of being first to hit an offer when he has a buy order to execute. When $N = 1$, this probability is equal to one. In contrast, when $N > 1$, this probability depends on the number of other market takers with a buy order to execute. This number is random, preventing us from writing down in a simple way the objective function of a market taker when $\kappa < \infty$ unless $N = 1$. This problem does not arise in the baseline model ($\kappa = \infty$) because, after a trade, a market taker immediately receives a new buy order. Thus, at any point in time, the number of market takers with a share to buy is N. When $N = 1$ and $M > 1$ the model is also not tractable since we cannot solve for traders' monitoring intensities as in Proposition 2.

[32] The closed-form solutions for the monitoring intensities in the baseline model when $M = N = 1$ is given in the Internet Appendix.

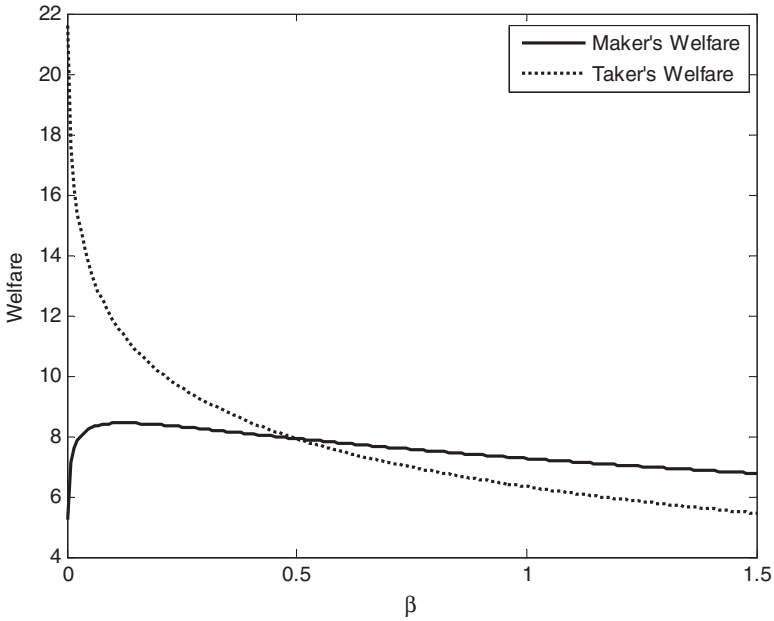

Figure 4. Monitoring costs and welfare. The figure plots the market maker's expected profit (solid line) and the market taker's expected profit (dashed line) in equilibrium as a function of the maker's monitoring cost (β). Parameter values are $v_0 = 300$ (expected payoff), $a = 304$ (ask price), $\Gamma = 25$ (gains from trade), $\Delta = 1$ (tick size), $\bar{c} = 0.1$ (total fee), $\gamma = 0.5$ (market taker's monitoring cost), and $\kappa = 2$ (arrival rate of trading needs). All monetary values are in cents.

take fees are

$$c_{ms}^* = s \cdot \Delta - \frac{\Gamma - \bar{c}}{1 + r^{\frac{1}{4}}} \quad \textit{and} \quad c_t^* = \bar{c} - c_m^*. \tag{29}$$

As in the baseline model and for the same reason, the optimal make fee declines in the relative monitoring cost of the market-taking side (r). An increase in κ increases the incentive to monitor the market for both market makers and market takers. For this reason, a change in κ does not alter the velocity ratio and the optimal make/take fee breakdown does not depend on κ.

In the baseline model ($\kappa = \infty$), a decrease in monitoring cost does not necessarily result in a Pareto improvement when fees are set optimally (see Section III.C). Figure 4 shows that this result continues to hold when $\kappa < \infty$. The figure plots the evolution of the market maker's expected profit (solid line) and the market taker's expected profit (dashed line) in equilibrium as a function of the maker's monitoring cost (β). Parameter values are $v_0 = 300$, $a = 304$, $\Gamma = 25$, $\Delta = 1$, $\bar{c} = 0.1$, $\gamma = 0.5$, and $\kappa = 2$ ($M = N = 1$ and all monetary values are in cents). As can be seen, the market maker's expected profit peaks for a strictly positive value of β. Thus, for small values of β, the market maker is worse off when its monitoring cost declines. Indeed, as in the baseline case, the platform

reacts to this decline by charging a higher make fee and this effect more than offsets the benefit of a smaller monitoring cost for the market maker.

B. Fast and Slow Traders

In the baseline model we assume that all traders on one side have identical monitoring costs. Hence, in equilibrium, they react with the same latency to a profit opportunity. In reality, latencies often differ across market makers because some have a technological edge over others. What are the effects of asymmetries in market makers' speed of access to the market?

To address this question, we consider the effect of reducing the monitoring cost of market maker 1 relative to other market makers (i.e., $\beta_1 < \beta_j$, for $j \neq 1$). When market makers' monitoring costs are heterogeneous, we cannot solve for equilibrium monitoring intensities in closed form. However, we can obtain a numerical solution for these intensities when the total number of traders $M+N$ is not too large.[33]

Consider the following numerical example. There are two market makers ($M=2$) and one market taker ($N=1$). Furthermore $v_0 = 300$, $a = 304$, $\Gamma = 25, \Delta = 1$, $c_m = -0.2$, and $c_t = 0.3$ (all monetary values are in cents). In Figure 5, we show the effect of decreasing the monitoring cost of market maker 1 (increasing $1/\beta_1$) on the equilibrium monitoring intensities of each trader (Figure 5A), the trading rate (Figure 5B), and the market share of each market maker (Figure 5C).[34] For this analysis we fix the monitoring costs of the other traders ($\beta_2 = 0.5$ and $\gamma_1 = 0.5$).

