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

March 14, 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 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 cost for market-takers to monitoring cost 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: Make/Take Spread, Duration Clustering, Algorithmic trading, Two-Sided Markets.

*We thank participants at seminars at University of Mannheim and University College in Dublin for their useful comments.

1 Introduction

Trading in securities, 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]

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, several empirical studies emphasize the importance of monitoring to understand the dynamics of trades and quotes in limit order markets (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 speed at which make/take liquidity cycles are completed.

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.[2] Hence, 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 make/take liquidity cycle. In the process, it addresses two sets of related issues.

Firstly, in recent years, algorithmic trading has considerably decreased the cost

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

[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.

of monitoring and revolutionized 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 the distribution of trading gains among market participants.

Secondly, the model sheds light 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 providing liquidity pay a lower fee than investors taking liquidity.

	Tape A		**Tape B**		**Tape C**	
	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	-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 U.S. for different groups of stocks (Tapes A, B, C); A minus sign indicates that the fee is a rebate. Source: Traders Magazine, July 2008

For instance, Table 1 gives the make/take fees charged on liquidity makers and liquidity takers for a few U.S. equity trading platforms, as of July 2008. All these platforms subsidize liquidity makers by paying a rebate on 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 market-

[3]For instance, in each transaction, BATS (a trading platform for U.S stocks) 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, limit order traders involved in these transactions earned about $2 million on this day only.

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]

The attention of market participants to these fees suggests that they alter the market microstructure of securities markets. Yet, to the best of our knowledge, the rationale for the make/take spread and its impact on the trading process have not been analyzed. In this paper, we show that they play an important role for liquidity formation.

We distinguish two sides: (i) traders who post quotes (the "market-makers") and (ii) traders who hit these quotes (the "market-takers"). Both sides must monitor the market to grab fleeting trading opportunities. In choosing their monitoring intensity, traders on each side trade-off the benefit from a higher likelihood of detecting a profit opportunity with the cost of paying more attention to the trading process. In equilibrium, traders' monitoring choices determine the trading rate.

Monitoring decisions of 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. Thus, the initial shock on market-takers' monitoring is amplified, and triggers a snowballing effect on trading activity.

This 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 latter equilibrium,

[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

[6] It is well-known that the lack of coordination in traders' decision to participate in a market can lead to multiple equilibria with differing levels of liquidity (see Admati and Pfleiderer (1988), Pagano (1989), and Dow (2005) for example). In our setting, the multiplicity of equilibria also stems from

monitoring decisions depend on the factors that determine the cost and benefit of monitoring, namely (i) the monitoring cost of each side; (ii) the number of participants on each side; (iii) the tick size (the minimum price increment between two quotes); and (iv) trading fees

For fixed trading fees, a decrease in the monitoring cost on one side increases traders' monitoring *on both sides* because of the complementarity in monitoring decisions. Now, consider an increase in the number of market-makers. On the one hand, the probability that market-takers find good prices when they check the market becomes higher. Thus, they monitor more intensively which, through the snowballing effect we described previously, induces market-makers to monitor more, other things equal. But competition among market-makers reduces each one's market share. This second effect reduces market-makers' incentives to monitor. In our model, the first effect dominates in equilibrium so that the total monitoring intensity of both sides increases in the number of market-makers. As a result the trading rate increases when (i) the monitoring cost decreases or (ii) the number of participants on either side become larger.

A larger tick size translates into larger gains from trade for market-makers.[7] 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.

Next, we analyze the optimal breakdown of its total fee per trade between market-

a coordination problem, but between traders posting quotes on the one hand and traders hitting quotes on the other hand. This type of effect 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).

[7] This is a feature of several models of trading in financial markets (e.g., Glosten (1994) or Large (2008)).

makers and market-takers for the trading platform. We show that this breakdown is not neutral because it alters monitoring decisions of both sides and thereby the trading rate.

For instance, suppose that the tick size is very small. If the total trading fee is equally split between both sides (a zero make/take spread), then market-makers monitor the market less than market takers since they obtain a very small fraction of the gains from trade. Thus, trade opportunities are lost because the market frequently lacks good prices when it is checked by market-takers. In this sense, there is an *excess of attention* by market takers. In this situation, it is optimal for the trading platform to increase its fee on market-takers and reduce the fee charged on market-makers. This shift in the make/take spread helps to balance the monitoring intensities of both sides, and thereby the demand and supply of liquidity. Ultimately, it increases the trading rate.

Using this logic, we find that the optimal fee charged on market-makers (resp. market-takers) increases (resp. decreases) with (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. Importantly, these findings do not depend on the trading platform's market power since they hold for all levels of the total fee earned by the trading platform. Hence, the make/take spread should not per se be construed as a sign of imperfect competition between trading platforms.

Interestingly, in line with the model, the practice of subsidizing market-makers developed after the tick size was reduced to a penny in 2001 in U.S. equity markets. The recent decision of some options markets in the U.S. to adopt a make/take pricing structure also coincides with a reduction in the tick size of these markets.[8] According to our model, the subsidy of the market-making side could also reflect (i) a relatively small number of firms engaged in electronic market-making relative to the number of investors demanding liquidity; or/and (ii) a faster automation of their search for

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

liquidity by these investors.

Our analyses 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 platforms' fees.

