Liquidity Cycles and Make/Take Fees in Electronic Markets[*]

Thierry Foucault
HEC School of Management, Paris
1 rue de la Liberation
78351 Jouy en Josas, France
foucault@hec.fr

Ohad Kadan
Olin Business School
Washington University in St. Louis
Campus Box 1133, 1 Brookings Dr.
St. Louis, MO 63130
kadan@wustl.edu

Eugene Kandel
School of Business Administration,
and Department of Economics,
Hebrew University,
91905, Jerusalem, Israel
mskandel@mscc.huji.ac.il

September 2009

Abstract

We develop a model of trading in securities markets with two specialized sides: traders posting quotes ("market makers") and traders hitting quotes ("market takers"). Liquidity cycles emerge naturally, as the market moves from phases with high liquidity to phases with low liquidity. Traders monitor the market to seize profit opportunities. Complementarities in monitoring decisions generate multiplicity of equilibria: one with high liquidity and another with no liquidity. The trading rate depends on the allocation of the trading fee between each side and the maximal trading rate is typically achieved with asymmetric fees. The difference in the fee charged on market-makers and the fee charged on market-takers ("the make-take spread") increases in (i) the tick-size, (ii) the ratio of the size of the market-making side to the size of the market-taking side, and (iii) the ratio of monitoring costs for market-takers to monitoring costs for market-makers. The model yields several empirical implications regarding the trading rate, the duration between quotes and trades, the bid-ask spread, and the effect of algorithmic trading on these variables.

Keywords: Liquidity, Monitoring, Make/Take Fees, Duration Clustering, Algorithmic trading, Two-Sided Markets.

[*] We thank Bruno Biais, Gideon Saar, Terrence Hendershott, Ernst Maug, Albert Menkveld, Marios Panayides, Jean-Charles Rochet, Elu Von Thadden, and seminar participants at the 1st FBF-IDEI conference on investment banking and financial markets in Toulouse, Humboldt University, University of Mannheim, and University College in Dublin, as well as participants at the 2009 WFA and 2009 EFA meetings for their useful comments.

1 Introduction

Securities trading, especially in equities markets, increasingly takes place in electronic limit order markets. The trading process in these markets feature high frequency cycles made of two phases: (i) a "make liquidity" phase during which traders post prices (limit orders) at which they are willing to trade, and (ii) a "take liquidity" phase during which limit orders are hit by market orders, generating a transaction. The submission of market orders depletes the limit order book of liquidity and ignites a new make/take cycle as it creates transient opportunities for traders submitting limit orders.[1] The speed at which these cycles are completed determines the trading rate, a dimension of market liquidity.

A trader reacts to a transient increase or decline in the liquidity of the limit order book only when she becomes aware of this trading opportunity. Accordingly, the dynamics of trades and quotes in limit order markets is in part determined by traders' monitoring decisions, as emphasized by some empirical studies (e.g., Biais et al. (1995), Sandås (2001) or Hollifield et al. (2004)). For instance, Biais, Hillion, and Spatt (1995) observe that (p.1688): "*Our results are consistent with the presence of limit order traders monitoring the order book, competing to provide liquidity when it is rewarded, and quickly seizing favorable trading opportunities.*" Hence, traders' attention to the trading process is an important determinant of the trading rate.

In practice, monitoring is costly because intermediaries (brokers, market-makers, as well as potentially patient traders who need to execute a large order) have limited monitoring capacity or choose to allocate limited attention to certain markets.[2] Hence, traders react with delay to trading opportunities and the trading rate depends on a trade-off between the benefit and cost of monitoring. Our goal in this paper is to study this trade-off and its impact on the trading rate. In the process, we address two sets of related issues.

Firstly, algorithmic trading (the automation of monitoring and orders submission) considerably decreases the cost of monitoring and revolutionizes the way liquidity is provided and consumed. We use our model to study the effects of this evolution on the trading rate, the bid-ask spread, and welfare. Secondly, the model sheds light

[1] These cycles are studied empirically in Biais, et al. (1995), Coopejans et al.(2003), Degryse et al.(2005), and Large (2007).

[2] For instance, Corwin and Coughenour (2008) show that limited attention by market-makers ("specialists") on the floor of the NYSE affects their liquidity provision.

	Tape A - NYSE Stocks		Tape B - Other Stocks		Tape C - NASDAQ Stocks	
	Make Fee	Take Fee	Make Fee	Take Fee	Make Fee	Take Fee
AMEX	-30	30	-30	30	-30	30
BATS	-24	25	-30	25	-24	25
LavaFlow	-24	26	-24	26	-24	26
NASDAQ-OMX	-20	30	-20	30	-20	30
NYSEArca	-25	30	-20	30	-20	26

Table 1: Fees per share (in cents for 100 shares) for limit orders (Make Fee) and market orders (Take Fee) on different trading platforms in the US. A minus sign indicates a rebate. Source: Traders Magazine, July 2008

on pricing schedules set by trading platforms. Increasingly, these platforms charge different fees on limit orders (orders "making liquidity") and market orders (orders "taking liquidity"). The difference between these fees is called the *make/take spread* and is usually negative. That is, traders posting quotes pay a lower fee than traders hitting these quotes.

For instance, Table 1 gives the make/take fees charged on liquidity makers and liquidity takers for several U.S. equity trading platforms, as of July 2008. At this time, all these platforms subsidize liquidity makers by paying a rebate on executed limit orders, and charge a fee on liquidity takers (so called "access fees").

This fee structure results in significant monetary transfers between traders taking liquidity, traders making liquidity, and the trading platforms.[3] For this reason, the make/take spread is closely followed by market participants, in particular by market-making firms using highly automated strategies.[4] Access fees are the subject of heated debates and, in its regulation NMS, the SEC decided to cap them at $0.003 per share (30% of the tick size) in equity markets.[5] Yet, to the best of our knowledge, the rationale for the make/take spread and its impact on liquidity have not been

[3] For instance, in each transaction, BATS charges a fee of 0.25 cents per share on market orders and rebates 0.24 cents on executed limit orders (see Table 1). On October 10, 2008, 838,488,549 shares of stocks listed on the NYSE were traded on BATS (about 9% of the trading volume in these stocks on this day); see BATS website: http://www.batstrading.com/. Thus, collectively on this day, limit order traders involved in these transactions collected about $2.01 million in rebates from BATS while traders submitting market orders paid about $2.09 million in fees to BATS.

[4] Some specialized magazines report the fees charged by the various electronic trading platforms in U.S. equity markets. See for instance the "Price of Liquidity" section published by "Traders magazine"; http://www.tradersmagazine.com.

[5] As an example of the controversies raised by these fees, see the petition for rule-making regarding access fees in option markets, addressed by Citadel at the SEC at http://www.sec.gov/rules/petitions/2008/petn4-562.pdf

analyzed. Our analysis provides an explanation for the make-take spread, makes predictions about its determinants, and shows that it serves to maximize the trading rate.

In our model we consider a trading platform on which two types of traders interact: (i) those who post quotes (the "market-makers") and (ii) those who hit these quotes (the "market-takers"). All market participants monitor the market to grab fleeting trading opportunities. Specifically, a market-maker wants to be first to post new quotes after a transient increase in the bid-ask spread and a market-taker wants to be first to hit quotes when the bid-ask spread is tight. An increase in traders' monitoring intensities shortens their reaction time to changes in the state of the market and thereby increases the trading rate. In choosing their monitoring intensity, traders on each side trade-off the benefit from a higher likelihood of being first to detect a profit opportunity with the opportunity cost of monitoring.

Monitoring decisions of traders on both sides reinforce each other. Indeed, suppose that an exogenous shock induces market-takers to monitor the market more intensively. Then, market-makers expect more frequent profit opportunities since good prices are hit more quickly. Hence, they have an incentive to monitor more and as a consequence the market features good prices more frequently, which in turn induces market-takers to monitor more. This cross-side complementarity in monitoring decisions creates a coordination problem, which results in two equilibria (i) an equilibrium with no monitoring and no trading; and (ii) an equilibrium with monitoring and trading.[6]

In the equilibrium with trading, the aggregate monitoring levels of each side are typically not equal. For instance, suppose that market-takers' monitoring cost is relatively small and suppose that gains from trade when a transaction occurs are equally split between market-makers and market-takers. In this case market-takers monitor the market more than market-makers, in equilibrium, since they have relatively small monitoring costs. Thus, good prices take relatively more time to be posted than it takes time for market-takers to hit these prices when they are posted. In this sense, there is an excess of liquidity demand relative to liquidity supply in

[6] It is well-known that liquidity externalities create coordination problems among traders, which lead to multiple equilibria with differing levels of liquidity (see Admati and Pfleiderer (1988), Pagano (1989), and Dow (2005) for example). In contrast to the extant literature, our model emphasizes the egg and chicken problem that exists between traders posting quotes on the one hand and traders hitting quotes on the other hand.

the market. In this situation, the relatively slow response of market-makers to a transient increase in the bid-ask spread slows down trading since trades happen when the bid-ask spread is tight. To achieve a higher trading rate, the trading platform can reduce its fee on market-makers while increasing its fee on market-takers so that its total profit per trade is unchanged. In this way, market-makers obtain a larger fraction of the gains from trade when a transaction occurs and have more incentives to quickly improve upon unaggressive quotes. Thus, good prices, hence trades, are more frequent.

Generally, the same logic implies that there is a level of the make-take spread that maximizes the trading rate. We show that the optimal make-take spread increases in (i) the tick size, (ii) the ratio of the number of market-makers to the number of market-takers, and (iii) the ratio of market-takers' monitoring cost to market-makers' monitoring cost. Indeed, in equilibrium, an increase in these parameters enlarges the speed at which good prices are posted relative to the speed at which they are hit. Thus, the imbalance between the supply and demand of liquidity narrows and therefore the need to incentivize market-makers is lower.

The model has a rich set of empirical implications. For instance, complementarities between market-makers and market-takers provide a new explanation for clustering in trade duration found in securities markets (see for instance Engle and Russell (1998)). Indeed, it implies that the aggregate monitoring intensity of both sides are positively related. Thus, an increase in the speed at which market-makers post good prices results in an increase in the speed at which market-takers hit these quotes and vice versa. This inter-dependence leads to periods in which trading activity is high because both sides are fast or periods in which trading activity is low because both sides are slow. The coexistence of an equilibrium with trading and an equilibrium without trading is an extreme manifestation of this phenomenon in our model.

Moreover, the model implies that the make-take spread increases in the tick size. Indeed, the higher the tick size, the higher the fraction of gains from trade for market-makers. Thus, market-makers have naturally more incentive to monitor markets with a large tick size. Hence, rebates for market-makers are more likely to appear in stocks or platforms with a low tick size. In line with this prediction, the practice of subsidizing market-makers in U.S equity markets and more recently in U.S. options markets developed after the tick size was reduced to a penny in these markets.[7]

Our model also has implications for the introduction and proliferation of algorithmic trading. Algorithmic trading reduces the monitoring costs for both market-makers and market-takers through the use of computers. Observe, however, that the same economic forces apply. Indeed, computerized monitoring is still costly since fixed computing capacities must be allocated among hundreds of stocks and millions of quotes and trades that require processing. Intense monitoring in one market may result in lost profit opportunities in another market. Our model predicts that the development of algorithmic trading should have a large positive impact on the trading rate (as found empirically in Hendershott, Jones, and Menkveld (2009)). The reduction in monitoring costs has direct positive impact on the level of monitoring by both sides. But this positive impact encourages market participants to monitor even more because of the complementarity in monitoring decisions. Eventually, through this chain reaction, the impact of the reduction in monitoring costs on the trading rate is amplified.

In contrast we find that the effect of algorithmic trading on the average bid-ask spread is ambiguous. Indeed, a decrease in market-makers' monitoring cost reduces the average bid-ask spread while a decrease in market-takers' monitoring cost has the opposite effect. Actually in the second case, the speed at which market-takers hit the quotes increase at a faster rate than the speed at which market-makers post good prices. These findings are consistent with

The increase in the bid-ask spread however has no material effect on market-takers' welfare as they only trade when the bid-ask spread is tight in our model. Instead, we show that market participants' welfare is inversely related to monitoring costs. Indeed, a decrease in monitoring costs results in a larger trading rate, which, in our setting, makes traders better off since positive gains from trade are realized each time there is a transaction. Thus, the model identifies one channel through which algorithmic trading could be welfare enhancing.

Our study is related to several strands of research. Foucault, Roëll and Sandås (2003) and Liu (2008) provide theoretical and empirical analyzes of market-making with costly monitoring. However, the effects in these models are driven by market-makers' exposure to adverse selection and they do not study the role of trading fees. It is also related to the burgeoning literature on two-sided markets (see Rochet and

[7] See "*Options maker-taker markets gain steam*," Traders Magazine, October 2007.

Tirole (2006) for a survey). Rochet and Tirole (2006) define a two-sided market as a market in which the volume of transactions depends on the allocation of the fee earned by the matchmaker (the trading platform in our model) between the end-users (the market-makers and the market-takers in our model).[8] Make-take fees strongly suggest that securities markets are two-sided. To our knowledge, our paper is first to study the cause and implications of the two-sided nature of securities markets. Our paper also contributes to the growing literature on the effects of algorithmic trading (see for instance (e.g., Biais and Weill (2008), Foucault and Menkveld (2008) or Hendershott, Jones, and Menkveld (2009)). Finally, our paper adds to the developing literature on limited attention and its rationale (e.g. Abel, Eberly, and Panageas (2009), Huang and Liu (2007), Iliev and Welch (2008), and Sims (2003)). We show how a fee structure can affect the optimal attention level of market participants, and derive welfare and liquidity implications.

Section 2 describes the model. In Section 3, we study the determinants of traders' equilibrium monitoring intensities for fixed fees of the trading platform. We endogenize these fees and derive the optimal fee structure for the trading platform in Section 4. We discuss the empirical implications of the model in Section 5 and Section 6 concludes. The proofs are in the Appendix.

2 Model

2.1 Market participants

We consider a market for a security with two sides: "market-makers" and "market-takers." Market-makers post quotes (limit orders) whereas market-takers hit these quotes (submit market orders) to complete a transaction.[9] The number of market-makers and market-takers is, respectively, M and N. All participants are risk neutral.

To simplify the analysis we assume that traders on one side cannot switch to the other side. This is the case in some markets (e.g., EuroMTS, a trading platform for government bonds in Europe) but, in reality, traders can often choose whether to post a quote or to hit a quote. However, even in this case, traders tend to specialize as assumed here. The market-making side can be viewed as electronic market-makers, such

[8] Examples of two-sided markets include videogames platforms, payment card systems etc...

[9] Some trading platforms refer to the market-making side as the "passive" side and to the market-taking side as the "active" (or aggressive) side. See for instance Chi-X at http://www.chi-x.com/Cheaper.html

as Automated Trading Desk (ATD), Global Electronic Trading company (GETCO), Tradebots Systems, or Citadel Derivatives, which specialize in high frequency market-making.[10] The market-taking side are institutional investors who break their large orders and feed them piecemeal when liquidity is plentiful to minimize their trading costs.[11] Electronic market-makers primarily use limit orders whereas the second type of traders primarily use market orders. Both types increasingly use highly automated algorithms to detect and exploit trading opportunities.