Not surprisingly, as market maker 1's monitoring cost declines, she monitors the market more intensively. More interestingly, as explained in Section I.D, this effect reduces the marginal expected return on monitoring for her competitor. For this reason, market maker 2's monitoring intensity becomes smaller as market maker 1's monitoring cost declines (Figure 5A). Accordingly, the market share of the fast market maker increases at the expense of the slow market maker who is progressively crowded out of the market (Figure 5C). Yet, the aggregate monitoring intensity of the market making side increases. This effect exerts a positive externality on the market taker who reacts by increasing his monitoring intensity, and ultimately the trading rate becomes higher (Figure 5B).

Thus, as high frequency market makers become increasingly fast, we should observe a simultaneous increase in the trading rate and an increase in the market share of high frequency market makers. In line with these predictions, many analysts have noticed that high frequency traders are responsible for

[33] Traders' equilibrium monitoring intensities solve a system of $M+N$ nonlinear equations corresponding to the first order conditions of the problem. In general, we cannot solve this system analytically, unless, as in the baseline model, all traders operating on one side have identical monitoring costs. Indeed in this case, the system of $M+N$ equations boils down to just two equations (see the proof of Proposition 2).

[34] The market share of each market-maker (the average fraction of trades in which she participates) is given by $\frac{\mu_i}{\mu_1+\mu_2}$, $i = 1, 2$.

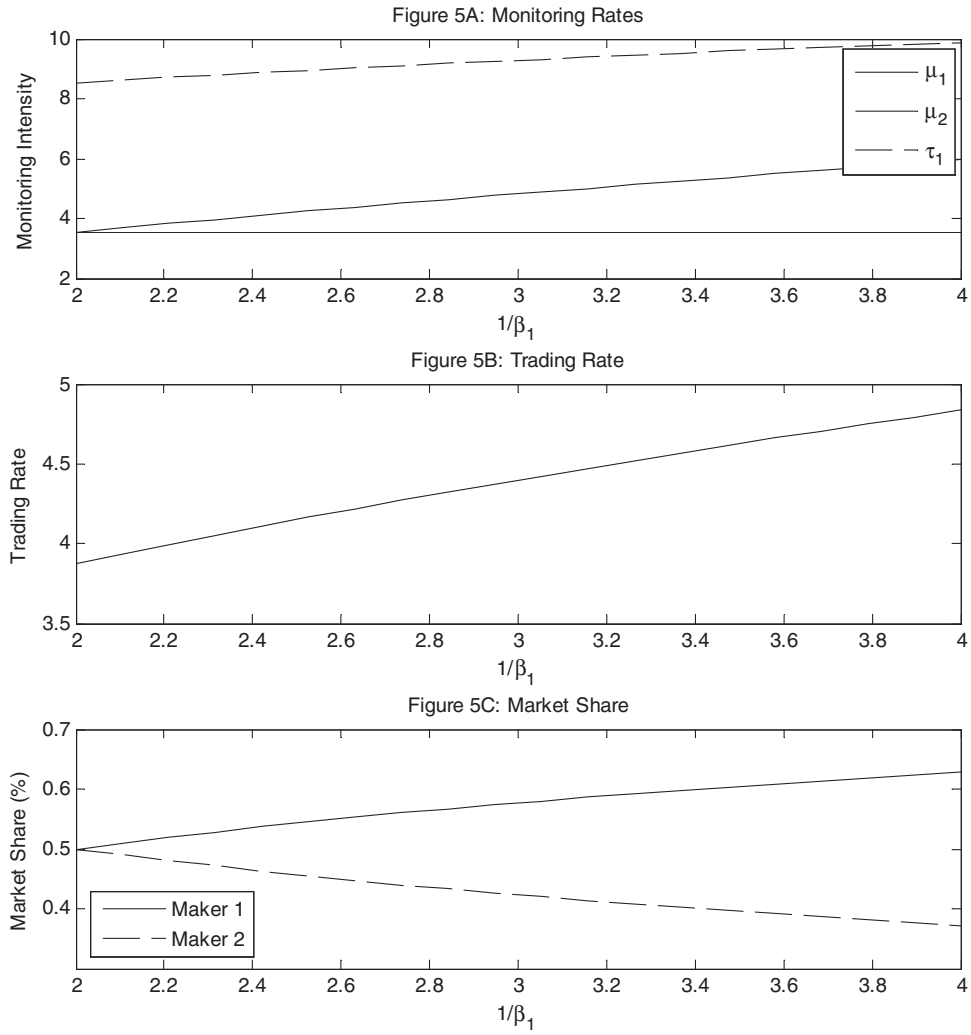

Figure 5. The effect of heterogeneous monitoring costs on monitoring, the trading rate, and market shares. Parameter values are: $M = 2$ (number of market makers), $N = 1$ (number of market takers), $v_0 = 300$ (expected payoff), $a = 304$ (ask price), $\Gamma = 25$ (gains from trade), $\Delta = 1$ (tick size), $c_m = -0.2$ (make fee), $c_t = 0.3$ (take fee). All monetary values are in cents. The horizontal axis is the reciprocal of the monitoring costs of market maker 1 ($1/\beta_1$).

an increasing fraction of trading volume (73% according to "SEC run eye over high-speed trading," *Financial Times*, July 29, 2009).

C. Endogenous Transaction Prices

So far we have fixed exogenously the price at which market makers trade with market takers. We now endogenize this price. In this way, we can study

whether the results hold when transaction prices adjust following a change in fees.

As market makers inspect the market at stochastic points in time, one may first make an offer at one price, which is subsequently improved by another market maker, and so on. Modeling this auction is beyond the scope of this paper. We take a simpler approach to model how gains from trade are divided between market makers and market takers. In particular, we assume that this division is given by the Nash bargaining solution in which market takers' market power is measured by $\theta \in (0, 1)$, under the constraint that the price at which market makers and market takers trade belongs to the grid of feasible prices $\mathcal{P}$ (see (1)). That is, the transaction price now solves

$$\max_{a \in \mathcal{P}} \mathcal{O}(a, c_m, \theta) \equiv \pi_t(a, c_m)^{\theta} \pi_m(a, c_m)^{1-\theta}, \tag{30}$$

where $\pi_m(a, c_m) = a - v_0 - c_m$ and $\pi_t(a, c_m) = v_0 + \Gamma - a - \bar{c} + c_m$ are the per trade profits of makers and takers, respectively. We denote by $a^*(c_m, \theta)$ the solution to (30), that is, the transaction price at which the traders agree to trade given the value of the make fee.

For fixed values of the fees, all the findings regarding traders' monitoring decisions (see Section II) are still valid. In particular, traders' monitoring decisions depend on their profit per trade, that is, $\pi_m(a^*(c_m, \theta), c_m)$ or $\pi_t(a^*(c_m, \theta), c_m)$, as in equations (14) and (15) in the baseline model.

We now turn our attention to the optimal fees set by the trading platform. As a benchmark, it is useful to first consider the polar case in which the tick size is zero. In this case the solution to equation (30) is

$$a^*(c_m, \theta) = v_0 + c_m + (1 - \theta)(\Gamma - \bar{c}), \tag{31}$$

and the division of gains from trade, $\Gamma - \bar{c}$, between makers and takers is

$$\pi_m(a^*(c_m, \theta), c_m) = (1 - \theta)(\Gamma - \bar{c}), \tag{32}$$

$$\pi_t(a^*(c_m, \theta), c_m) = \theta(\Gamma - \bar{c}). \tag{33}$$

The platform controls the ask price at which market makers and market takers trade, since there is a one-to-one mapping between the ask price and the make fee (equation (31)). However, when the tick size is zero, the make fee does not affect the share of the gains from trade $(\Gamma - \bar{c})$ captured by the market makers, which only depends on their relative market power, $1 - \theta$ (see equation (32)). Indeed, when the tick size is zero, traders fully neutralize the effect of a change in the make fee on the division of gains from trade by a one-for-one adjustment in the transaction price. For example, a one cent decrease in the make fee combined with a one cent increase in the take fee is fully neutralized by a one cent decrease in the ask price (so that, cum fee, the price paid by takers is unchanged). Accordingly, the platform cannot affect traders' monitoring decisions and the trading rate using its make and take fees. Hence, the make/take fee breakdown is neutral, as claimed by

Angel, Harris, and Spatt (2011). We state this irrelevance result in the next proposition.

PROPOSITION 6: [Benchmark] *When the tick size is zero, the make/take fee breakdown has no effect on monitoring decisions or the trading rate.*

This is unfortunate since, in general, the division of gains from trade achieved by traders when the tick size is zero does not maximize the trading rate: the side capturing a small share of the gains from trade relative to its monitoring cost tends to react too slowly to trading opportunities. By allocating a higher share of gains from trade to the relatively slow side, one could increase the speed at which transactions get executed.

This irrelevance result breaks down, however, when the tick size is strictly positive. Indeed, in this case, the ask price must be a multiple of the tick size. This prevents traders from fully neutralizing a change in make/take fees by adjusting the price at which they trade. Therefore, the platform can influence traders' monitoring decisions with its fees and raise the trading rate relative to the case in which the tick size is zero. Proposition 7 below (proved in the Internet Appendix) characterizes the optimal pricing policy of the platform for any value of the tick size. As in the baseline model, we fix the total fee at $\bar{c}$ and we focus on the breakdown of this fee between makers and takers. For given values of the parameters ($\Delta, r,$ and q), we denote the optimal make and take fees by $c_m^*(\Delta, r, q)$ and $c_t^*(\Delta, r, q)$, respectively. Recall that $\ell = \frac{\Gamma}{\Delta}$ and that c_{ms}^* is the optimal make fee from the baseline model in Section III.

PROPOSITION 7: *For each* $s \in \{1, 2, \ldots, \ell - 1\}$,

(i) *There exists a unique make fee* $\hat{c}_{ms}$ *such that, for all* $c_m \in [\hat{c}_{ms}, \hat{c}_{ms} + \Delta]$, *trades take place at* $a^*(c_m, \theta) = v_0 + s \cdot \Delta$.[35]

(ii) *The following pricing policy is such that trade takes place at price* $v_0 + s \cdot \Delta$, *and is optimal for the trading platform:*

$$c_m^*(\Delta, r, q) = \begin{cases} \hat{c}_{ms} & \text{if } c_{ms}^* \leq \hat{c}_{ms} \\ c_{ms}^* & \text{if } \hat{c}_{ms} < c_{ms}^* < \hat{c}_{ms} + \Delta \\ \hat{c}_{ms} + \Delta & \text{if } c_{ms}^* \geq \hat{c}_{ms} + \Delta \end{cases}$$

and $c_t^*(\Delta, r, q) = \bar{c} - c_m^*(\Delta, r, q)$.

In contrast to the case with zero tick size, the price at which makers and takers choose to trade on the grid does not increase one-for-one in the make fee (first part of the proposition). Indeed, suppose that the make fee is such that traders choose to trade at price $v_0 + s \cdot \Delta$. If the platform increases its make fee by, say, one tick, then traders neutralize the effect of this increase on the division of gains from trade by trading at a price one tick higher, that is, at $v_0 + (s + 1) \cdot \Delta$. However, traders cannot neutralize the change in the make fee

[35] For each value of s, the threshold $\hat{c}_{ms}$ is a function of θ (see the proof of Proposition 7 in the Internet Appendix). We do not make this relationship explicit to simplify notations.

if this change is small enough relative to the tick size as this would require trading at a price that is not on the grid. For this reason, the ask price at which traders choose to trade is increasing stepwise in the make fee rather than continuously.

To understand the second part of the proposition, suppose that the platform decides to induce makers and takers to trade at $v_0 + s \cdot \Delta$. Then, in choosing its make/take fees, the platform solves the same problem as in the baseline case with one additional constraint: the make fee must be in the interval $[\hat{c}_{ms}, \hat{c}_{ms} + \Delta]$, as otherwise traders will choose to trade at another price (first part of the proposition). This constraint is not binding if the make fee chosen in the baseline case when the market makers' spread is $s \cdot \Delta$, that is, c^*_{ms} is already in the interval $[\hat{c}_{ms}, \hat{c}_{ms} + \Delta]$. Thus, in this case the platform chooses the same fee as in the baseline case. Otherwise, the constraint is binding and the platform chooses the appropriate corner solution (either $\hat{c}_{ms}$ or $\hat{c}_{ms} + \Delta$).