The expected payoff of the security is v_0. Market-takers value the security at $v_0 + L$, where $L > 0$ while market-makers value the security at v_0. Heterogeneity in traders' valuation creates gains from trade as in other models of trading in securities markets (e.g., Duffie et al. (2005) or Hollifield et al. (2006)).[12] As market-takers have a higher valuation than market-makers, they buy the security from market-makers. Thus, our model captures "the upper half" of the market characterized by limit sell orders and market buy orders. In a more complex model, market-takers could have either high or low valuations relative to market-makers, so that they can be buyers or sellers. This possibility adds some mathematical complexity to the model, but provides no additional economic insight.

Market-makers and market-takers meet on a trading platform with a positive tick-size denoted by $\Delta > 0$ and the first price on the grid above v_0 is half a tick above v_0. Let $a \equiv v_0 + \frac{\Delta}{2}$ be this price. All trades take place at this price because market-takers' valuation is less than $a + \Delta$ (specifically, $\frac{\Delta}{2} < L \leq \Delta$) and market-makers lose money if they trade at a smaller price than a on the grid. Thus, we focus on a "one tick market" similar, for example, to Parlour (1998) or Large (2008). Finally, we assume that a large number of shares is offered for sale at price $a + \Delta$ by a fringe of competitive traders, as in Seppi (1997) or Parlour (1998). The cost of liquidity provision for these traders is higher than for the electronic market-makers

[10] According to analysts, electronic market-makers now account for a very high fraction of the total liquidity provision on electronic markets. For instance, Schack and Gawronski (2008) write on page 74 that: "*based on our knowledge of how they do business [...], we believe that they [electronic market-makers] may be generating two-thirds or more of total daily volume today, dwarfing the activity of institutional investors.*"

[11] Bertsimas and Lo (1998) solve the dynamic optimization of such traders, assuming that they exclusively use market orders as we do here.

[12] Hollifield et al. (2004) and Hollifield et al. (2006) show empirically that heterogeneity in traders' private values can explain the flow of orders in limit order markets. In reality, as noted in Duffie et al.(2005), differences in traders' private values may stem from differences in hedging needs (endowments), liquidity needs or tax treatments.

and therefore they cannot intervene profitably at price a.

There is an upper bound (normalized to one) on the number of shares that can be offered at price a. This upper bound rules out the uninteresting case in which a single market-maker or multiple market-makers offer an infinite quantity at price a. In a more complex model, the upper bound could derive from an upward marginal cost of liquidity provision due, for instance, to exposure to informed trading as in Glosten (1994) or Sandås (2001).[13]

The trading platform charges trading fees each time a trade occurs. The fee (per share) paid by a market-maker is denoted c_m, whereas the fee paid by a market-taker is denoted c_t. These fees can be either positive or negative (rebates). We normalize the cost of processing trades for the trading platform to zero so that, per transaction, the platform earns a profit of

$$\bar{c} \equiv c_m + c_t.$$

Introducing an order processing cost per trade is straightforward and does not change the results.

Thus, the gains from trade in each transaction (i.e., L) are split between the parties to the transaction and the trading platform as follows: the market-taker obtains

$$\pi_t = L - \frac{\Delta}{2} - c_t, \tag{1}$$

the market-maker obtains

$$\pi_m = \frac{\Delta}{2} - c_m, \tag{2}$$

and the platform obtains $\bar{c}$.

Thus, the gains from trade accruing to market-makers and market-takers are $L-\bar{c}$. We focus on the case $\bar{c} \leq L$ since otherwise traders on at least one side lose money on each trade, and would therefore choose not to trade at all. Moreover, we assume that $c_m > -\frac{\Delta}{2}$, so that $a - \Delta - c_m - v_0 < 0$. Thus, a market-maker cannot profitably post an offer at a price strictly less than a, even if she receives a subsidy from the platform ($c_m < 0$). As shown below this constraint is not binding for the platform (see Section 4).

Notice that since quotes must be on the price grid set by the platform, market-makers cannot fully neutralize a small change in the fee structure by adjusting their

[13] Empirically, several papers document a reduction in quoted depth after a reduction in tick size (e.g., Goldstein and Kavajecz (2000)). This observation is consistent with an upward marginal cost of liquidity provision.

quotes. For instance, suppose that the fee charged on market-makers increases by a small amount, say 1% of the tick size. Market-makers cannot neutralize this increase by posting a higher offer at $a + 1\% \cdot \Delta$, as this price is not on the grid.

Our setup is clearly very stylized. Yet, it captures in the simplest possible way the essence of the liquidity cycles described in the introduction. Specifically, when there is no quote at a, the market lacks liquidity and there is a profit opportunity for market-makers. Indeed, the first market-maker who submits an offer at a will serve the next buy market order and earns π_m. Conversely, when there is an offer at a, liquidity is plentiful and there is a profit opportunity (worth π_t) for a market-taker. After a trade, the market switches back to a state in which liquidity is scarce. Consequently, the market oscillates between a state in which there is a profit opportunity for market-makers and a state in which there is a profit opportunity for market-takers. Thus, market-makers and market-takers have an incentive to monitor the market. Market-makers are looking for periods when liquidity is scarce and market-takers are looking for periods when liquidity is plentiful.

2.2 Cycles, Monitoring, and Timing

We now define the notion of "cycles," discuss the monitoring activities of market participants, and explain the timing of the game.

Cycles. This is an infinite horizon model with a continuous time line. At each point in time the market can be in one of two states:

1. State E – liquidity is low: no offer is posted at a.

2. State F – Liquidity is high: an offer for one share is posted at a.

Thus F (for Full) is the state in which the (half) bid-ask spread (i.e., $a - v_0$) is competitive whereas E (for Empty) is the state in which the bid-ask spread is not competitive. The market moves from state F to state E when a market-taker hits the best offer. The bid-ask spread then widens until a market-maker sets the competitive offer. At this point the bid-ask spread reverts to the competitive level, i.e., the market moves from state E to state F. Then, 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.

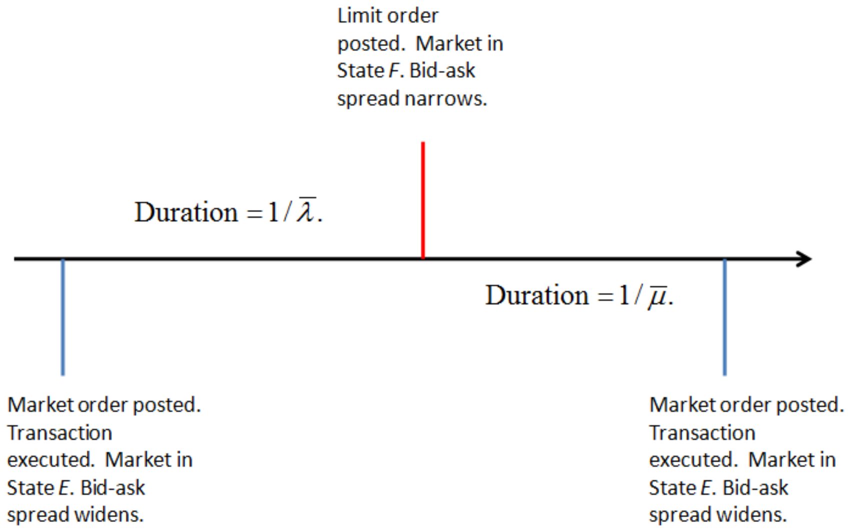
Limit order posted. Market in State *F*. Bid-ask spread narrows.
Duration $=1/\overline{\lambda}$.
Duration $=1/\overline{\mu}$.
Market order posted. Transaction executed. Market in State *E*. Bid-ask spread widens.
Market order posted. Transaction executed. Market in State *E*. Bid-ask spread widens.

Figure 1: Time Structure of a Cycle

Monitoring. Market-makers and market-takers have an incentive to monitor the market to be the first to detect a profit opportunity for their side. We formalize monitoring as follows. Each market-maker $i = 1, ..., M$ inspects the market according to a Poisson process with parameter λ_i, that characterizes her monitoring intensity. As a result, the time between one inspection of the market to the next by market-maker i is distributed exponentially with an average inter-inspection time of $\frac{1}{\lambda_i}$. Similarly, each market-taker $j = 1, ..., N$ inspects the market according to a Poisson process with parameter μ_j.[14] The total inspection frequency of all market-makers is

$$\bar{\lambda} \equiv \lambda_1 + ... + \lambda_M,$$

and the total inspection frequency of market-takers is

$$\bar{\mu} \equiv \mu_1 + ... + \mu_N.$$

When a market-maker inspects the market she learns whether it is in state E or F. If the bid-ask spread is not competitive (state E) then she posts an offer at a. If it is competitive (state F), the market-maker stays put until her next inspection.

[14] This approach rules out deterministic monitoring such as inspecting the market exactly once every certain number of minutes. 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.

Similarly, a market-taker can respond to the state of the market only upon inspection. He submits a market order when he observes that the bid-ask spread is competitive, and stays put until the next inspection otherwise.[15] Figure 1 illustrates this process.

The expected duration of a cycle depends on aggregate monitoring levels. To see this, suppose that a trade just took place so that the bid-ask spread just widened. Then, the average time it takes for the market-making side to post a new offer at a is $\mathcal{D}_m \equiv \frac{1}{\lambda_1+\ldots+\lambda_M} = \frac{1}{\bar{\lambda}}$. Once a market-maker posts an offer at a, so that the market enters in state F, it takes then on average $\mathcal{D}_t \equiv \frac{1}{\mu_1+\ldots+\mu_N} = \frac{1}{\bar{\mu}}$ units of time for a market-taker to hit this offer. Thus, the average duration of a cycle is

$$\mathcal{D}\left(\bar{\lambda}, \bar{\mu}\right) \equiv \mathcal{D}_m + \mathcal{D}_t = \frac{1}{\bar{\lambda}} + \frac{1}{\bar{\mu}} = \frac{\bar{\lambda}+\bar{\mu}}{\bar{\lambda}\cdot\bar{\mu}}. \tag{3}$$

Similarly, the trading rate, defined as the average number of transactions per unit of time, is given by

$$\mathcal{R}\left(\bar{\lambda}, \bar{\mu}\right) \equiv \frac{1}{\mathcal{D}\left(\bar{\lambda}, \bar{\mu}\right)} = \frac{\bar{\lambda}\cdot\bar{\mu}}{\bar{\lambda}+\bar{\mu}}. \tag{4}$$

The aggregate monitoring levels, $\bar{\lambda}$ and $\bar{\mu}$, determine market liquidity. Indeed, $\bar{\mu}$ determines the speed at which the market-taking side responds to a competitive offer made by the market-making side whereas $\bar{\lambda}$ determines the speed at which the market-making side reinjects liquidity into the market after a transaction. This speed determines the *resiliency* of the market.[16] Thus, $\bar{\lambda}$ and $\bar{\mu}$ are measures of liquidity supply and demand respectively. Ultimately, the trading rate increases when either $\bar{\lambda}$ or $\bar{\mu}$ increase. As $\mathcal{D}_m = \frac{1}{\bar{\lambda}}$ and $\mathcal{D}_t = \frac{1}{\bar{\mu}}$, the inter-trade average durations ($\mathcal{D}_m$ and $\mathcal{D}_t$) can be used as proxies for the aggregate monitoring level. Alternatively, $\bar{\lambda}$ or $\bar{\mu}$ could be estimated directly using the empirical technique described in Large (2007). We use these observations to develop several empirical implications of the model in Section 5.

In practice, monitoring can be manual, by looking at a computer screen, or automated by using automated algorithms. For humans, the need to monitor several stocks contemporaneously limits the monitoring capacity and constrains the amount of attention dedicated to a specific stock. Computers also have fixed capacity that

[15] Hall and Hautsch (2007) model the arrival of buy and sell market orders as a Poisson Process with state-dependent intensities. They find empirically that these intensities are higher when the bid-ask spread is tight. This empirical finding is consistent with our assumption that market takers submit their market orders when the bid-ask spread is competitive.

[16] See, for instance, Foucault et al.(2005) for a theoretical analysis of resiliency and Large (2007) for an empirical analysis.

must be allocated over potentially hundreds of stocks and millions of pieces of information that require processing. Prioritization of this process is conceptually similar to the allocation of attention across different stocks by a human market-marker. Hence, in all cases, monitoring one market is costly, because it reduces the monitoring capacity available for other markets.

To account for this cost, we assume that, over a time interval of length T, a market-maker choosing a monitoring intensity λ_i bears a monitoring cost:

$$C_m(\lambda_i) \equiv \frac{1}{2}\beta\lambda_i^2 T \quad \text{for } i = 1, ..., M. \tag{5}$$

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

$$C_t(\mu_j) \equiv \frac{1}{2}\gamma\mu_j^2 T \quad \text{for } j = 1, ..., N. \tag{6}$$

Thus, the cost of monitoring is proportional to the time interval and convex in the monitoring intensity.

Parameters $\beta, \gamma > 0$ control the level of monitoring costs for a given monitoring intensity. We say that market-makers' (resp. market-takers') monitoring costs become lower when β (resp. γ) decreases. Such a decline in monitoring costs can be a result, for example, of automation of the monitoring process. Thus, below, we analyze the effect of algorithmic trading on the trading process by considering the effect of a reduction in β and γ.

Timing. In reality, traders can change their monitoring intensities as market conditions change, whereas trading fees are usually fixed over a longer period of time. For this reason, it is natural to assume that traders choose their monitoring intensities after observing the fees set by the trading platform. Hence the trading game unfolds in three stages as follows:

Stage 1: The trading platform chooses the fees c_m and c_t.

Stage 2: Market-makers and market-takers simultaneously choose their individual monitoring intensities λ_i and μ_j.

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.

2.3 Objective functions and equilibrium

We now describe market participants' objective functions and define the notion of equilibrium that we use to solve for players' optimal actions in each stage.

Objective functions. Recall that a make/take cycle is a flow of events from the time the market is in state E until it goes back to this state. 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 p_i that she is first to post a competitive offer at price a after the market entered in state E. Given our assumptions on the monitoring process, this probability is $\frac{\lambda_i}{\lambda_1+\ldots+\lambda_M} = \frac{\lambda_i}{\bar{\lambda}}$. Thus, in each cycle, the expected profit (gross of monitoring costs) for market-maker i is $\frac{\lambda_i}{\bar{\lambda}} \cdot \pi_m = \frac{\lambda_i}{\bar{\lambda}} \left(\frac{\Delta}{2} - c_m\right)$.

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

$$\Pi_i(T) = E_{\tilde{n}_T}(\sum_{k=1}^{\tilde{n}_T} \frac{\lambda_i}{\bar{\lambda}} \pi_m) - \frac{1}{2}\beta\lambda_i^2 T,$$

where the expectation is taken over the number of completed cycles up to time T.

As is common in infinite horizon Markovian models, we assume that each player maximizes his/her long-term (steady-state) payoff per unit of time. Thus, market-maker i chooses his monitoring intensity to maximize

$$\Pi_{im} \equiv \lim_{T\to\infty} \frac{\Pi_i(T)}{T} = \lim_{T\to\infty} \frac{E_{\tilde{n}_T}(\sum_{k=1}^{\tilde{n}_T} p_i \pi_m)}{T} - \frac{1}{2}\beta\lambda_i^2. \tag{7}$$

Recall that $\mathcal{D}\left(\bar{\lambda}, \bar{\mu}\right)$ is the expected duration of a cycle. A standard theorem from the theory of stochastic processes (often referred to as the "Renewal Reward Theorem" see Ross (1996), p. 133) implies that

$$\lim_{T\to\infty} \frac{E_{\tilde{n}_T}(\sum_{k=1}^{\tilde{n}_T} p_i \pi_m)}{T} = \frac{\frac{\lambda_i}{\bar{\lambda}} \cdot \pi_m}{\mathcal{D}\left(\bar{\lambda}, \bar{\mu}\right)},$$

which is simply the expected profit for market maker i per make/take cycle divided by the expected duration of a cycle. An immediate implication is that the objective function of market-maker i (equation (7)) can be rewritten in a very intuitive way,

$$\Pi_{im} = \frac{\lambda_i}{\bar{\lambda}} \cdot \pi_m \cdot \mathcal{R}\left(\bar{\lambda}, \bar{\mu}\right) - \frac{1}{2}\beta\lambda_i^2. \tag{8}$$

That is, per unit of time, market-maker i maximizes the expected profit from a transaction ($\frac{\lambda_i}{\bar{\lambda}} \cdot \pi_m$) times the transaction rate, less monitoring costs. In a similar way, the objective function of market-taker j can be written as

$$\Pi_{jt} = \frac{\mu_j}{\bar{\mu}} \cdot \pi_t \cdot \mathcal{R}\left(\bar{\lambda}, \bar{\mu}\right) - \frac{1}{2}\beta\mu_j^2. \tag{9}$$

Finally, in each cycle, the trading platform earns a fee $\bar{c}$. Again, similar arguments show that the objective function of the exchange is given by the total fees per transaction times the transaction rate,

$$\Pi_e \equiv \bar{c} \cdot \mathcal{R}\left(\bar{\lambda}, \bar{\mu}\right) = (c_m + c_t) \cdot \mathcal{R}\left(\bar{\lambda}, \bar{\mu}\right). \tag{10}$$

Liquidity Externalities and Cross-Side Complementarities. An increase in the aggregate monitoring level of one side exerts a positive externality on the other side. To see this point, observe that $\frac{\partial \Pi_{im}}{\partial \bar{\mu}} > 0$ and $\frac{\partial \Pi_{jt}}{\partial \bar{\lambda}} > 0$. Intuitively, an increase in the aggregate monitoring intensity of market-makers (resp., market-takers) enlarges the rate at which market-takers (resp., market-makers) find trading opportunities and therefore renders them better-off. Moreover, the marginal benefit of monitoring for traders on one side increases in the aggregate monitoring level of traders on the other side since $\frac{\partial^2 \Pi_{im}}{\partial \bar{\mu} \partial \lambda_i} > 0$ and $\frac{\partial^2 \Pi_{jt}}{\partial \bar{\lambda} \partial \mu_j} > 0$. For this reason, market-makers (resp., market-takers) will inspect the state of the market more frequently when they expect market-takers (resp. market-makers) to inspect the state of the market more frequently. Thus, market-makers and market-takers' monitoring decisions reinforce each other. In other words, liquidity supply begets liquidity demand and vice versa. As we shall see, this complementarity in traders' decisions on both sides has important implications.

In contrast, an increase in the monitoring level of a trader hurts the traders who are on his or her side. That is, $\frac{\partial \Pi_{im}}{\partial \lambda_j} < 0$ and $\frac{\partial \Pi_{it}}{\partial \mu_j} < 0$ (for $j \neq i$). This effect captures the fact that traders on the same side are engaged in a "horse race" to be first to detect a trading opportunity when it appears. In reality, this aspect is a key reason for automating order submission.[17]

Equilibrium. The strategies for the market-makers and market-takers are their monitoring intensities λ_i and μ_j respectively. A strategy for the trading platform is

[17] Dee for instance "Tackling latency-the algorithmic arms race," IBM Global Business Services report.

a menu of fees (c_m, c_t) for a fixed total fee level $\bar{c} = c_m + c_t$. We solve the model backwards. First, for a given set of fees (c_m, c_t), we look for Nash equilibria in monitoring intensities in Stage 2.[18] Using (8) and (9), a Nash equilibrium in this stage is a vector of monitoring intensities $(\lambda_1^*, \ldots, \lambda_M^*, \mu_1^*, \ldots, \mu_N^*)$ such that for all $i = 1, \ldots, M$, λ_i^* maximizes (8), and for all $j = 1, ..., N$, μ_j^* maximizes (9), taking the monitoring intensities of all other traders as fixed.

Note that λ_i and μ_j affect (8) and (9) both directly and indirectly through their effect on $\bar{\lambda}$ and $\bar{\mu}$, and therefore through the trading rate. Thus, when optimizing, individual traders trade-off the marginal effect of increased monitoring on their probability of winning and on the trading rate, against the marginal cost of monitoring. Thus, the strategic complementarity between the two sides plays an important role in the determination of equilibrium.

Given a Nash equilibrium in the monitoring intensities, we solve for the fee structure (c_m^*, c_t^*) that maximizes the trading platform's expected profit (equation (10)). In most of the paper we assume that $\bar{c}$ is fixed to better focus the analysis on the fee structure. It is straightforward to endogenize $\bar{c}$, as shown in Section 4.1.

3 Monitoring, Liquidity, and Welfare with Fixed Fees

In this section we study the equilibrium monitoring intensities for a given set of fees (c_m, c_t). We will show that for all parameters values, the model has two equilibria: (i) an equilibrium with no trading; and (ii) an equilibrium with trading. This multiplicity of equilibria is due to the complementarity in monitoring decisions discussed in the previous section, which leads to a coordination problem between the two sides.

To see this point, consider how the no-trade equilibrium arises. If market-takers do not monitor the quotes on the trading platform, then maker-makers do not expect any arrival of market-orders. Given that monitoring is costly, each market-maker optimally sets $\lambda_i^* = 0$. Similarly, if market-makers do not monitor, then market-takers expect no competitive quotes to be posted. Again, since monitoring is costly, each market-taker will optimally set $\mu_j^* = 0$. Thus, traders' beliefs that the other side will not be active are self-fulfilling and result in a no-monitoring, no-trade equilibrium.

Proposition 1 *:For any given set of fees, there is an equilibrium in which traders do not monitor: $\lambda_i^* = \mu_j^* = 0$ for all $i \in \{1, \ldots, M\}$ and $j \in \{1, \ldots, N\}$. The trading*

[18] Note that c_m and c_t affect the optimization in (8) and (9) through their effect on π_m and π_t.

volume in this equilibrium is zero.

A second equilibrium does involve monitoring and trade. To describe this equilibrium, let $r \equiv \frac{\gamma}{\beta}$ be the relative monitoring costs of market takers vs. market makers. And, let

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

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

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

$$\lambda_i^* = \frac{M + (M-1)\,\Omega^*}{(1+\Omega^*)^2} \cdot \frac{\pi_m}{M\beta} \qquad i = 1, \ldots, M \tag{11}$$

$$\mu_j^* = \frac{\Omega^*\left((1+\Omega^*)\,N - 1\right)}{(1+\Omega^*)^2} \cdot \frac{\pi_t}{N\gamma} \qquad j = 1, \ldots, N \tag{12}$$

where Ω^* *is the unique positive solution to the cubic equation*

$$\Omega^3 N + (N-1)\Omega^2 - (M-1)\,z\Omega - Mz = 0. \tag{13}$$

Moreover, in equilibrium, $\frac{\bar{\lambda}^*}{\bar{\mu}^*} = \Omega^*$.

It is interesting to note that the no-trade equilibrium is highly unstable, whereas the equilibrium with trade is stable. That is, although no-trade is an equilibrium, a slight deviation from zero monitoring by one side will attract a large amount of monitoring on the other side, which in turn will generate a cascade of monitoring that will end up in the equilibrium with trade. To illustrate this, consider the case $M = N = 1$. Figure 2 plots the reaction functions, denoted by $\rho_m(\mu_1)$ and $\rho_t(\lambda_1)$, for the market-maker and market-taker respectively, with $\pi_m = \pi_t = 0.5$, $\beta = \gamma = 0.5$.[19] Note that $\rho_m(\cdot)$ is plotted as a function of μ_1 (the horizontal axis) whereas $\rho_t(\cdot)$ is plotted as a function of λ_1 (the vertical axis). The two reaction functions meet at two points (0,0) and (0.25,0.25), which are the two equilibria in this example (from Propositions 1 and 2). It can be verified that the slope of both reaction functions at zero is infinite. Thus, even a slight deviation by one side from zero monitoring (for example due to an exogenous injection of order flow) will lead to a cascade

[19] That is, $\rho_m(\mu_1)$ is the best response of the market-maker given that the market-taker's monitoring level is μ_1, and vice-versa.

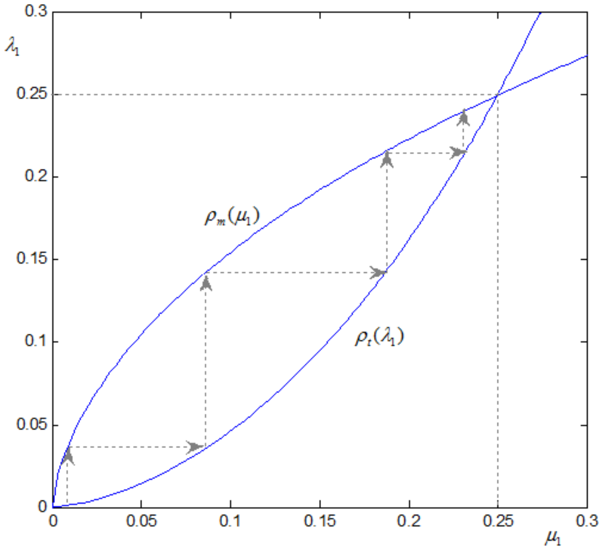

Figure 2: Monitoring Recation Functions

of larger deviations. The figure illustrates the outcome of such a deviation using the dashed arrows. A small increase in μ_1 optimally attracts a relatively large λ_1, which in turn optimally attracts an even larger μ_1, and so on. This process ends only when the reactions converge to the unique equilibrium with trade. Of course, the complementarity between the two sides is key to this process in which increased monitoring by one side reinforces monitoring by the other side.

Given the fragility of the no-trade equilibrium, we dedicate the rest of the paper to studying the properties of the equilibrium with trade derived in Proposition 2.

Trading rate, aggregate monitoring, and cross-side effects. We first use Proposition 2 to study how a change in the exogenous parameters (the number of participants on either side, the trading fees, and the monitoring costs) affect the aggregate monitoring levels of both sides and the trading rate.

Corollary 1 *In the unique equilibrium with trade,*

1. *The aggregate monitoring level of each side increases in the number of participants on either side (*$\frac{\partial\bar{\lambda}^*}{\partial N} > 0$, $\frac{\partial\bar{\lambda}^*}{\partial M} > 0$, $\frac{\partial\bar{\mu}^*}{\partial N} > 0$, $\frac{\partial\bar{\mu}^*}{\partial M} > 0$*) and decreases in*

(i) monitoring costs ($\frac{\partial\bar{\lambda}^*}{\partial\beta} < 0$ *,* $\frac{\partial\bar{\lambda}^*}{\partial\gamma} < 0$*,* $\frac{\partial\bar{\mu}^*}{\partial\beta} < 0$ *,* $\frac{\partial\bar{\mu}^*}{\partial\gamma} < 0$*) or (ii) the fee per trade charged on either side (*$\frac{\partial\bar{\lambda}^*}{\partial c_m} < 0$ *,* $\frac{\partial\bar{\lambda}^*}{\partial c_t} < 0$*,* $\frac{\partial\bar{\mu}^*}{\partial c_m} < 0$ *,* $\frac{\partial\bar{\mu}^*}{\partial c_t} < 0$*).*

2. *The trading rate decreases in (i) the monitoring costs (*$\frac{\partial\mathcal{R}(\bar{\lambda}^*,\bar{\mu}^*)}{\partial\beta} < 0$ $\frac{\partial\mathcal{R}(\bar{\lambda}^*,\bar{\mu}^*)}{\partial\gamma} < 0$*) or the trading fees (*$\frac{\partial\mathcal{R}(\bar{\lambda}^*,\bar{\mu}^*)}{\partial c_m} < 0$ *and* $\frac{\partial\mathcal{R}(\bar{\lambda}^*,\bar{\mu}^*)}{\partial c_t} < 0$*) and (ii) increases in the number of participants on either side (*$\frac{\partial\mathcal{R}(\bar{\lambda}^*,\bar{\mu}^*)}{\partial M} > 0$ *and* $\frac{\partial\mathcal{R}(\bar{\lambda}^*,\bar{\mu}^*)}{\partial N} > 0$*).*

The cross-side complementarity in monitoring decisions (discussed at the end of Section 2.3) is key for this finding. Indeed, this corollary implies that a change in a parameter that directly affects the aggregate monitoring level of one side also affects the aggregate monitoring level of the other side. To see this, consider a decrease in the monitoring cost for market-makers. This decrease directly raises their individual monitoring levels, other things equal. Consequently, the marginal benefit of monitoring for market-takers is higher as they are more likely to find a good price when they inspect the market. Thus, market-takers monitor the market more intensively. This indirect effect reinforces market-makers' attention and thereby triggers a chain reaction that raises the trading rate.

Figure 3 uses a simple example to illustrate the importance of the cross-side complementarities. We assume $M = N = 10$, $L = \Delta = 1$, and $c_m = c_t = 0.05$. We also fix $\gamma = 1$ and study how a reduction in monitoring costs on market-makers (β) affects the trading rate $\mathcal{R}$. The horizontal axis is $1/\beta$, and the vertical axis is the trading rate. The solid curve depicts the equilibrium trading rate calculated using (11) and (12) for any given level of β. A reduction in β affects $\bar{\lambda}^*$ directly and $\bar{\mu}^*$ indirectly (through the cross-side complementarities). Both these effects are reflected in the solid line. By contrast, the dotted line depicts the direct effect only. To plot this curve we use the equilibrium value of $\bar{\lambda}^*$, but keeps $\bar{\mu}^*$ at its original value (calculated for $1/\beta = 0.5$). Thus, this curve reflects a hypothetical trading rate that ignores any cross-side complementarities. By comparing the two curves in the figure it is evident that cross-side complementarities have a material impact on monitoring intensities and on the trading rate.