The maximum trading rate that the platform can achieve does not depend on the price, $a^*(c_m, \theta)$, resulting from the choice of its make fee. Indeed, suppose that c^*_m is an optimal make fee. If the platform increases or decreases this fee by one tick, then traders will neutralize this shift by adjusting upward or downward by one tick the price at which they choose to trade. As a result, the division of gains from trade and therefore the trading rate are the same whether the platform sets its make fee at c^*_m, $c^*_m + \Delta$, or $c^*_m - \Delta$. Thus, the optimal pricing policy for the platform is defined up to one tick. In reality, trading platforms may have a preference for displaying small bid-ask spreads. Such a preference would pin down the optimal make fee uniquely in our model: for instance, if the platform wants a maker's spread equal to one tick then it must choose its optimal make fee in $[\hat{c}_{m1}, \hat{c}_{m1} + \Delta]$.

As in the baseline case, differentiating the make and take fees is optimal for the platform since, in general, $c^*_m(\Delta, r, q) \neq \frac{\bar{c}}{2}$. When the market is thick, c^*_{ms} is given explicitly by Proposition 3, and Proposition 7 yields a closed-form characterization of the optimal make/take fees. Furthermore, as $\hat{c}_{ms}$ does not depend on q and r, the following result is immediate.

COROLLARY 6: *In the thick market case, the optimal make fee for the platform* $(c^*_m(\Delta, r, q))$ *weakly increases with the relative size of the market-making side*, q, *and the relative monitoring cost for the market-taking side*, r.

Thus, our baseline results regarding the effects of a change in r and q on the optimal take/make spread in the thick market case are robust when the makers' spread is endogenous.[36] As in the baseline case, when the market is not thick, we cannot obtain a closed-form solution for the optimal make/take fees for arbitrary values of M and N. However, numerical simulations show that the results of Corollary 6 are valid even when the market is not thick.

[36] A reduction in the tick size, Δ, affects the optimal make/take fee breakdown, but the direction of the effect can be positive or negative depending on parameter values.

Table III

Trading Rate, Welfare, and Fee Revenue for Different Tick Sizes

This table reports the trading rate, the expected welfare of each category of participants per second (the platform, the makers, and the takers), the sum of makers and takers' welfare, the total welfare (the sum of all participants' welfare per second), and the annualized revenue of the platform for three different levels of the tick size and two scenarios for make/take fees: (i) "uniform": the total fee is equally split between makers and takers, and (ii) "optimal": the make and take fees are optimally chosen by the platform. Annualized revenues are computed assuming that there are 2,000 stocks traded on the platform, with 250 trading days per year, each consisting of 6.5 trading hours. Parameter values are $M = 10$, $N = 20$, $\Gamma = 50$, $v_0 = 600$, $\beta = 0.4$, $\gamma = 0.1$, $\Delta = 1$, and $\bar{c} = 0.1$.

	Tick Size = \$1/8		Tick Size = \$1/16		Tick Size = \$1/100	
	Uniform Pricing	Optimal Pricing	Uniform Pricing	Optimal Pricing	Uniform Pricing	Optimal Pricing
Trading Rate (shares/second)	330.6	355.7	330.6	346.4	330.6	333.4
Platform's Profits (\$/second)	0.3306	0.3557	0.3306	0.3464	0.3306	0.3334
Makers' Welfare (per second)	42.3	57.2	42.3	49.9	42.3	43.5
Takers' Welfare (per second)	42.8	34.4	42.8	39.2	42.8	42.3
Makers and Takers Welfare (per second)	85.1	91.6	85.1	89.1	85.1	85.8
Total Welfare (per second)	85.4	92.0	85.4	89.5	85.4	86.1
Estimated Annual Platform Fee Revenue (\$ in millions)	3,868	4,162	3,868	4,053	3,868	3,901

To sum up, as long as the tick size is not zero, the make/take pricing model has true economic consequences: it affects monitoring intensities, the trading rate, and market participants' welfare. We illustrate this point with a numerical example. Assume that $M = 10$, $N = 20$; $\beta = 0.4$, $\gamma = 0.1$, $\theta = 0.5$, $\Gamma = 50$, $v_0 = 600$, and $\bar{c} = 1/10$ (all monetary amounts in cents). Table III provides the aggregate welfare per unit of time of each market participant and the sum of these welfares ("Total Welfare") for various values of the tick size. As in Table II, the time unit is one second, and we annualize trading revenues for the platform using the same assumptions as in Section III.B. For each value of the tick size, we compare the case in which there is no differentiation of make/take fees ("uniform pricing") and the case in which make/take fees are chosen optimally by the platform ("optimal pricing").

The parameters for Table III are such that, with uniform pricing, market makers' aggregate monitoring intensity is smaller than market takers' monitoring intensity. Indeed, both sides have equal market power but market makers bear a higher monitoring cost and they are fewer. It is therefore optimal for the platform to lower the make fee while increasing the take fee relative to the case in which there is no differentiation in make and take fees. In this way, the platform increases the trading rate by reducing the gap in the aggregate monitoring intensities of both sides. For this reason, in Table III market makers' welfare is higher when the platform sets its fees optimally compared to the case with uniform pricing. In contrast, market takers'

welfare is smaller since they end up paying a higher fraction of the total fee. However, market makers' welfare gains more than offsets market takers' welfare loss. Thus, the change in the make/take fee breakdown is not just a redistribution of gains from trade from takers to makers. It raises aggregate welfare by increasing the rate at which takers are matched with makers. Thus, restricting trading platforms' ability to differentiate their make and take fees by, for instance, capping the take fee can impair investors' aggregate welfare.