The same reasoning explains why an increase in the trading fee charged on market-makers (resp. market-takers) has a negative impact on the aggregate monitoring levels of **both** sides other things equal, although the cost of trading for market-takers (resp. market-makers) does not change. The trading platform must therefore

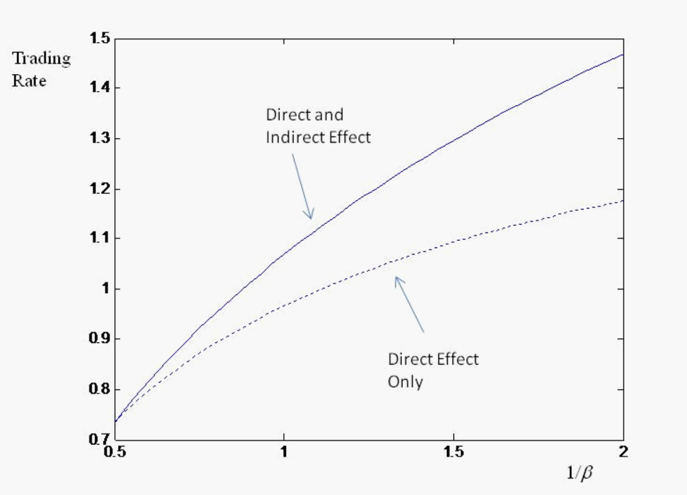

Figure 3: Cross-Side Complimentarities and the Trading Rate

account for cross-side complementarities in solving for its optimal fees, as shown in the next section.

An increase in the number of participants on one side has a positive impact on the aggregate monitoring on both sides for the same reason. In this case, however, the individual monitoring levels of the market participants on the side that becomes more populated may decrease. Indeed, as more traders on one side compete for trading opportunities, the likelihood of winning a profit opportunity declines. This competition effect decreases the incentive to monitor of each participant on the side that becomes thicker. Yet, this competition effect remains small as it is offset by the cross-side complementarity effect, which is conducive to more monitoring by each participant.

Welfare and algorithmic trading. The aggregate expected profit (per unit of time) for all market participants is a measure of welfare. We denote it by W. Using equations (8), (9), and (10), we obtain:

$$\begin{aligned} W(\gamma,\beta,c_m,c_t,M,N) &\equiv \textstyle\sum_{i=1}^{M}\Pi_{im}+\sum_{j=1}^{N}\Pi_{jt}+\Pi_e \\ &= \mathcal{R}\left(\bar{\lambda}^{*},\bar{\mu}^{*}\right)\cdot L - M\cdot C_m(\lambda_1^{*}) - N\cdot C_t(\mu_1^{*}). \end{aligned}$$

Thus, *other things being equal*, aggregate welfare enlarges with the trading rate. For this reason, a decrease in market participants' monitoring costs or in trading fees raise aggregate welfare, as shown in the next corollary.

Corollary 2 *The following hold:*

1. *The total expected profit of each class of participants (the market-makers, the market-takers, and the platform) decreases with the monitoring cost on either side (β or γ). Thus, aggregate welfare decreases in monitoring costs.*

2. *Aggregate welfare decreases in trading fees on either side (c_m or c_t).*

The first part of the proposition implies that algorithmic trading can be socially useful. As monitoring costs decrease, both market-makers and market-takers complete their trades more quickly. Consequently, the trading rate per unit of time increases. This means that the rate at which gains from trade are realized is higher, which makes all participants better-off.

A higher trading fee on one side results in a smaller trading rate. Thus, it leads to a loss in aggregate welfare as the rate at which gains from trade are realized is smaller. Of course, a higher trading fee may be beneficial for the trading platform. But overall, the increase in expected profit for the platform is more than offset by the decline in expected profits for the traders.

The balance between liquidity supply and demand. As explained in the previous section, we can view $\bar{\lambda}^*$ as a measure of liquidity supply and $\bar{\mu}^*$ as a measure of liquidity demand. The next corollary shows that liquidity supply and demand are not necessarily balanced in equilibrium (that is, in general, $\bar{\lambda}^* \neq \bar{\mu}^*$). This finding is important as the optimal pricing policy for the trading platform consists of choosing fees to reduce imbalances in the "supply" and "demand" of liquidity, as explained in the next section.

Corollary 3 *In equilibrium, for fixed fees, the market-making side monitors the market more intensively (less) than the market-taking side ($\bar{\lambda}^* > \bar{\mu}^*$) if and only if $\frac{z(2M-1)}{2N-1} > 1$. If $\frac{z(2M-1)}{2N-1} = 1$, the market-making and the market-taking sides have identical monitoring intensities.*

Thus, in equilibrium, there is excess attention by the market-making side (resp. market-taking side) when $\frac{z(2M-1)}{(2N-1)} > 1$ (resp. $\frac{z(2M-1)}{(2N-1)} < 1$). For instance, if $M = N$ and $\frac{\pi_m}{\beta} > \frac{\pi_t}{\gamma}$, the market-making side inspects the market more frequently than the market-taking side because market-makers' cost of missing a trading opportunity is relatively higher. In this case, liquidity supply is abundant but in part useless since market-takers check the market relatively infrequently.

Small and large markets. In general we do not have an explicit solution for traders' monitoring levels because we cannot solve for Ω^* in closed-form (Ω^* is the unique positive root of equation (13)). However, there are two polar cases in which we can do so. The analysis of these cases will be useful to form intuition about the optimal pricing policy of the trading platform in the next section.

In the first case, the market features one market-maker and one market-taker ($M = 1$ and $N = 1$). We refer to this case as "the small market." In this case, (13) gives $\Omega^* = z^{\frac{1}{3}}$. Thus, using equations (11) and (12), we obtain,

Corollary 4 *The monitoring intensities when $M = N = 1$ are given by,*

$$\lambda_1^* = \frac{1}{\left(1 + z^{\frac{1}{3}}\right)^2} \cdot \left(\frac{\pi_m}{\beta}\right), \tag{14}$$

$$\mu_1^* = \frac{1}{\left(1 + z^{-\frac{1}{3}}\right)^2} \cdot \left(\frac{\pi_t}{\gamma}\right). \tag{15}$$

In the second case, that we term "the large market," the number of participants on both sides is very large (both M and N tend to infinity) yet the ratio $q \equiv \frac{M}{N}$ is fixed. This ratio measures the size of the market-making side *relative* to the size of the market-taking side. First, it is easily verified from (13) that

$$\Omega^\infty \equiv \lim_{M \to \infty} \Omega^* = (zq)^{\frac{1}{2}}.^{20} \tag{16}$$

Furthermore, the individual monitoring intensities converge to finite levels given in the next corollary.

Corollary 5 *Let $q > 0$ be fixed, and assume $N = \frac{M}{q}$. Then,*

$$\lambda_i^\infty \equiv \lim_{M \to \infty} \lambda_i^* = \frac{1}{1 + (zq)^{\frac{1}{2}}} \cdot \frac{\pi_m}{\beta} \qquad i = 1, 2, 3, ... \tag{17}$$

$$\mu_j^\infty \equiv \lim_{M \to \infty} \mu_j^* = \frac{1}{1 + (zq)^{-\frac{1}{2}}} \cdot \frac{\pi_t}{\gamma} \qquad j = 1, 2, 3, ...$$

[20] To see this note that from (13), $z = \frac{\Omega^{*3}\frac{M}{q} + (\frac{M}{q} - 1)\Omega^{*2}}{(M-1)\Omega^* + M}$. Then, take the limit as $M \to \infty$.

It can be formally shown that the convergence to this limiting case is pretty fast.[21] That is, using the closed-form solutions in Lemma **??** to study markets with a finite number of traders provides good approximations even for relatively low values of M and N. Notice the similarities in the expressions for the monitoring rates in the two extreme cases in corollaries (4) and (5).

4 The determinants of the make/take spread

Now, we study the fees set by the trading platform. In most of the analysis, we fix the total fee charged by the trading platform, $\bar{c}$, as we are mainly interested in the fee structure, (c_m, c_t). We refer to $c_m - c_t$ as the *make/take spread.* The make/take spread is zero when the fee structure is flat (i.e., $c_m = c_t$) and positive (negative) if the market-making side pays a higher (lower) fee than the market-taking side. Our goal is to understand how the exogenous parameters of the model (the tick size, the monitoring costs, and the relative number of participants on each side) affect the make-take spread. For instance, we study the conditions under which the optimal make-take spread is negative ($c_m < c_t$), as often observed in reality.

As explained in Section 2.3, for a given total fee $\bar{c}$, the objective function of the trading platform is to find a fee structure (c_m^*, c_t^*) that solves,

$$\begin{aligned} \max_{c_m, c_t} \Pi_e &= (c_m + c_t)\,\mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*), \\ s.t\text{:} \quad & c_m + c_t = \bar{c} \end{aligned} \tag{18}$$

Trading fees affect traders' monitoring decisions and thereby the trading rate (see Corollary 1). Since $c_m + c_t = \bar{c}$ is fixed, the first order conditions for the trading platform's optimization problem impose that

$$\frac{\partial \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_m} = \frac{\partial \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_t}. \tag{19}$$

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.

An increase in the fee charged on market-makers has a negative effect on the aggregate monitoring levels of the market-makers and the market-takers. We denote the elasticity of these levels to the fee charged on market-makers by η_{mm} and η_{mt}.

[21] Formally, we can show that $\Omega^{\infty} - \Omega^*$ is $O\left(\frac{1}{M}\right)$, which means that the error in using Ω^{∞} to approximate for Ω^* is on the order of magnitude of $\frac{1}{M}$. The proof is available upon request.

Similarly, η_{tt} and η_{tm} are the elasticities of the aggregate monitoring levels of the market-taking side and the market-making side to the fee charged on market-takers. Thus

$$\begin{aligned}\eta_{mm} &\equiv \left(\frac{\partial\log(\bar{\lambda}^*)}{\partial c_m}\right)c_m \text{ and } \eta_{mt} \equiv \left(\frac{\partial\log(\bar{\lambda}^*)}{\partial c_t}\right)c_t, \\ \eta_{tt} &\equiv \left(\frac{\partial\log(\bar{\mu}^*)}{\partial c_t}\right)c_t \text{ and } \eta_{tm} \equiv \left(\frac{\partial\log(\bar{\mu}^*)}{\partial c_m}\right)c_m.\end{aligned} \tag{20}$$

Using equation (19), we obtain the following result.

Proposition 3 *For each level $\bar{c}$ of the total fee charged by the platform, the optimal fee structure must satisfy:*

$$\begin{aligned}c_m^* &= \left(\frac{h}{h+1}\right)\bar{c}, \\ c_t^* &= \bar{c} - c_m^* = \left(\frac{1}{h+1}\right)\bar{c},\end{aligned} \tag{21}$$

where $h \equiv \frac{(\bar{\lambda}^*)^{-1}\eta_{mm}+(\bar{\mu}^*)^{-1}\eta_{tm}}{(\bar{\lambda}^*)^{-1}\eta_{mt}+(\bar{\mu}^*)^{-1}\eta_{tt}}$.

Thus, it is optimal to charge different fees on market-makers and market-takers, unless $h = 1$. Moreover, the optimal fee structure depends on the elasticities of the aggregate monitoring levels to the fees and cross-side elasticities (η_{mt} and η_{tm}). This finding implies that estimating these elasticities is important to determine the optimal fee structure.

The previous lemma does not provide a closed-form solution for the trading fees since the elasticities of monitoring levels to trading fees depend on the fees. We can obtain analytical solutions in two particular cases: the large market case and the small market case. We first study the effects of the exogenous parameters on the make-take spread in these two cases. We then show using numerical simulations that the conclusions obtained in these two polar cases generalize to intermediate values of M and N.

4.1 Fees in the Large Market

Consider the case of the "large market" introduced in the previous section: both M and N tend to infinity, and yet $\frac{M}{N} = q$, where $q > 0$ reflects the relative size of the market-making vs. the market-taking sides. Recall from Corollary 5 that the individual monitoring frequencies converge to a finite limit. While the total level of monitoring diverges, the fees that maximize the trading rate converge to a finite limit. This allows us to obtain a closed form solution for the fee structure in this case.

Proposition 4 *In the large market case, the trading platform optimally allocates its fee $\bar{c}$ between the market-making side and the market-taking side as follows:*

$$c_m^* = \frac{1}{2}\left(\Delta - \frac{2(L-\bar{c})}{(1+(qr)^{\frac{1}{3}})}\right) \quad \textit{and } c_t^* = \bar{c} - c_m^*. \tag{22}$$

For these fees,

$$\pi_m^* = \frac{L-\bar{c}}{(1+(qr)^{\frac{1}{3}})} \quad \textit{and} \quad \pi_t^* = \frac{L-\bar{c}}{(1+(qr)^{-\frac{1}{3}})}, \tag{23}$$

and the equilibrium monitoring intensities are:

$$\lambda_i^\infty = \frac{L-\bar{c}}{\beta\left(1+(qr)^{\frac{1}{3}}\right)^2} \quad \textit{and} \quad \mu_j^\infty = \frac{L-\bar{c}}{\gamma\left(1+(qr)^{-\frac{1}{3}}\right)^2} \quad \textit{for } i,j = 1,2,... \tag{24}$$

Using these results we can explore how the tick size, the monitoring costs and the ratio of market participants on both sides determine the optimal fee structure Let $\bar{\Delta}(q,r) \equiv 2(L-\bar{c})\left(1+(qr)^{\frac{1}{3}}\right)^{-1} + \bar{c}$. Using equation (22), we obtain the following result.

Corollary 6 *In the large market, the make-take spread increases with (i) the tick size, Δ; (ii) the relative size of the market-making side, q; and (iii) the relative monitoring cost for the market-taking side, r. Moreover the make-take spread is negative if and only if $\Delta < \bar{\Delta}(q,r)$.*

Figure 4 illustrates the set of parameters for which the make-take spread is negative or positive.

The make-take spread is more likely to be negative when (i) the tick size is small, (ii) the number of market-makers is relatively small or (iii) the monitoring cost for market-makers is relatively large. These findings follow from the same general principle. Namely, when a change in parameters increases the level of monitoring of one side relative to the level of monitoring of the other side, the trading platform raises its fee on the side whose monitoring increases. In other words, the trading platform

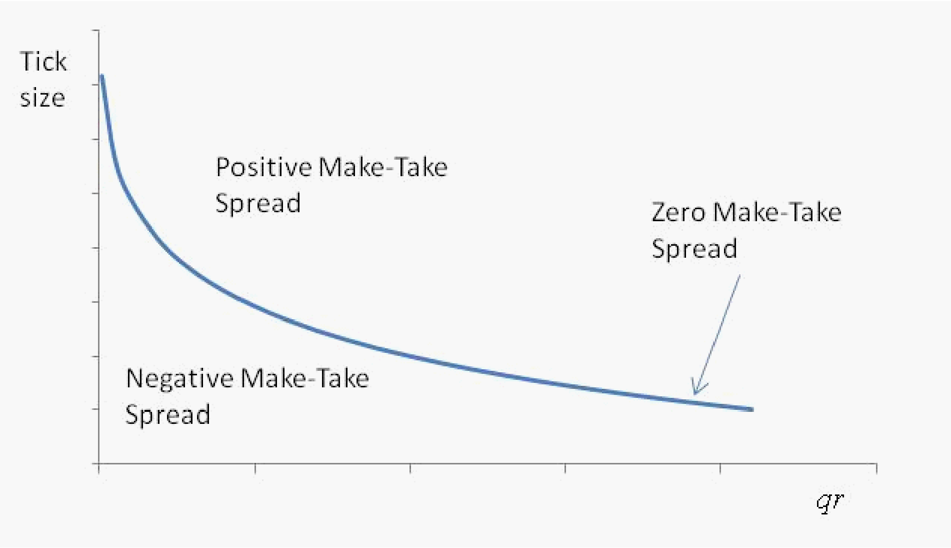

Figure 4: Determinants of the Sign of the Make/Take Spread

uses its fee structure to balance the level of attention of the market-making side and the market-taking side.