Table III further shows that the differentiation of make/take fees has greater welfare effects when the tick size is large. For instance, for the trading platform, the difference in annualized revenue between the optimal pricing policy and the uniform pricing policy is largest ($294 million per year) when the tick size is $1/8, even though it remains significant ($33 million per year) when the tick size is one penny. Intuitively, as the tick size declines, the ability of the platform to affect the trading rate is reduced since a variation in the make fee cannot affect the division of gains from trade by more than one tick per trade. Hence, the platform has a preference for a coarse price grid. In practice, however, the tick size is set by regulators rather than by trading platforms. For instance, in U.S. equity markets, the SEC imposed a one penny tick size for all stocks trading above one dollar in 2001.

VI. Conclusion

We have proposed an explanation for the maker/taker pricing model and show how this pricing scheme interacts with algorithmic trading. Our theory yields a rich set of empirical implications regarding the factors affecting make/take fees and the effects of algorithmic trading on liquidity, volume, and traders' welfare.

The model could be extended in many directions. In our model, traders do not choose the side on which they are active. An interesting extension would be to endogenize the number of makers and takers by allowing traders to choose their side. Furthermore, we focus on a single trading platform. In reality, securities often trade on multiple platforms. The economic forces analyzed in our paper should still hold in a multi-market environment as long as monitoring is costly. In particular, the make/take fees charged by a platform should still affect its trading rate as they will affect makers and takers' incentives to monitor the market. Yet, intermarket competition may add other considerations to the choice of make/take fees. Last, the joint CFTC-SEC task force on the flash crash of May 2010 has recently advocated varying make and take fees in real time in order to attract liquidity suppliers when the market momentarily lacks liquidity (see "Summary report of the joint CFTC-SEC Advisory Committee on Emerging Regulatory issues," p. 9[37]). Our model offers a starting point to analyze how make and take fees could be used to this end.

[37] Available at http://www.sec.gov/spotlight/sec-cftcjointcommittee/021811-report.pdf.

Initial submission: November 4, 2009; Final version received: August 15, 2012
Editor: Campbell Harvey

Appendix

Derivation of Traders' Payoffs

Let $\tilde{n}_T$ be the number of completed transactions (cycles) until time T. The expected payoff to market maker i until time T, net of monitoring costs, is

$$\Pi_i(T) = E_{\tilde{n}_T}\left(\sum_{k=1}^{\tilde{n}_T} \frac{\mu_i}{\bar{\mu}}\pi_m\right) - \frac{1}{2}\beta\mu_i^2 T. \tag{A1}$$

Thus, the steady-state expected profit of market maker i per unit of time is

$$\Pi_{im} \equiv \lim_{T\to\infty} \frac{\Pi_i(T)}{T} = \lim_{T\to\infty} \frac{E_{\tilde{n}_T}\left(\sum_{k=1}^{\tilde{n}_T} \frac{\mu_i}{\bar{\mu}}\pi_m\right)}{T} - \frac{1}{2}\beta\mu_i^2. \tag{A2}$$

A standard theorem from the theory of stochastic processes (the "Renewal Reward Theorem," see Ross (1996), p. 133) implies that

$$\lim_{T\to\infty} \frac{E_{\tilde{n}_T}(\sum_{k=1}^{\tilde{n}_T} \frac{\mu_i}{\bar{\mu}}\pi_m)}{T} = \frac{\frac{\mu_i}{\bar{\mu}}\cdot\pi_m}{\mathcal{D}(\bar{\mu},\bar{\tau})} = \frac{\mu_i}{\bar{\mu}}\cdot\pi_m\cdot\mathcal{R}(\bar{\mu},\bar{\tau}). \tag{A3}$$

We conclude from Equation (A2) that

$$\Pi_{im} = \frac{\mu_i}{\bar{\mu}}\cdot\pi_m\cdot\mathcal{R}(\bar{\mu},\bar{\tau}) - \frac{1}{2}\beta\mu_i^2, \tag{A4}$$

as claimed in the text. Expressions for the expected profit per unit of time for market takers and the trading platform are obtained in a similar way.

Proof of Proposition 1: Direct from the argument in the text. Q.E.D.

Proof of Proposition 2: Using equations (7) and (10), the first-order condition for market maker i is

$$\frac{\bar{\tau}(\bar{\tau}+\bar{\mu}-\mu_i)}{(\bar{\mu}+\bar{\tau})^2}\frac{\pi_m}{\beta} = \mu_i. \tag{A5}$$

Summing over all $i = 1,\ldots M$, we obtain

$$\frac{\bar{\tau}((\bar{\tau}+\bar{\mu})M-\bar{\mu})}{(\bar{\mu}+\bar{\tau})^2}\frac{\pi_m}{\beta} = \bar{\mu}. \tag{A6}$$

Similarly for market takers we obtain,

$$\frac{\bar{\mu}((\bar{\tau}+\bar{\mu})N-\bar{\tau})}{(\bar{\mu}+\bar{\tau})^2}\frac{\pi_t}{\gamma} = \bar{\tau}. \tag{A7}$$

Let $\mathcal{V} \equiv \frac{\bar{\mu}}{\bar{\tau}}$. Using (A6) and (A7), we obtain

$$\frac{M + (M-1)\,\mathcal{V}}{(1+\mathcal{V})^2}\frac{\pi_m}{\beta} = \bar{\mu}, \tag{A8}$$

$$\frac{\mathcal{V}\left((1+\mathcal{V})\,N - 1\right)}{(1+\mathcal{V})^2}\frac{\pi_t}{\gamma} = \bar{\tau}. \tag{A9}$$

Dividing these two equations gives

$$\frac{(M + (M-1)\,\mathcal{V})}{\mathcal{V}^2\left((1+\mathcal{V})\,N - 1\right)}z = 1, \tag{A10}$$

or equivalently,

$$\mathcal{V}^3 N + (N-1)\mathcal{V}^2 - (M-1)\,z\mathcal{V} - Mz = 0. \tag{A11}$$