For instance, consider an increase in the tick size. This increase reinforces market-makers' incentive to monitor since, other things equal, they get a larger fraction of the gains from trade when they participate to a trade. In contrast, market-takers' incentive to inspect the state of the market is lower. Thus, to better balance the level of attention of both sides, it is optimal for the platform to charge a larger fee on the market-makers and a smaller fee on the market-takers.

The effect of an increase in the relative size of the market-making side (q) or the ratio of market-takers to market-makers' monitoring cost ($r = \frac{\gamma}{\beta}$) on the make-take spread can be understood in the same way. Intuitively, an increase in the relative size of the market-making side or a decrease in its relative monitoring cost enlarge the monitoring intensity of this side relative to the market-taking side, other things equal. Thus, to balance the level of attention on both sides, it is optimal for the the trading platform to raise its fee on the market-making side when q or r increase.

Equation (22) implies that market-makers (resp. market-takers) are optimally subsidized (they pay a negative trading fee) when the tick size is small (resp. large) enough. Observe however that in all cases $c_m^* > -\frac{\Delta}{2}$ since $L \leq \Delta$. Thus, even if she receives a rebate, it is not optimal for a market-maker to post a quote below her valuation of the security, i.e., at $a - \Delta$ as this would result in an expected loss for the dealer ($a - \Delta - v_0 - c_m^* < 0$).

The previous findings about the optimal fee structure hold for any level $\bar{c}$. Thus, they would hold even if the total fee earned by the platform on each trade is arbitrarily capped at some level. If the trading platform is free to choose its total fee, $\bar{c}$, then it faces the standard price-quantity trade-off for a monopolist. That is, by raising $\bar{c}$, the trading platform gets a larger revenue per trade but it decreases the rate at which trades occur (Corollary 1). The next corollary provides the optimal value of $\bar{c}$ for the trading platform in this case.

Corollary 7 *The trading platform maximizes its expected profit by setting its total trading fee at $\bar{c} = L/2$ and by splitting this fee between both sides as described in Proposition 4.*

Thus, in contrast to the fee structure, the optimal fee for the platform is independent of the tick size, traders' monitoring costs and the relative size of the market-making side. Thus, our results regarding the effect of Δ, q, and r hold even if $\bar{c}$ is set by the trading platform.

4.2 The small market ($M = N = 1$)

We now consider the case with one market-maker and one market-taker. Using the expressions for monitoring levels on each side (Equations (14) and (15)), we can solve for the optimal fee structure of the platform in this case. We obtain the following result.

Proposition 5 *When $M = N = 1$, the trading platform optimally allocates its fee $\bar{c}$ between the market-making side and the market-taking side as follows:*

$$c_m^* = \frac{1}{2}\left(\Delta - \frac{2(L-\bar{c})}{(1+r^{\frac{1}{4}})}\right) \quad \text{and } c_t^* = \bar{c} - c_m^*. \tag{25}$$

For these fees,

$$\pi_m^* = \frac{L-\bar{c}}{(1+r^{\frac{1}{4}})} \quad \text{and} \quad \pi_t^* = \frac{L-\bar{c}}{(1+r^{-\frac{1}{4}})}, \tag{26}$$

and the equilibrium monitoring intensities are:

$$\lambda_1^* = \frac{L-\bar{c}}{\beta\left(1+r^{\frac{1}{4}}\right)^3} \quad \text{and} \quad \mu_1^* = \frac{L-\bar{c}}{\gamma\left(1+r^{-\frac{1}{4}}\right)^3}. \tag{27}$$

Clearly, this result is qualitatively similar to Proposition 4. In particular, it is readily checked that our findings regarding the effects of the tick size, and the relative monitoring cost of market-takers (Corollary 6) still hold in this case. Moreover, in this case as well, it is optimal for the trading platform to set $\bar{c} = L/2$.

4.3 General Case

The fact that Ω^* is only given implicitly for arbitrary M and N prevents us from obtaining an analytical solution for the optimal trading fees for arbitrary values of the parameters. However, we have checked through extensive numerical simulations that the comparative static results obtained in the large market and small market cases are robust. As an illustration, consider the following baseline values for the parameters: $M = N = 10$; $\gamma = \beta = 1$; $\Delta = 1$ (1 penny), $L = 1$ (1 penny). $\bar{c} = 0.1$ (0.1 pennies).

Figure 5a shows how the market-taking fee (dotted line), the market-making fee (plain line) and the make-take spread (dashed line) change as the tick size increases. As found in the large market and small market cases, the make-take spread increases as the tick size gets larger. As before, the intuition is that a larger tick-size benefits market-makers, and hence they monitor more. To maximize the trading-rate, the exchange penalizes the market-makers by increasing the maker-take spread.

Figure 5b considers the effect of an increase in $r = \frac{\gamma}{\beta}$ on the trading fees and the make-take spread. As expected, the make-take spread increases when the monitoring cost becomes relatively larger for market-takers. Finally Figure 5c considers the effect of an increase in $q = M/N$ on the trading fees and the make-take spread. As expected, the make-take spread increases as the number of market-makers increases relative to the number of market-takers. The intuition as again as in the small and large markets.

5 Implications

We now discuss the implications of the model in more details. Throughout, we focus on the large market case. But the implications discussed here hold more generally.

Duration Clustering and Cross-Side Complementarities. As explained in Section 2.3, market-makers' monitoring decisions and market-takers' monitoring decisions reinforce each other. This complementarity naturally leads to a positive cor-

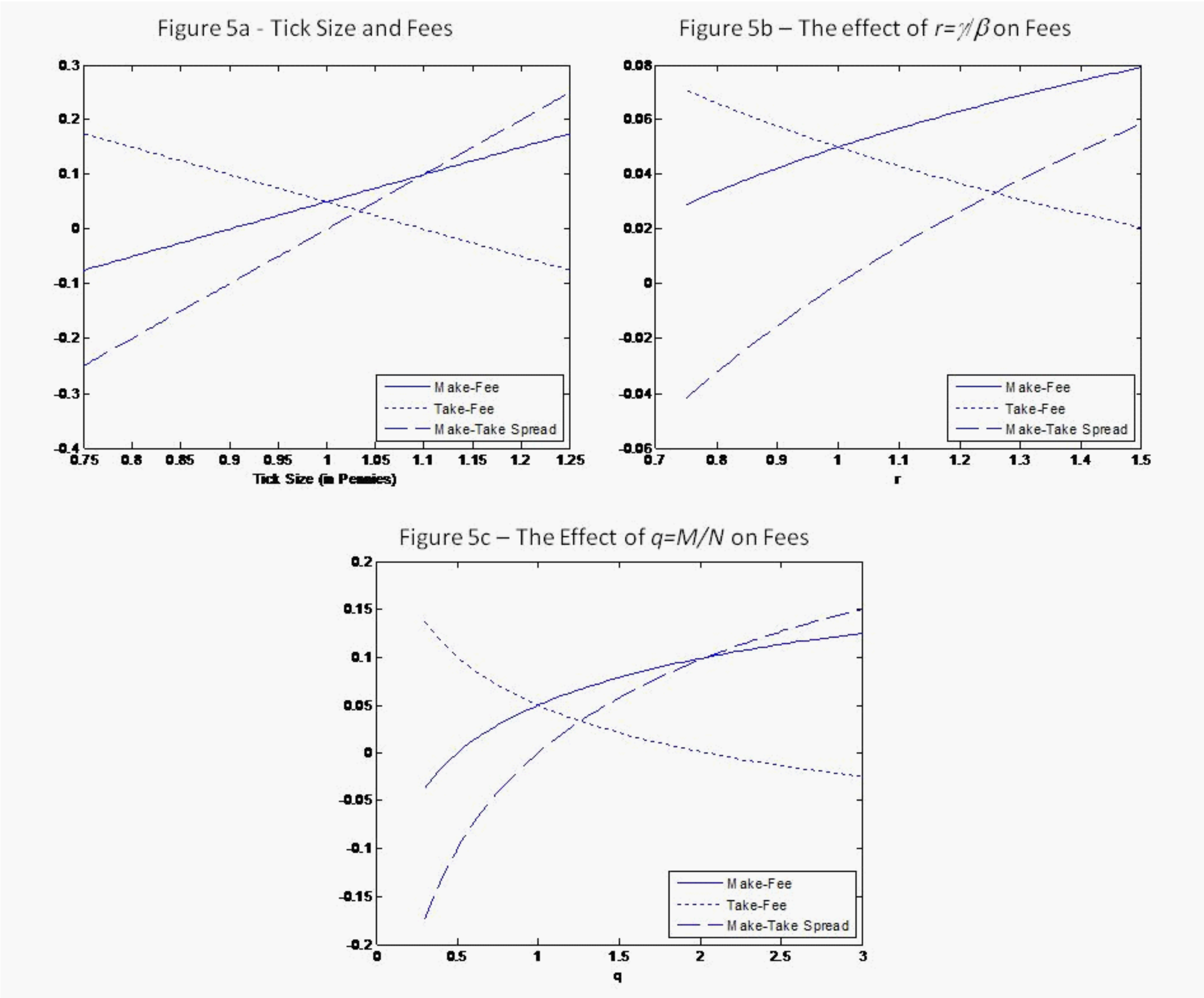

Figure 5: Determinants of the Make/Take Spread

relation between (i) the average time it takes for the bid-ask spread to revert to its competitive level after a trade (denoted by $\mathcal{D}_m \equiv \frac{1}{\lambda}$) and (ii) the average time it takes for a trade to occur when the bid-ask spread is competitive (denoted by $\mathcal{D}_t \equiv \frac{1}{\mu}$).

For instance, consider an increase in the number of market-takers. In equilibrium, this shock triggers a decrease in the reaction times of (i) the market-taking side (as they monitor more) and (ii) the market-making side (as more monitoring by market-takers encourages more monitoring by market-makers). Thus, both $\mathcal{D}_m$ and $\mathcal{D}_t$ fall. As a consequence, the duration between trades ($\mathcal{D}_m + \mathcal{D}_t$) falls as well (these claims follow directly from Corollary 1).[22]

This positive correlation between the average durations of each phase in a cycle echoes the clustering in the time intervals between consecutive transactions (trade durations) found in several empirical papers (e.g., Engle and Russell (1998)). In general, this clustering has been interpreted in light of models of trading with asymmetric information (e.g., Admati and Pfleiderer (1988)). In these models, clustering arises as liquidity traders optimally choose to trade at the same point in time. Instead, our model suggests that clustering in trade durations could stem from the complementarity in monitoring decisions between liquidity suppliers and liquidity demanders. In this case, a factor shortening the reaction time of one side shortens the reaction time of the other side as well. Thus, time-variations in this factor (e.g., the number of market-takers during the trading day) create a positive correlation between the various components ($\mathcal{D}_m$ and $\mathcal{D}_t$) of the total duration of a cycle and results in clustering in trade durations.

Our model implies that cross-sectional variations or exogenous shocks to determinants of monitoring can help estimate liquidity externalities. In fact, our model offers a direct way to structurally estimate this externality, which has proved hard to identify empirically (see for example Barclay and Hendershott (2004)).

Time Structure of a Cycle. Let $\mathcal{I} \equiv \frac{\mathcal{D}_t}{\mathcal{D}_m} = \frac{\overline{\lambda}}{\overline{\mu}}$. This is the average duration from a trade to a competitive quote divided by the average duration from a competitive quote to a trade (see Figure 1). This ratio can serve as an empirical proxy for the ratio $\frac{\overline{\lambda}}{\overline{\mu}}$, which is unobservable empirically. A value of $\mathcal{I}$ larger (resp. smaller) than 1 indicates that market-makers monitor the market more than market-takers. Thus, after a trade, the speed at which the bid-ask spread reverts to its competitive level is higher than the speed at which competitive quotes are hit by market-takers. In this sense $\mathcal{I}$ is a measure of the imbalance between liquidity supply and liquidity demand. We obtain the following result.

Corollary 8 *In equilibrium, for fixed fees, the imbalance between liquidity supply and*

[22] Thus, complementarity in the actions of market-makers and market-takers could explain why limit order markets exhibit sudden and short-lived booms and busts in trading rates during the trading day (see Hasbrouck (1999) or Coopejans, Domowitz and Madhavan (2001) for empirical evidence).

demand in the large market is:

$$\mathcal{I}(r,q,c_m,c_t) = (\frac{\pi_m rq}{\pi_t})^{1/2}. \tag{28}$$

Thus, the imbalance increases in (i) the relative size of the market-making side, (ii) the relative monitoring cost of the market-taking side, (iii) the fee charged on market-makers, and (iv) the fee charged on market-takers.

The first two implications (those regarding the effect of q and r on $\mathcal{I}$) also hold when fees are set at their optimal level. Indeed, using Proposition 4 and equation (28), we obtain that:

$$\mathcal{I}(r,q,c_m^*,c_t^*) = (\frac{\pi_m^* rq}{\pi_t^*})^{1/2} = (rq)^{2/3}. \tag{29}$$

The optimal make-take spread is also positively related to r and q (see Corollary 6). Thus, if fees are set optimally, the model implies a positive correlation between the make-take spread and imbalance. This prediction is interesting as the make-take spread varies (i) across securities for a given trading platform (see Table 1 in the introduction) and (ii) across trading platforms, for a given security (in which case q may differ across platforms). These variations provide a way to test whether the make-take spread co-varies positively with the imbalance.

Tick size and Make-Take Spread. The model also implies a positive association between the make-take spread and the tick size. Trading platforms' pricing policies are consistent with this implication. Indeed, the proliferation of negative make-take spreads in U.S. equity trading platforms (and even rebates paid to limit order traders) coincides with a reduction in the tick size on these platforms. Moreover, this practice was introduced by ECNs such as Archipelago and Island in the 90s which, at this time, were operating on much finer grids than their competitors (Nasdaq and NYSE).[23] Since January 2007, the tick size has been reduced for a list of options in U.S. option markets (so called "The Penny Pilot Program"). For these options, as implied by the model, a few trading platforms (e.g., NYSE Arca Options and the Boston Options Exchange) now charge a negative make-take spread. Lastly, in 2009, BATS decided to charge a positive make-take spread in stocks with a relative large tick size (i.e., low priced stocks).

[23] Biais, Bisière and Spatt (2002) stress the importance of the finess of the grid on Island for the competitive interactions between this platform and Nasdaq, Island's main competitor at the time of their study.

The model suggests two other reasons for the low make-take spreads that are observed in reality (see Corollary 6). This configuration could also arise because the size of the market-making sector is relatively small and/or because monitoring costs for this sector are relatively high. This situation is not implausible. First, in recent years, the burden of liquidity provision seems to rest on a relatively small number of market participants (GETCO, ATD, Citadel Derivatives etc...) who specialize in high-frequency market-making by actively monitoring the market. Thus, q could be small in reality. Moreover, brokers who must take a position in a list of stocks on behalf of their clients need to focus only on trading opportunities in this list of names. In contrast, electronic market-makers monitor the entire universe of stocks, unless they decide to specialize. Thus, their opportunity cost of monitoring one stock is likely to be higher than for market-takers.

Trading Activity and the Tick Size. The model also implies that, for **fixed trading fees**, there is a value of the tick size that maximizes the trading rate, as shown in the next corollary. For this corollary, let $c^*_m(L, q, r)$ denotes the optimal fee charged on market-makers in the special case when $\Delta = L$.

Corollary 9 *1. For fixed trading fees, the tick size that maximizes the trading rate is: $\Delta^* = 2(c_m - c^*_m(L, q, r)) + L$. Thus, Δ^* increases in (i) the fee charged on market-makers (c_m), and decreases in (ii) the number of market-makers relative to the number of market-takers (q) or (iii) market-takers' monitoring cost relative to market-makers' monitoring cost (r).*

2. In contrast, if the fees are set optimally, then a change in the tick size has no effect on the trading rate.

A larger tick size translates into larger gains from trade for market-makers. Thus, other things being equal, an increase in the tick size is conducive to more monitoring by market-makers. Hence, market-takers (i) obtain less surplus per transaction but (ii) expect more frequent trading opportunities when the tick size is larger. In equilibrium, the first effect dominates. Thus, an increase in the tick size enlarges market-makers' monitoring intensity, but it decreases market-takers' monitoring intensities. For this reason, the effect of a change in the tick size on the trading rate is not monotonic, and the trading rate is maximal for a strictly positive tick size. There are very few empirical studies that consider the effect of the tick size on the trading rate. Chakravarty et al. (2004) find a significant drop in the trading frequency for all trade sizes categories after the implementation of decimal pricing on the NYSE.

Interestingly, the trading platform fully neutralizes the effect of a change in the tick size on the trading rate through the choice of its trading fees (second part of the corollary). Thus, parts 1 and 2 of the corollary jointly suggest that the short run and long run effects (after adjustment of fees) of a change in the tick size are different. In the short run, a change in tick size should affect the trading rate whereas in the long run, after the adjustment of trading fees, the effect should disappear (if the fees are set optimally).

Trading Volume, Algorithmic Trading, and Trading Fees. Trading volume has considerably increased in the recent years. For instance, from 2005 to 2007, the number of shares traded on the NYSE rose by 111%, despite the loss in market share of the NYSE over the same period. The same trend is observed in other markets (e.g., the trading volume on the LSE increased by 69% in 2007).

In reality, average trading volume is the average trading rate multiplied by the average order size. Since all orders in our model have the same size, the trading rate proxies for volume. Thus, the model suggests two possible causes for the evolution in volume: (i) the development of algorithmic trading, and (ii) the evolution of the pricing policy used by trading platforms.[24] Indeed, as shown by Corollary 1, a decrease in the monitoring cost for the market-making side or the market-taking side triggers an increase in the trading rate. The result also holds when the fee structure is endogenous. Intuitively, a reduction in monitoring cost accelerates the speed at which market-takers and market-makers respond to each other and thereby results in more trades per unit of time. Second, the model implies that there is one split of trading fees between market-makers and market-takers that maximizes the trading rate. When the tick size is small, this split is such that market-makers are charged less than market-takers. Thus, the widespread adoption of a negative make-take spread should also enhance trading activity.

Bid-Ask Spread and Algorithmic Trading. Quoted bid-ask spreads are often used as a measure of liquidity. The (half) bid-ask spread (the best offer less v_0) is

[24] Of course, there might be other causes such as the development of institutional trading. See Chordia et al. (2008) for an empirical analysis of the evolution of the trading volume in U.S. equity markets and its determinants.

either a (in state F) or $a+\Delta$ (in state E).[25] During a cycle, the market is in state F for an average duration $\mathcal{D}_t$ and in state E for an average duration $\mathcal{D}_m$. Thus, the average half bid-ask spread (denoted ES) is:

$$ES = \theta a + (1-\theta)(a+\Delta) - v_0 = \frac{\Delta}{2} + (1-\theta)\Delta. \tag{30}$$

where

$$\theta \equiv \frac{\mathcal{D}_t}{\mathcal{D}_m + \mathcal{D}_t} = \frac{\mathcal{I}}{1+\mathcal{I}}. \tag{31}$$

Thus the average bid-ask spread decreases when liquidity supply increases relative to liquidity demand, that is when thee ratio $\frac{\bar{\lambda}}{\bar{\mu}}$ is large. An increase in $\bar{\mu}$ relative to $\bar{\lambda}$ means that liquidity demand pressure develops in the sense that it accelerates the speed at which the best offer is hit relative to the speed at which market-makers reinstate the best offer at the competitive pressure. In this case the bid-ask spread enlarges.

In equilibrium, $\frac{\bar{\lambda}}{\bar{\mu}}$ increases in the relative monitoring cost ratio, $r = \frac{\gamma}{\beta}$. For instance, in the large market, $\frac{\bar{\lambda}}{\bar{\mu}} = (\frac{\pi_m rq}{\pi_t})^{1/2}$ for given fees and $\frac{\bar{\lambda}}{\bar{\mu}} = (rq)^{2/3}$ when fees are set optimally. Hence, a decrease in β reduces the bid-ask spread whereas a decrease in γ enlarges the bid-ask spread. Thus, in considering the impact of algorithmic trading on the bid-ask spread, it is important to distinguish between algorithmic traders acting mainly as liquidity suppliers and algorithmic traders acting mainly as liquidity demanders. In the latter case, algorithmic trading increases price pressure by liquidity demanders and results in larger bid-ask spread on average.

Hendershott, Jones, and Menkveld (2009) consider a change in the organization of the NYSE that made algorithmic trading easier for liquidity suppliers. This is consistent with our model which implies that in this case the resulting increase in algorithmic trading should yield a smaller bid-ask spread. On the other hand, Hendershott and Moulton (2009) study an event in which monitoring and execution costs decreased for market-takers on NYSE. As predicted by our model, this resulted in an increase in in the bid-ask spread. Furthermore, our model offers an alternative explanation to this phenomenon, which is not related to changes in adverse-selection.

The model also implies that considering the effect of algorithmic trading on the trading rate is important. For instance when γ decreases, the bid-ask spread enlarges on average, a symptom of illiquidity. But this change in market-takers' monitoring

[25] Recall that a large number of shares is offered for sale at price $a+\Delta$ by a fringe of competitive traders.

cost makes all market participants better off, other things equal, as it results in a higher trading rate (Corollary 2). Thus, for fixed trading fees, the change in the trading rate is a better indicator of the impact of algorithmic trading on traders' welfare than the bid-ask spread. This is related to Boehmer (2005) who empirically shows a distinction between transaction costs and the speed at which transactions are executed.

6 Conclusion

This paper considers a model in which traders must monitor the market to seize trading opportunities. One group of traders ("market-makers") specializes in posting quotes while another group of traders ("market-takers") specializes in hitting quotes. Market-makers monitor the market to be the first to submit a new competitive quote after a transaction. Market-takers monitor the market to be the first to hit a competitive quote. In this way, we model the high frequency make/take liquidity cycles observed in electronic security markets. We show that the monitoring decisions of market-makers on the one hand and market-takers on the other hand are self-reinforcing. This feature has several implications. For instance, it implies that a trading platform can be trapped in a no trade equilibrium in which market-takers pay no attention to the platform because they expect market-makers to be inactive and vice versa. It also implies that the speed at which liquidity demanders respond to new quotes is positively related to the speed at which new quotes are posted on the platform and vice versa. This complementarity between liquidity demand and liquidity supply offers a new explanation for the clustering in the duration between trades.

In this set-up, we study the role of make/take fees. We show that these fees can be used by a trading platform to control traders' monitoring decisions and therefore the trading rate. In particular, it is optimal for the trading platform to reduce its fee on market-makers and increase its fee on market-takers when (i) the tick size decreases, (ii) the number of market-makers relative to the number of market-takers decrease or (iii) the monitoring cost of market-takers relative to the monitoring costs of market-makers increases.

We also use the model to study the effect of algorithmic trading (a drastic reduction in monitoring costs). The model implies that algorithmic trading should lead to a sharp increase in trading rate and has an ambiguous effect on the bid-ask spread. Interestingly, for fixed trading fees, we also find that algorithmic trading results in a Pareto improvement since it makes all market participants (including the trading platform) better off.

7 Appendix

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

Proof of Proposition 2: From (8) and (4), the first order condition for market-maker i is:

$$\frac{\bar{\mu}\left(\bar{\mu}+\bar{\lambda}-\lambda_i\right)}{\left(\bar{\lambda}+\bar{\mu}\right)^2}\frac{\pi_m}{\beta}=\lambda_i. \tag{32}$$

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

$$\frac{\bar{\mu}\left(\left(\bar{\mu}+\bar{\lambda}\right)M-\bar{\lambda}\right)}{\left(\bar{\lambda}+\bar{\mu}\right)^2}\frac{\pi_m}{\beta}=\bar{\lambda}. \tag{33}$$

Similarly, for market-takers we obtain,

$$\frac{\bar{\lambda}\left(\left(\bar{\mu}+\bar{\lambda}\right)N-\bar{\mu}\right)}{\left(\bar{\lambda}+\bar{\mu}\right)^2}\frac{\pi_t}{\gamma}=\bar{\mu}. \tag{34}$$

Let $\Omega\equiv\frac{\bar{\lambda}}{\bar{\mu}}$. Dividing the left-hand-side of (33) and (34) by $\bar{\mu}^2$ we obtain,

$$\frac{M+(M-1)\,\Omega}{(1+\Omega)^2}\frac{\pi_m}{\beta} \quad = \quad \bar{\lambda}. \tag{35}$$

$$\frac{\Omega\left((1+\Omega)\,N-1\right)}{(1+\Omega)^2}\frac{\pi_t}{\gamma} \quad = \quad \bar{\mu} \tag{36}$$

Dividing these two equations gives,

$$\frac{\left(M+(M-1)\,\Omega\right)}{\Omega^2\left((1+\Omega)\,N-1\right)}z=1, \tag{37}$$

or equivalently,

$$\Omega^3N+(N-1)\Omega^2-(M-1)\,z\Omega-Mz=0.$$

We argue that this cubic equation has a unique positive solution. Indeed, this equation is equivalent to

$$\Omega=g(\Omega,M,N,z). \tag{38}$$

with

$$g(\Omega, M, N, z) = \frac{(M-1)z}{\Omega N} + \frac{Mz}{N\Omega^2} - \frac{N-1}{N}. \tag{39}$$

Function $g(\cdot, M, N, z)$ decreases in Ω. It tends to plus infinity as Ω goes to zero, and to $-\frac{N-1}{N}$ as Ω goes to infinity. Thus, (38) has a unique positive solution that we denote by Ω^*.

To obtain a full characterization of the aggregate monitoring levels in equilibrium, insert this root into Equations (35) and (36).

To obtain traders' individual monitoring levels note that the equilibrium trading strategies are symmetric among the market-makers and market-takers. That is, $\lambda_1 = \lambda_2 = \ldots = \lambda_M$ and $\mu_1 = \mu_2 = \ldots = \mu_N$.[26] Hence, the individual equilibrium monitoring levels are obtained from $\lambda_i = \bar{\lambda}/M$ and $\mu_j = \bar{\mu}/N$ for all i, j. This completes the proof. ■

Proof of Corollary 1: Recall that Ω^* is such that:

$$\Omega^* = g(\Omega^*, M, N, z), \tag{40}$$

where $g(\cdot)$ is defined in equation (39). It is immediate that $g(\cdot)$ increases in M, decreases in N, and increases in z. As $g(\cdot)$ decreases in Ω, we have

$$\frac{\partial \Omega^*}{\partial M} > 0, \tag{41}$$

$$\frac{\partial \Omega^*}{\partial N} < 0. \tag{42}$$

Now, using Equations (41) and (11), we conclude that:

$$\frac{\partial \lambda_i^*}{\partial N} = \frac{-\frac{\partial \Omega^*}{\partial N} \cdot ((M+1) + (M-1)\Omega^*)}{(1+\Omega^*)^3} \left(\frac{\pi_m}{M\beta}\right) > 0.$$

Hence, $\frac{\partial \bar{\lambda}^*}{\partial M} > 0$. Similarly, using equations (42) and (12), we deduce that

$$\frac{\partial \mu_j^*}{\partial M} > 0. \tag{43}$$

[26] Indeed, suppose for example that $\lambda_1 > \lambda_2$. Then, from (32),

$$\frac{\bar{\mu}\left(\bar{\mu} + \bar{\lambda} - \lambda_1\right)}{\left(\bar{\lambda} + \bar{\mu}\right)^2} \frac{\pi_m}{\beta} > \frac{\bar{\mu}\left(\bar{\mu} + \bar{\lambda} - \lambda_2\right)}{\left(\bar{\lambda} + \bar{\mu}\right)^2} \frac{\pi_m}{\beta},$$

which simplifies to $\lambda_1 < \lambda_2$ - a contradiction.

Hence, $\frac{\partial \bar{\mu}^*}{\partial M} > 0$. We also have

$$\Omega^* = \frac{\bar{\lambda}^*}{\bar{\mu}^*}.$$

Thus, using equations (41) and (42), we conclude that $\frac{\bar{\lambda}^*}{\bar{\mu}^*}$ increases in M and decreases in N. Equation (43) implies that $\bar{\mu}^*$ increases in M. Thus it must be the case that $\bar{\lambda}^*$ increases in M as well. A similar argument shows that $\bar{\mu}^*$ increases in N.