This equation is equivalent to $\mathcal{V} = h(\mathcal{V}, M, N, z)$, where the function $h(\cdot)$ is defined by

$$h(\mathcal{V}, M, N, z) = \frac{(M-1)z}{\mathcal{V}N} + \frac{Mz}{N\mathcal{V}^2} - \frac{N-1}{N}. \tag{A12}$$

The function $h(\cdot, M, N, z)$ decreases in $\mathcal{V}$. It tends to plus infinity as $\mathcal{V}$ goes to zero, and to $-\frac{N-1}{N}$ as $\mathcal{V}$ goes to infinity. Thus, equation (A11) has a unique positive root that we denote by $\mathcal{V}^*$. We obtain $\bar{\mu}^*$ and $\bar{\tau}^*$ by inserting $\mathcal{V}^*$ into (A8) and (A9). Note that $\mu_1^* = \ldots = \mu_M^*$ and $\tau_1^* = \ldots = \tau_N^*$. Hence, $\mu_i^* = \bar{\mu}^* M$ and $\tau_j^* = \bar{\tau}^*$[38] N for all i, j. This completes the proof. Q.E.D.

Proof of Corollary 2: Recall that $\frac{\bar{\mu}^*}{\bar{\tau}^*} = \mathcal{V}^*$. Using equation (16), it is readily checked that $\mathcal{V}^* = 1$ if and only if $z = \frac{2N-1}{2M-1}$. Thus, $\bar{\mu}^* = \bar{\tau}^*$ if and only if $z = \frac{2N-1}{2M-1}$. Moreover, as shown in the proof of Corollary 1 (in the Internet Appendix), $\mathcal{V}^*$ increases in z. Hence, $\bar{\mu}^* > \bar{\tau}^*$ iff $z > \frac{2N-1}{2M-1}$. Q.E.D.

Proof of Corollary 3: Recall that $\mathcal{V}^* \equiv \frac{\mathcal{D}_t}{\mathcal{D}_m} = \frac{\bar{\mu}^*}{\bar{\tau}^*}$. We know from the proof of Corollary 1 (in the Internet Appendix) that $\mathcal{V}^*$ increases in z and M and decreases in N. The corollary is then immediate from equation (13). Q.E.D.

[38] Indeed, suppose for example that $\mu_1^* > \mu_2^*$. Then, from equation (A5),

$$\frac{\bar{\tau}^*\left(\bar{\tau}^* + \bar{\mu}^* - \mu_1^*\right)}{(\bar{\mu}^* + \bar{\tau}^*)^2}\frac{\pi_m}{\beta} > \frac{\bar{\tau}^*\left(\bar{\tau}^* + \bar{\mu}^* - \mu_2^*\right)}{(\bar{\mu}^* + \bar{\tau}^*)^2}\frac{\pi_m}{\beta},$$

which simplifies to $\mu_1^* < \mu_2^*$—a contradiction.

Proof of Corollary 4: Using Proposition 2, we deduce that

$$\begin{aligned}\mu_i^{\infty} &\equiv \lim_{M\to\infty}\mu_i^* = \lim_{M\to\infty}\left(\frac{M+(M-1)\mathcal{V}^*}{M(1+\mathcal{V}^*)^2}\right)\left(\frac{\pi_m}{\beta}\right)\\ &= \lim_{M\to\infty}\left(\frac{1+\frac{M-1}{M}\mathcal{V}^*}{(1+\mathcal{V}^*)^2}\right)\left(\frac{\pi_m}{\beta}\right)\\ &= \frac{1}{1+\mathcal{V}^{\infty}}\left(\frac{\pi_m}{\beta}\right) = \frac{1}{1+(zq)^{\frac{1}{2}}}\frac{\pi_m}{\beta} \quad \text{(using equation (17))}.\end{aligned}$$

A similar argument is used to derive τ_j^{∞}. Q.E.D.

Proof of Proposition 3: We fix $q>0$, and let $N=\frac{M}{q}$. Note that there is a one-to-one mapping between the fees charged by the platform and the trading profits π_m and π_t. Thus, instead of using c_m and c_t as the decision variables of the platform, we can use π_m and π_t. It turns out that this is easier. We also know that

$$\mathcal{R}(\bar{\mu}^*,\bar{\tau}^*)=\frac{\bar{\mu}^*\bar{\tau}^*}{\bar{\mu}^*+\bar{\tau}^*}=\frac{\bar{\mu}^*}{1+\mathcal{V}^*}. \tag{A13}$$

Moreover, as $\pi_m=\Gamma-\bar{c}-\pi_t$, we can express the trading rate, $\mathcal{R}(\bar{\mu}^*,\bar{\tau}^*)$, as a function of π_t only. Thus, for a fixed $\bar{c}$, we rewrite the platform's problem as

$$Max_{\pi_t}\quad \frac{\bar{\mu}^*}{1+\mathcal{V}^*}. \tag{A14}$$

The first-order condition with respect to π_t gives

$$\frac{\partial\bar{\mu}^*}{\partial\pi_t}=\mathcal{R}(\bar{\mu}^*,\bar{\tau}^*)\frac{\partial\mathcal{V}^*}{\partial\pi_t}. \tag{A15}$$

Since $\bar{\mu}^*=M\mu_1^*$, we can divide both sides by M and obtain

$$\frac{\partial\mu_1^*}{\partial\pi_t}=\frac{\mathcal{R}(\bar{\mu}^*,\bar{\tau}^*)}{M}\frac{\partial\mathcal{V}^*}{\partial\pi_t}. \tag{A16}$$

As the first order condition holds for any M, we can take limits on both sides to obtain

$$\lim_{M\to\infty}\frac{\partial\mu_1^*}{\partial\pi_t}=\lim_{M\to\infty}\frac{\mathcal{R}(\bar{\mu}^*,\bar{\tau}^*)}{M}\cdot\lim_{M\to\infty}\frac{\partial\mathcal{V}^*}{\partial\pi_t}. \tag{A17}$$

Straightforward calculations show that

$$\lim_{M\to\infty}\frac{\partial\mathcal{V}^*}{\partial\pi_t}=\frac{d\mathcal{V}^{\infty}}{d\pi_t}=-\frac{q}{2\mathcal{V}^{\infty}}\frac{\Gamma-c}{\pi_t^2}\frac{\gamma}{\beta},\ \text{and} \tag{A18}$$

$$\lim_{M\to\infty}\frac{\partial\mu_1^*}{\partial\pi_t}=-\frac{1}{\beta(1+\mathcal{V}^{\infty})}-\frac{1}{(1+\mathcal{V}^{\infty})^2}\cdot\frac{\partial\mathcal{V}^{\infty}}{\partial\pi_t}\cdot\frac{\Gamma-\bar{c}-\pi_t}{\beta}, \tag{A19}$$

where $\mathcal{V}^{\infty}$ is given by equation (17). Furthermore, it is direct from equation (A13) that

$$\lim_{M\to\infty} \frac{\mathcal{R}(\bar{\mu}^*, \bar{\tau}^*)}{M} = \frac{\mu_1^{\infty}}{1+\mathcal{V}^{\infty}}, \tag{A20}$$

where μ_1^{∞} is given by (18). Using (A18), (A19), and (A20), we obtain that equation (A17) is equivalent to

$$\frac{\pi_t}{\Gamma - \bar{c}} = \frac{\mathcal{V}^{\infty}}{1+\mathcal{V}^{\infty}}. \tag{A21}$$

Denote

$$w \equiv \frac{\pi_t}{\Gamma - \bar{c}}. \tag{A22}$$

Equation (A21) imposes

$$w = \frac{\mathcal{V}^{\infty}}{1+\mathcal{V}^{\infty}} = \frac{1}{1+(zq)^{\frac{1}{2}}}. \tag{A23}$$

Now, observe that $z = r(\frac{1-w}{w})$. Thus, we can rewrite equation (A23) as

$$w = \frac{1}{1+\left(e\frac{1-w}{w}\right)^{-0.5}(rq)^{-0.5}}. \tag{A24}$$

It is immediate that the unique solution of equation (A24) is

$$w^* = \frac{(rq)^{\frac{1}{3}}}{1+(rq)^{\frac{1}{3}}}. \tag{A25}$$

Accordingly, from equation (A22) we obtain

$$\pi_t = \frac{\Gamma - \bar{c}}{1+(rq)^{-\frac{1}{3}}}, \tag{A26}$$

which implies

$$\pi_m = \Gamma - \bar{c} - \pi_t = \frac{\Gamma - \bar{c}}{1+(rq)^{\frac{1}{3}}}. \tag{A27}$$

As, by definition, $\pi_m = a - v_0 - c_m$, we deduce from equation (A27) that, when $a - v_0 = s \cdot \Delta$, the optimal make fee in the thick market case is

$$c_{ms}^* = s\cdot\Delta - \pi_m = s\cdot\Delta - \frac{\Gamma - \bar{c}}{1+(qr)^{\frac{1}{3}}} \tag{A28}$$

and the optimal take fee follows from the fact that $c_{ms}^* + c_{ts}^* = \bar{c}$. Q.E.D.

Proof of Corollary 5: The result follows directly from equation (22). Q.E.D.

Proof of Proposition 4: Using equation (A8) in the proof of Proposition 2, we deduce that, for a fixed $\bar{\tau}$, market makers' aggregate monitoring level, $\bar{\mu}$, solves

$$F(\bar{\mu}; \bar{\tau}, \beta, c_m, M) = 0, \tag{A29}$$

where

$$F(\bar{\mu}; \bar{\tau}, \beta, c_m, M) \equiv M + \frac{(M-1)\pi_m \bar{\mu}}{\bar{\tau}} - \beta \bar{\mu} \left(1 + \frac{\bar{\mu}}{\bar{\tau}}\right)^2. \tag{A30}$$

It is easily shown that, for all parameter values, equation (A29) has a unique positive solution $\bar{\mu}$. Let $\bar{\mu} = \varphi(\bar{\tau}; \beta, c_m, M)$ be this solution. Using the implicit function theorem,

$$\frac{d\varphi}{d\bar{\tau}} = -\frac{\frac{\partial F}{\partial \bar{\tau}} \mid_{\bar{\mu}=\varphi(\bar{\tau};\beta,c_m,M)}}{\frac{\partial F}{\partial \bar{\mu}} \mid_{\bar{\mu}=\varphi(\bar{\tau};\beta,c_m,M)}}. \tag{A31}$$

Using the expression for $F(\cdot)$, we obtain $\frac{\partial F}{\partial \bar{\tau}} \mid_{\bar{\mu}=\varphi(\bar{\tau};\beta,c_m,M)} > 0$ and $\frac{\partial F}{\partial \bar{\mu}} \mid_{\bar{\mu}=\varphi(\bar{\tau};\beta,c_m,M)} < 0$. Hence, $\frac{d\varphi}{d\bar{\tau}} > 0$. Now, since we have $\bar{\mu} = \varphi(\bar{\tau}; \beta, c_m, M)$, we deduce that

$$\mathcal{D}_m = f(\mathcal{D}_t; \beta, M, c_m), \tag{A32}$$

with $f(\mathcal{D}_t; \beta, M, c_m) = \frac{1}{\varphi(\frac{1}{\mathcal{D}_t}; \beta, c_m, M)}$. Then, since $\frac{d\varphi}{d\bar{\tau}} > 0$ we have that $\frac{\partial f}{\partial \mathcal{D}_t} > 0$. In a similar way, we can show that there exists a function $g(\cdot)$ such that $\mathcal{D}_m = g(\mathcal{D}_t; \gamma, N, c_t)$ and $\frac{\partial g}{\partial \mathcal{D}_m} > 0$. This completes the first part of the proposition. The second part is obtained using similar arguments (again applying the implicit function theorem). We omit the details for brevity. Q.E.D.

Proof of Proposition 6: See the discussion before the proposition.

REFERENCES

Admati, Anat R., and Paul Pfleiderer, 1988, A theory of intraday patterns: Volume and price variability, *Review of Financial Studies* 1, 3–40.