Now, consider the effect of a change in β on market-takers' monitoring intensities. We have (see Proposition 2),

$$\mu_j^* = \zeta(\Omega^*) \left(\frac{\pi_t}{N\gamma} \right),$$

where

$$\zeta(\Omega^*) = \left(\frac{\Omega^* \left((1 + \Omega^*) N - 1 \right)}{(1 + \Omega^*)^2} \right).$$

Thus

$$\frac{\partial \mu_j^*}{\partial \beta} = \left(\frac{\partial \zeta(\Omega^*)}{\partial \Omega^*} \frac{\partial \Omega^*}{\partial z} \frac{\partial z}{\partial \beta} \right) \left(\frac{\pi_t}{N\gamma} \right)$$

We have $\frac{\partial \zeta(\Omega^*)}{\partial \Omega^*} > 0$. Moreover $\frac{\partial \Omega^*}{\partial z} > 0$ and $\frac{\partial z}{\partial \beta} < 0$. Thus

$$\frac{\partial \mu_j^*}{\partial \beta} < 0,$$

which implies that $\frac{\partial \overline{\mu}^*}{\partial \beta} < 0$. Now, since $\bar{\lambda}^* = \Omega^* \bar{\mu}^*$, we have:

$$\frac{\partial \bar{\lambda}^*}{\partial \beta} = \Omega^* \frac{\partial \overline{\mu}^*}{\partial \beta} + \frac{\partial \Omega^*}{\partial z} \frac{\partial z}{\partial \beta} \bar{\mu}^* < 0,$$

which implies $\frac{\partial \lambda_j^*}{\partial \beta} < 0$. The impact of other parameters on the aggregate monitoring levels of the market-makers and the market-takers is obtained in the same way. The second part of the corollary directly follows from the first part. ■

Proof of Corollary 2: Consider first the aggregate expected profit for market-takers. We have:

$$\Pi_t(\mu_1^*, .., \mu_j^*, ..., \mu_N^*, \bar{\lambda}^*; \gamma, \beta, c_m, c_t) = \textstyle\sum_j \Pi_{jt}(\mu_j^*, \bar{\lambda}^*; \gamma, \beta, M, N)$$

Thus,

$$\begin{aligned} \frac{d\Pi_t}{d\gamma} &= \textstyle\sum_j \left(\frac{\partial \Pi_{jt}}{\partial \mu_j^*} \frac{\partial \mu_j^*}{\partial \gamma} + \frac{\partial \Pi_{jt}}{\partial \bar{\lambda}^*} \frac{\partial \bar{\lambda}^*}{\partial \gamma} + \frac{\partial \Pi_{jt}}{\partial \gamma} \right) \\ \frac{d\Pi_t}{d\beta} &= \textstyle\sum_j \left(\frac{\partial \Pi_{jt}}{\partial \mu_j^*} \frac{\partial \mu_j^*}{\partial \beta} + \frac{\partial \Pi_{jt}}{\partial \bar{\lambda}^*} \frac{\partial \bar{\lambda}^*}{\partial \beta} + \frac{\partial \Pi_{jt}}{\partial \beta} \right) \end{aligned}$$

Now, the envelope theorem implies that $\frac{\partial \Pi_{jt}}{\partial \mu_j^*} = 0$ for all j. Moreover, the cross-side complementarity implies $\frac{\partial \Pi_{jt}}{\partial \bar{\lambda}^*} > 0$ for all j, and Corollary 1 yields $\frac{\partial \bar{\lambda}^*}{\partial \gamma} < 0$ and $\frac{\partial \bar{\lambda}^*}{\partial \beta} < 0$. Last, for all j, $\frac{\partial \Pi_{jt}}{\partial \gamma} = -\frac{1}{2}\left(\mu_j^*\right)^2 < 0$ and $\frac{\partial \Pi_{jt}}{\partial \beta} = 0$. Thus, $\frac{d\Pi_t}{d\gamma} < 0$ and $\frac{d\Pi_t}{d\beta} < 0$. This establishes the first part of the proposition for the market-taking side. The proof for the market-makers is parallel. Last, we have proved in Corollary 1 that the trading rate decreases when β or γ increases. It follows that the expected profit of the platform decreases with β or γ. Thus, the first part of the proposition is proved.

For the second part of the proposition, observe that

$$\begin{aligned}
\frac{d\Pi_t}{dc_t} &= \textstyle\sum_j \left(\frac{\partial \Pi_{jt}}{\partial \mu_j^*} \frac{\partial \mu_j^*}{\partial c_t} + \frac{\partial \Pi_{jt}}{\partial \bar{\lambda}^*} \frac{\partial \bar{\lambda}^*}{\partial c_t} + \frac{\partial \Pi_{jt}}{\partial c_t} \right) \\
\frac{d\Pi_m}{dc_t} &= \textstyle\sum_i \left(\frac{\partial \Pi_{im}}{\partial \lambda_i^*} \frac{\partial \lambda_i^*}{\partial c_t} + \frac{\partial \Pi_{im}}{\partial \bar{\mu}^*} \frac{\partial \bar{\mu}^*}{\partial c_t} \right) \\
\frac{d\Pi_e}{dc_t} &= \mathcal{R}(\overline{\lambda}^*, \overline{\mu}^*) + \frac{\partial \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_m} \overline{c}
\end{aligned}$$

The envelope theorem implies that $\frac{\partial \Pi_{jt}}{\partial \mu_j^*} = 0$ and $\frac{\partial \Pi_{im}}{\partial \lambda_i^*} = 0$ for all i and j. Moreover, $\frac{\partial \Pi_{jt}}{\partial c_t} = -\frac{\mu_j^* \bar{\lambda}^*}{\bar{\lambda}^* + \bar{\mu}^*}$ and $\frac{\partial \Pi_{jt}}{\partial \bar{\lambda}^*} = \frac{\mu_j^* \bar{\mu}^*}{(\bar{\lambda}^* + \bar{\mu}^*)^2} \pi_t$. Last, $\frac{\partial \Pi_{im}}{\partial \bar{\mu}^*} = \frac{\lambda_i^* \bar{\lambda}^*}{(\bar{\lambda}^* + \bar{\mu}^*)^2}$. Thus,

$$\frac{d\Pi_t}{dc_t} = \frac{\bar{\mu}^*}{(\bar{\lambda}^* + \bar{\mu}^*)^2} \pi_t \frac{\partial \bar{\lambda}^*}{\partial c_t} - \frac{\bar{\mu}^* \bar{\lambda}^*}{\bar{\lambda}^* + \bar{\mu}^*}, \tag{44}$$

and

$$\frac{d\Pi_m}{dc_t} = \frac{\bar{\lambda}^{*2}}{(\bar{\lambda}^* + \bar{\mu}^*)^2} \frac{\partial \bar{\mu}^*}{\partial c_t}. \tag{45}$$

Moreover,

$$\frac{d\Pi_e}{dc_t} = \frac{\bar{\mu}^* \bar{\lambda}^*}{\bar{\lambda}^* + \bar{\mu}^*} + \frac{\bar{\mu}^{*2}}{(\bar{\lambda}^* + \bar{\mu}^*)^2} \left(\frac{\partial \bar{\lambda}^*}{\partial c_t} \left(\frac{1}{\bar{\lambda}^*} \right)^2 + \frac{\partial \bar{\mu}^*}{\partial c_t} \left(\frac{1}{\bar{\mu}^*} \right)^2 \right). \tag{46}$$

Adding up (44), (45), and (46), we obtain,

$$\begin{aligned}
\frac{dW}{dc_t} &= \frac{\bar{\mu}^*}{(\bar{\lambda}^* + \bar{\mu}^*)^2} \pi_t \frac{\partial \bar{\lambda}^*}{\partial c_t} + \frac{\bar{\lambda}^{*2}}{(\bar{\lambda}^* + \bar{\mu}^*)^2} \frac{\partial \bar{\mu}^*}{\partial c_t} \\
&\quad + \frac{\bar{\mu}^{*2}}{(\bar{\lambda}^* + \bar{\mu}^*)^2} \left(\frac{\partial \bar{\lambda}^*}{\partial c_t} \left(\frac{1}{\bar{\lambda}^*} \right)^2 + \frac{\partial \bar{\mu}^*}{\partial c_t} \left(\frac{1}{\bar{\mu}^*} \right)^2 \right),
\end{aligned}$$

which is negative since $\frac{\partial \bar{\mu}^*}{\partial c_t} < 0$ and $\frac{\partial \bar{\lambda}^*}{\partial c_t} < 0$. A similar argument applies to changes in c_m. ■

Proof of Corollary 3: Using equation (13), it is readily checked that $\Omega^* = 1$ if and only if $z = \frac{2N-1}{2M-1}$. Thus, $\bar{\lambda}^* = \bar{\mu}^*$ if and only if $z = \frac{2N-1}{2M-1}$. Moreover, as shown in the proof of Corollary **??**, Ω^* increases in z. Hence, $\bar{\lambda}^* > \bar{\mu}^*$ iff $z > \frac{2N-1}{2M-1}$. ■

Proof of Corollary 5: Using Proposition (2),

$$\begin{aligned}
\lambda_i^{\infty} &\equiv \lim_{M\to\infty} \lambda_i^* = \lim_{M\to\infty} \left(\frac{M + (M-1)\,\Omega^*}{M\,(1+\Omega^*)^2} \right) \left(\frac{\pi_m}{\beta} \right) = \lim_{M\to\infty} \left(\frac{1 + \frac{M-1}{M}\Omega^*}{(1+\Omega^*)^2} \right) \left(\frac{\pi_m}{\beta} \right) \\
&= \frac{1}{1+\Omega^{\infty}} \left(\frac{\pi_m}{\beta} \right) = \frac{1}{1+(zq)^{\frac{1}{2}}} \frac{\pi_m}{\beta} \qquad \text{(using (16))}.
\end{aligned}$$

A similar argument is used to derive μ_j^{∞}. ■

Proof of Proposition 3: We have

$$\begin{aligned}
\frac{\partial \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_m} &= \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)^2 \left(\frac{\partial \bar{\lambda}^*}{\partial c_m} \frac{1}{\bar{\lambda}^{*2}} + \frac{\partial \bar{\mu}^*}{\partial c_m} \frac{1}{\bar{\mu}^{*2}} \right) \\
&= \frac{\mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)^2}{c_m} \left(\frac{\eta_{mm}}{\bar{\lambda}^*} + \frac{\eta_{tm}}{\bar{\mu}^*} \right). \qquad (47)
\end{aligned}$$

and

$$\frac{\partial \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_t} = \frac{\mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)^2}{c_t} \left(\frac{\eta_{mt}}{\bar{\lambda}^*} + \frac{\eta_{tt}}{\bar{\mu}^*} \right). \qquad (48)$$

The optimal fee structure is such that

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

Thus, using equations (47) and (48), we deduce that

$$\frac{(\bar{\lambda}^*)^{-1}\eta_{mm} + (\bar{\mu}^*)^{-1}\eta_{tm}}{(\bar{\lambda}^*)^{-1}\eta_{mt} + (\bar{\mu}^*)^{-1}\eta_{tt}} = \frac{c_m}{c_t}.$$

Now, (21) follows directly from this equation and the fact that $c_m + c_t = \bar{c}$. ■

Proof of Proposition 4: 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 trading platform and the per trade trading profits obtained by the market-making side and the market-taking side, 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. Thus, for a fixed $\bar{c}$, we rewrite the platform's problem as:

$$\begin{aligned}
&Max_{\pi_m,\pi_t} \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*) \\
&s.t \;\; \pi_t + \pi_m = L - \bar{c}.
\end{aligned}$$

Moreover, using that $\pi_m = L - \bar{c} - \pi_t$, we can present $\mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)$ as a function of π_t only. We know that

$$\mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*) = \frac{\bar{\lambda}^* \bar{\mu}^*}{\bar{\lambda}^* + \bar{\mu}^*} = \frac{\bar{\lambda}^*}{1 + \Omega^*}. \tag{49}$$

The first order condition with respect to π_t gives

$$\frac{\partial \bar{\lambda}^*}{\partial \pi_t}(1 + \Omega^*) - \frac{\partial \Omega^*}{\partial \pi_t}\bar{\lambda}^* = 0,$$

or equivalently,

$$\frac{\partial \bar{\lambda}^*}{\partial \pi_t} = \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)\frac{\partial \Omega^*}{\partial \pi_t}.$$

Since $\bar{\lambda}^* = M\lambda_1^*$, we can divided both sides by M and obtain

$$\frac{\partial \lambda_1^*}{\partial \pi_t} = \frac{\mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{M}\frac{\partial \Omega^*}{\partial \pi_t}. \tag{50}$$

Since the first order condition holds for any M, we can take limits on both sides to obtain a necessary condition for the large market:

$$\lim_{M\to\infty} \frac{\partial \lambda_1^*}{\partial \pi_t} = \lim_{M\to\infty} \frac{\mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{M} \cdot \lim_{M\to\infty} \frac{\partial \Omega^*}{\partial \pi_t}. \tag{51}$$

Straightforward calculations show that

$$\begin{aligned} \lim_{M\to\infty} \frac{\partial \Omega^*}{\partial \pi_t} &= \frac{d\Omega^\infty}{d\pi_t} = -\frac{q}{2\Omega^\infty}\frac{L-c}{\pi_t^2}\frac{\gamma}{\beta}, \text{ and} \\ \lim_{M\to\infty} \frac{d\lambda_1^*}{d\pi_t} &= -\frac{1}{\beta\,(1+\Omega^\infty)} - \frac{1}{(1+\Omega^\infty)^2} \cdot \frac{d\Omega^\infty}{d\pi_t} \cdot \frac{L - \bar{c} - \pi_t}{\beta}, \end{aligned}$$

where Ω^∞ is given by (16). Furthermore, it is direct from (49) that

$$\lim_{M\to\infty} \frac{\mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{M} = \frac{\lambda_1^\infty}{1 + \Omega^\infty},$$

where λ_1^∞ is given by (17). Plugging back into (51) and simplifying yields

$$1 - \frac{q\,(L - \bar{c} - \pi_t)}{\Omega^\infty\,(1+\Omega^\infty)} \cdot \frac{L - \bar{c}}{\pi_t^2} \cdot \frac{\gamma}{\beta} = 0,$$

or,

$$1 - \frac{zq}{(1+\Omega^\infty)\,\Omega^\infty}\frac{L-\bar{c}}{\pi_t} = 0, \tag{52}$$

which simplifies to

$$\frac{\pi_t}{L - \bar{c}} = \frac{\Omega^\infty}{1 + \Omega^\infty}. \tag{53}$$

Denote

$$w \equiv \frac{\pi_t}{L-\bar{c}}. \tag{54}$$

Then (53) imposes

$$w = \frac{\Omega^\infty}{1+\Omega^\infty} = \frac{1}{1+(zq)^{\frac{1}{2}}}. \tag{55}$$

Now, observe that

$$z = r\frac{1-w}{w}.$$

Thus, we can rewrite equation (55) as

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

It is immediate that this equation has a unique solution:

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

From (54) we obtain,

$$\pi_t = \frac{L-\bar{c}}{1+(rq)^{-\frac{1}{3}}}, \tag{56}$$

and,

$$\pi_m = L-\bar{c}-\pi_t = \frac{L-\bar{c}}{1+(rq)^{\frac{1}{3}}}. \tag{57}$$

Given this,

$$z = \frac{\frac{L-\bar{c}}{1+(qr)^{\frac{1}{3}}}}{\frac{L-\bar{c}}{1+(qr)^{-\frac{1}{3}}}}r = q^{-\frac{1}{3}}r^{\frac{2}{3}}.$$