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

Barclay, Michael J., and Terrence Hendershott, 2004, Liquidity externalities and adverse selection: Evidence from trading after hours, *Journal of Finance* 59, 681–710.

Biais, Bruno, Lawrence R. Glosten, and Chester Spatt, 2005, Market microstructure: A survey of microfoundations, empirical results and policy implications, *Journal of Financial Markets* 8, 111–264.

Biais, Bruno, Pierre Hillion, and Chester S. Spatt, 1995, An empirical analysis of the limit order book and the order flow in the Paris Bourse, *Journal of Finance* 50, 1655–1689.

Biais, Bruno, Johan Hombert, and Pierre O. Weill, 2010, Trading and liquidity with limited cognition, Working paper, Toulouse University, IDEI.

Brogaard, Jonathan A., 2010, High frequency trading and its impact on market quality, Working paper, Kellogg School of Management, Northwestern University.

Chaboud, Alain, Benjamin Chiquoine, Erik Hjalmarsson, and Clara Vega, 2010, Rise of the machines: Algorithmic trading in the foreign exchange market, International Finance Discussion Papers, Board of Governors of the Federal Reserve System.

Chordia, Tarun, Richard Roll, and Avanidhar Subrahmanyam, 2011, Recent trends in trading activity, *Journal of Financial Economics* 101, 243–263.
Colliard, Jean-Edouard, and Thierry Foucault, 2012, Trading fees and efficiency in limit order markets, *Review of Financial Studies* 25, 3305–3350.
Coopejans, Mark T., Ian H. Domowitz, and Ananth Madhavan, 2001, Liquidity in an automated auction, Working paper, BlackRock; Barclays Global Investors.
Degryse, Hans, Frank De Jong, Maarten Van Ravenswaaij, and Gunther Wuyts, 2005, Aggressive orders and the resiliency of a limit order market, *Review of Finance* 9, 201–242.
Dow, James, 2004, Is liquidity self-fulfilling? *Journal of Business* 77, 895–908.
Duffie, Darrell, 2010, Presidential address: Asset price dynamics with slow moving capital, *Journal of Finance* 65, 1237–1267.
Duffie, Darrell, Nicolae Gârleanu, and Lasse H. Pedersen, 2005, Over-the-counter markets, *Econometrica* 73, 1815–1847.
Easley, David, and Maureen O'Hara, 1992, Time and the process of security price adjustment, *Journal of Finance* 47, 577–605.
Engle, Robert F., 2000, The econometrics of ultra high-frequency data, *Econometrica* 68, 1–22.
Engle, Robert F., and Jeffrey R. Russell, 1998, Autoregressive conditional duration: A new model for irregularly spaced transaction data, *Econometrica* 66, 1127–1162.
Foucault, Thierry, Ohad Kadan, and Eugene Kandel, 2005, Limit order book as a market for liquidity, *Review of Financial Studies* 18, 1171–1217.
Foucault, Thierry, and Albert J. Menkveld, 2008, Competition for order flow and smart order routing systems, *Journal of Finance* 63, 119–158.
Foucault, Thierry, Ailsa Roëll, and Patrick Sandås, 2003, Market making with costly monitoring: An analysis of SOES trading, *Review of Financial Studies* 16, 345–384.
Glosten, Lawrence R., 1994, Is the electronic open limit order book inevitable? *Journal of Finance* 49, 1127–1161.
Goettler, Ronald L., Christine A. Parlour, and Uday Rajan, 2005, Equilibrium in a dynamic limit order market, *Journal of Finance* 60, 2149–2192.
Hasbrouck, Joel, and Gideon Saar, 2010, Low latency trading, Working paper, *Johnson School*, Cornell University.
Hendershott, Terrence J., Charles M. Jones, and Albert J. Menkveld, 2011, Does algorithmic trading improve liquidity? *Journal of Finance* 66, 1–33.
Hendershott, Terrence J., and Haim Mendelson, 2000, Crossing networks and dealer markets: Competition and performance, *Journal of Finance* 55, 2071–2115.
Hendershott, Terrence J., and Pamela C. Moulton, 2011, Speed and stock market quality: The NYSE's hybrid, *Journal of Financial Markets* 14, 568–604.
Hendershott, Terrence J., and Ryan Riordan, 2009, Algorithmic trading and information, Working paper, University of California, Berkeley.
Hollifield, Burton, Robert A. Miller, Patrik Sandås, and Joshua Slive, 2006, Estimating the gains from trade in limit order markets, *Journal of Finance* 61, 2753–2804.
Kandel, Eugene, and Leslie M. Marx, 1999, Payments for order flow on NASDAQ, *Journal of Finance* 49, 35–66.
Large, Jeremy, 2007, Measuring the resiliency of an electronic limit order book, *Journal of Financial Markets* 10, 1–25.
Liu, Wai-Man, 2009, Monitoring and limit order submission risks, *Journal of Financial Markets* 12, 107–141.
Menkveld, Albert J., 2010, High frequency trading and the new market makers, Working paper, Vrije Universiteit.
Pacurar, Maria, 2006, Autoregressive Conditional Duration (ACD) models in finance: A survey of the theoretical and empirical literature, Working paper, Dalhousie University.
Pagano, Marco, 1989, Trading volume and asset liquidity, *Quarterly Journal of Economics* 104, 255–276.
Parlour, Christine A., and Uday Rajan, 2003, Payment for order flow, *Journal of Financial Economics* 68, 379–411.

Rochet, Jean Charles, and Jean Tirole, 2006, Two sided markets: A progress report, *Rand Journal of Economics* 37, 645–667.

Ross, Sheldon M., 1996, *Stochastic Processes* (John Wiley & Sons, Inc. New York).

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

Roşu, Ioanid, 2010, Liquidity and information in order driven markets, Working paper, Booth School of Business, University of Chicago.

Sandås, Patrik, 2001, Adverse selection and competitive market making: Empirical evidence from a limit order market, *Review of Financial Studies* 14, 705–734.