And,

$$\Omega^\infty = (zq)^{\frac{1}{2}} = (rq)^{\frac{1}{3}}\,. \tag{58}$$

The optimal fees in the large market are:

$$\begin{aligned} c_m &= \frac{\Delta}{2} - \pi_m = \frac{\Delta}{2} - \frac{L-\bar{c}}{1+(qr)^{\frac{1}{3}}} \\ c_t &= L - \frac{\Delta}{2} - \pi_t = L - \frac{\Delta}{2} - \frac{L-\bar{c}}{1+(qr)^{-\frac{1}{3}}}. \end{aligned}$$

Finally, the monitoring intensities in the large market are obtained by plugging these expressions into Corollary 5. ■

Proof of Corollary 6: The result follows directly from equation (22) ■

Proof of Corollary 7: We fix $q > 0$, and let $N = \frac{M}{q}$.For any given M, maximizing $\mathcal{R}(\lambda^*, \mu^*)\bar{c}$ is equivalent to maximizing $\frac{\mathcal{R}(\lambda^*,\mu^*)}{M}\bar{c}$, which in turn (using (49) and that $\bar{\lambda}^* = M\lambda_1^*$) is equivalent to maximizing $\frac{\lambda_1^*}{1+\Omega^*}\bar{c}$. Denote $\mathcal{H}(\bar{c}) \equiv \frac{\lambda_1^*}{1+\Omega^*}$. Then, to find the optimal total fee $\bar{c}$ in the large market case we need to find the limit as M tends to infinity of

$$\arg\max_{\bar{c}\geq 0} \mathcal{H}(\bar{c})\,\bar{c}.$$

The FOC for a given M is

$$\mathcal{H}(\bar{c}) + \mathcal{H}'(\bar{c})\,\bar{c} = 0. \tag{59}$$

Note that $\mathcal{H}$ depends on $\bar{c}$ only through its dependence on λ_1^* and Ω^*. It follows that

$$\mathcal{H}'(\bar{c}) = \frac{\partial\mathcal{H}}{\partial\lambda_1^*}\frac{\partial\lambda_1^*}{\partial\bar{c}} + \frac{\partial\mathcal{H}}{\partial\Omega^*}\frac{\partial\Omega^*}{\partial\bar{c}} = \frac{1}{1+\Omega^*}\frac{\partial\lambda_1^*}{\partial\bar{c}} - \frac{\lambda_1^*}{(1+\Omega^*)^2}\frac{\partial\Omega^*}{\partial\bar{c}}. \tag{60}$$

Since (59) holds for any M, we can take the limit as $M \to \infty$. We have,

$$\lim_{M\to\infty} \mathcal{H}(\bar{c}) = \frac{\lambda_1^\infty}{1+\Omega^\infty} = \frac{L-\bar{c}}{\beta\left(1+(qr)^{\frac{1}{3}}\right)^3} \quad \text{(using (24) and (58))}.$$

It can also be verified using (24) and (58) that

$$\begin{aligned}\lim_{M\to\infty}\frac{\partial\lambda_1^*}{\partial\bar{c}} &= \frac{\partial\lambda_1^\infty}{\partial\bar{c}} = -\frac{1}{\beta\left(1+(qr)^{\frac{1}{3}}\right)^2}, \text{ and}\\ \lim_{M\to\infty}\frac{\partial\Omega^*}{\partial\bar{c}} &= \lim_{M\to\infty}\frac{\partial\Omega^\infty}{\partial\bar{c}} = 0.\end{aligned}$$

Thus, from (60),

$$\lim_{M\to\infty}\mathcal{H}'(\bar{c}) = -\frac{1}{\beta\left(1+(qr)^{\frac{1}{3}}\right)^3}.$$

And, in the limit (59) becomes

$$\frac{L-\bar{c}}{\beta\left(1+(qr)^{\frac{1}{3}}\right)^3} - \frac{1}{\beta\left(1+(qr)^{\frac{1}{3}}\right)^3}\bar{c} = 0,$$

which gives $\bar{c} = \frac{L}{2}$. ■

Proof of Proposition 5: As in the proof of Proposition 4, we can use π_{mm} and π_{mt}. Thus, for a fixed $\bar{c}$, when $M = N = 1$, the platform problem is:

$$\begin{gathered}Max_{\pi_m,\pi_t}\ \frac{\lambda_1^*\mu_1^*}{\lambda_1^*+\mu_1^*}\bar{c}\\ s.t\ \ \pi_t + \pi_m = L - \bar{c}.\end{gathered}$$

From equations (35) and (36),

$$\frac{\lambda_1^*}{\mu_1^*} = z^{\frac{1}{3}} = (\frac{\pi_m}{\pi_t}\frac{\gamma}{\beta})^{\frac{1}{3}}$$

and

$$\lambda_1^* = \frac{\pi_m}{\beta}\frac{1}{\left(1+z^{\frac{1}{3}}\right)^2}$$

Thus, we can rewrite the previous optimization problem as:

$$Max_{\pi_m,z}\frac{\lambda_1^*}{1+z^{\frac{1}{3}}}\bar{c} \tag{61}$$

$$s.t \quad \pi_m\left(1+\frac{\gamma}{\beta z}\right) = L-\bar{c}. \tag{62}$$

$$\text{and } \lambda_1^* = \frac{L-\bar{c}}{\beta\left(1+z^{\frac{1}{3}}\right)^2\left(1+\frac{\gamma}{\beta z}\right)} \tag{63}$$

This problem is equivalent to finding z that minimizes

$$\left(1+z^{\frac{1}{3}}\right)^3\left(\beta+\frac{\gamma}{z}\right).$$

The first order condition to this problem imposes

$$-\frac{1}{z^2}\left(\gamma - z^{\frac{4}{3}}\beta\right)\left(z^{\frac{1}{3}}+1\right)^2 = 0.$$

Hence, the solution is

$$z = \left(\frac{\gamma}{\beta}\right)^{\frac{3}{4}} = r^{\frac{3}{4}}. \tag{64}$$

Using the constraint (62), we have,

$$\pi_m^* = \frac{L-\bar{c}}{1+r^{\frac{1}{4}}}. \tag{65}$$

It follows that,

$$\pi_t^* = L-\bar{c}-\pi_m = \frac{L-\bar{c}}{1+r^{-\frac{1}{4}}}. \tag{66}$$

Then, plugging (64), (65), and (66) into equations (14) and (15), we obtain the required expressions for λ_1^* and μ_1^*. ■

Proof of Corollary 9: Define $\hat{c}_m = \frac{L}{2} - \frac{\Delta}{2} + c_m$ and $\hat{c}_t = \frac{\Delta}{2} - \frac{L}{2} + c_t$. Observe that we can write market-makers' and market-takers' payoffs as:

$$\begin{aligned}\Pi_{im} &= \frac{\lambda_i\bar{\mu}\left(\frac{\Delta}{2}-c_m\right)}{\bar{\lambda}+\bar{\mu}} - \frac{1}{2}\beta\lambda_i^2 = \frac{\lambda_i\bar{\mu}\left(\frac{L}{2}-\hat{c}_m\right)}{\bar{\lambda}+\bar{\mu}} - \frac{1}{2}\beta\lambda_i^2 \\ \Pi_{jt} &= \frac{\mu_j\bar{\lambda}\left(L-\frac{\Delta}{2}-c_t\right)}{\bar{\lambda}+\bar{\mu}} - \frac{1}{2}\beta\mu_j^2 = \frac{\mu_j\bar{\lambda}\left(\frac{L}{2}-\hat{c}_t\right)}{\bar{\lambda}+\bar{\mu}} - \frac{1}{2}\beta\mu_j^2\end{aligned}$$

These payoffs are those obtained when $\Delta = L$ and fees are set at $\hat{c}_m$ and $\hat{c}_t$. Thus, the values of $\hat{c}_m$ and $\hat{c}_t$ that maximize the trading rate are:

$$\begin{aligned} \hat{c}^* &= c_m^*(L, q, r) \\ \hat{c}_t &= c_t^*(L, q, r). \end{aligned}$$

Observe that $\hat{c}_m^*$ and $\hat{c}_t^*$ do not depend on the tick size. Thus, when the platform optimally chooses its trading fees, it does so that eventually traders' payoffs do not depend on the tick size. Thus, in this case, the maximal trading rate does not depend on the tick size, which proves the second part of the corollary.

For arbitrary fees c_m and c_t, $\hat{c}_m^*$ and $\hat{c}_t^*$ are obtained by choosing a tick size Δ^* such that $c_m^*(L, q, r) = \frac{L}{2} - \frac{\Delta^*}{2} + c_m$. This proves the first part of the corollary. ■

Proof of Corollary 8: By definition we have $\mathcal{I} = \frac{\overline{\lambda}}{\mu} = \Omega^*$. In equilibrium, in the large market case, $\Omega^* = \Omega^\infty = (zq)^{1/2}$ (see the proof of Lemma **??**). The proposition follows. ■

References

[1] Abel A. B, J.C. Eberly, and S. Panageas, 2009, Optimal Inattention to the Stock Market with Information Costs and Transactions Costs, working paper, University of Pennsilvanya.

[2] Admati, A.R., and Pfleiderer, (1988), "A Theory of Intraday Patterns : Volume and Price Variability", *The Review of Financial Studies*, 1, 3-40.

[3] Barclay M. J., and T. Hendershott. 2004, Liquidity externalities and adverse selection: evidence from trading after hours, Journal of Finance 59, 681-710.

[4] Bertsimas, D. and Lo, A.(1998) "Optimal control of execution costs," *Journal of Financial Markets* 1, 1-50.

[5] Biais, B., Hillion, P., and Spatt, C. (1995) "An empirical analysis of the limit order book and the order flow in the Paris bourse". *Journal of Finance* 50, 1655-1689.

[6] Biais, B., Bisière, C. and Spatt (2002) "Imperfect competition in financial markets: Island vs. Nasdaq," Working Paper, Toulouse University.

[7] Biais, B. and Weill, P.O. (2008) "Algorithmic Trading and the Dynamics of the Order Book," Manuscript, Toulouse University, IDEI.

[8] Bloomfield, R., O'Hara, M., and Saar, G., (2005) "The "make or take" decision in an electronic market: Evidence on the evolution of liquidity", *Journal of Financial Economics* 75, 165-199.

[9] Boehmer, E., 2005, "Dimensions of execution quality: Recent evidence for US equity markets," *Journal of Financial Economics* 78, 553–582.

[10] "Decimals and liquidity: a study of the NYSE," Journal of Financial Research,

[11] Chordia, T., Roll, R. and Subrahmanyam, A. (2008) "Why has trading volume increased," Working paper, UCLA.

[12] Coopejans, M, Domowitz, I. and Madhavan A. (2001) "Liquidity in an automated auction," Working paper, ITG.

[13] Corwin, S. and Coughenour, J.(2008), "Limited attention and the allocation of effort in securities trading," forthcoming in *Journal of Finance*.

[14] Degryse, H., De Jong F., Van Rvenswaaij, M. and Wuyts, G.(2005), "Aggressive orders and the resiliency of a limit order market," *Review of Finance*, 9, 201-242.

[15] Dow, J., (2005). "Self-sustaining liquidity in an asset market with asymmetric information." *Journal of Business* 78,

[16] Duffie, D, Garlenau, N. and Pedersen, L.H (2005) "Over-the-counter markets," *Econometrica* 73, 1815-1847.

[17] Engle, R.F. and J.R. Russell (1998), "Autoregressive conditional duration: a new model for irregularly spaced transaction data," *Econometrica*, 66, 1127-1162.

[18] Foucault, T., Roëll, A., and Sandås, P. (2003) "Market Making With Costly Monitoring: An Analysis of SOES Trading", *Review of Financial Studies* 16, 345-384.

[19] Foucault, T., Kadan, O., and Kandel, E. (2005) "Limit order book as a market for liquidity," *Review of Financial Studies*, 18, 1171-1217.

[20] Glosten, L. R. (1994) “Is the electronic open limit order book inevitable?” *Journal of Finance* 49, 1127-1161.

[21] Goldstein, M. and Kavajecz, K. (2000), “Eighths, Sixteenths and Market Depth: Changes in Tick Size and Liquidity Provision on the NYSE,” *Journal of Financial Economics* 56, 125-149

[22] Hasbrouck (1999) “Trading fast and slow: security markets in real time,” mimeo, NYU.

[23] Hall, A. and Hautsch, N. (2007) “Modelling the buy and sell intensity in a limit order book market,” *Journal of Financial Markets* 10, 249-286.

[24] Hendershott, T., Jones, C. and Menkveld, A. (2009) “Does algorithmic trading improve liquidity,” mimeo, U.C. Berkeley.

[25] Hendershott, T., and P. C. Moulton, 2009, Speed and Stock Market Quality: The NYSE’s Hybrid, Working Paper, University of California, Berkley.

[26] Hollifield, B., Miller, R. A., Sandas, P. (2004) “Empirical analysis of limit order markets”. *Review of Economic Studies* 71, 1027-1063.

[27] Huang, L., and H. Liu, 2007, Rational Inattention and Portfolio Selection, *Journal of Finance* 62, 1999-2040.

[28] Iliev P, and I. Welch, 2008, A Model of Operational Slack: The Short-Run, Medium-Run, and Long-Run Consequences of Limited Attention, working paper, Penn-State University and Brown University.

[29] Large, J. (2007), “Measuring the resiliency of an electronic limit order book,” *Journal of Financial Markets*, 1-25.

[30] Large, J. (2009), “A market clearing role for inefficiency on a limit order book,” *Journal of Financial Economics*, 102-117.

[31] Liu, W. (2007) “Monitoring and limit order submission risks”, Forthcoming *Journal of Financial Markets.*

[32] Pagano, M. (1989), “Trading volume and asset liquidity”, *Quarterly Journal of Economics*, 104, 255-276.

[33] Parlour, C. (1998), "Price dynamics in limit order markets," *Review of Financial Studies*, 11, 789-816.

[34] Rochet, JC and Tirole, J.(2006) "Two sided markets: a progress report," *Rand Journal of Economics, 3*7, 645-667.

[35] Rochet, JC and Tirole, J.(2003) "Platform competition in two sided markets," *Journal of the European Economic Association*, 1, 990-1029

[36] Ross, S. M., 1996, Stochastic Processes, John Wiley & Sons, Inc.

[37] Sandås, P., 2001, "Adverse selection and competitive market making: Empirical evidence from a limit order market". *Review of Financial Studies* 14, 705-734.

[38] Schack, J. and Gawronski, J. (2008) "History does not repeat itself, it rhymes: The coming revolution in European market structure," *The Journal of Trading*, Fall, 71-81.

[39] Seppi, D. J. (1997) "Liquidity provision with limit orders and a strategic specialist." *Review of Financial Studies* 10, 103-150.

[40] Sims C. A., 2003, "Implications of Rational Inattention", *Journal of Monetary Economics* 50, 665-690.