Hendershott et al.(2008) find empirically that the development of algorithmic trading is associated with a reduction in bid-ask spreads and an increase in the trading rate. In line with their findings, our model implies that a decrease in monitoring cost can lead to a significant increase in trading rates (through the snowballing effect described previously). Thus, it implies a sharp increase in trading volume after upgrades in trading platforms facilitating algorithmic trading.[9]

Interestingly, the model also implies that asymmetries in the speed of automation of the market-making sector relative to the market-taking sector should affect the make-take spread and thereby the distribution of trading profits between these sectors. Indeed, a reduction in the cost of monitoring for market-takers shifts the division of the trading surplus in favor of market-makers (and vice versa). Indeed, the trading platform optimally reacts to a decrease in monitoring cost for market-takers by charging a larger fee on market-takers, and a smaller fee on market-makers. Thus, paradoxically, automation of the monitoring process by one side benefits to the other side after adjustment of trading fees.

Our analysis also contributes to the burgeoning literature on two-sided markets (see Rochet and 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).[10]

[9] In 2007, the trading volume on the London Stock Exchange (LSE) has increased by a stunning 69%. Market observers attribute this increase to upgrades in the LSE trading platform enabling algorithmic traders to get faster access to this platform.

[10] Rochet and Tirole (2006) provide several examples of two-sided markets. For instance videogames platforms, payment card systems etc...

The fee differential between liquidity providers and liquidity suppliers suggest that securities markets are two-sided markets and our model develops the implications of this feature.

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 5and Section 6 concludes. The proofs are in the Appendix.

2 Model

2.1 Market Participants

We consider a market for a security with two distinct sides: "market-makers" and "market-takers." Market-makers are those who post prices (limit orders) whereas market-takers are those who hit the quotes (submit market orders) to complete a transaction.[11] The number of market-makers and market-takers is, respectively, M and N.

In reality, traders can choose whether to post a quote or to hit a quote. Here we simplify the analysis by assuming that traders' roles are fixed. The market-making side can be viewed as electronic market-makers, such as Automated Trading Desk (ATD), Global Electronic Trading company (GETCO), Tradebots Systems, Citadel Derivatives, which specialize in high frequency market-making.[12] 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.[13] Electronic market-makers

[11] Trading platforms use various terminologies for designing each side. For instance, in limit order markets such as the Paris Bourse or the London Stock Exchange, traders submitting limit orders constitute the market-making side whereas traders submitting market orders constitute the market-taking side. Sometimes, the market-making and market-taking sides are designated respectively as the passive and active (or aggressive) side. See for instance Chi-X at http://www.chi-x.com/Cheaper.html

[12] 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 may be generating two-thirds or more of total daily volume today, dwarfing the activity of institutional investors.*"

[13] Bertsimas and Lo (1998) solve the dynamic optimization of such traders, assuming that they

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. Thus, market-makers and market-takers differ in their private values for the security. 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)).[14]

As market-takers have a higher valuation than market-makers, they buy the security from market-makers. In a more complex model, we could assume that market-takers have either high or low valuations relative to market-makers so that they can be buyers or sellers. This possibility adds 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 refuse to trade at a larger price on the grid (as $\frac{\Delta}{2} < L \leq \Delta$) and market-makers would 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). For the problem to be interesting, we assume that a fixed number of shares (normalized to one) can be profitably offered at price a. In a more complex model, this limit could follow for instance from exposure to informed trading as in Glosten (1994).[15]

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

exclusively use market orders as we do here.

[14] Hollifield et al.(2004) and Hollifield et al.(2006) show empirically that heterogeneity in traders' private values is needed to 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.

[15] 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 liquidity supply curve, as in Glosten (1994)'s model.

is denoted c_t. Thus, per transaction, the platform earns a revenue of

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

We assume that the cost of processing trades for the trading platform is zero. Introducing an order processing cost per trade is straightforward and does not change the results.

Thus, the gains for trade in each transaction (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 both traders are $L - \bar{c}$. We focus on the case $\bar{c} < L$ since otherwise traders on at least one side lose money on each trade, and would choose not to trade at all.

This 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 scarce: no offer is not posted at a.

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

We shall sometimes say that in State F the book is "full" whereas in State E the book is empty at a, or for brevity "empty." The market moves from state E to state F when a market-maker notices the profit opportunity and posts a quote at a. The market moves from state F back to state E when a market-taker notices the profit opportunity and hits the quote. Then, the process starts over again. 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."

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 $j = 1, ..., N$ chooses a monitoring intensity μ_j, which means that he inspects the market according to a Poisson process with parameter

μ_j.[16] 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 the book is empty (state E) or full (state F). If the book is empty the market-maker posts an offer at a, whereas if the book is full she stays put until her next inspection. Similarly, a market-taker submits a market order when he learns that the book is full, and stays put until the next inspection otherwise. Thus, market-makers compete against each other for seizing occasional profit opportunities reflected in empty books, and market-takers compete against each other for seizing profit opportunities reflected in full books. Market-makers and market-takers provide liquidity to one another as profitable trades can only be realized after an offer is hit by a market order.

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 have also a fixed computing capacity that 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

[16] Note that we restrict attention to stochastic monitoring policies. This rules out deterministic monitoring such as inspecting the market exactly once every certain number of minutes. The time interval between two inspections is random as 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 a similar constraints as periods of high transaction volume, and unexpectedly high traffic on communication lines prevent monitoring at exact points in time.

market-maker choosing a monitoring intensity λ_i bears a monitoring cost:

$$C_m(\lambda_i) \equiv \frac{1}{2}\beta\lambda_i^2 T \quad \text{for } i = 1, ..., M. \tag{3}$$

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

$$C_t(\mu_j) \equiv \frac{1}{2}\gamma\mu_j^2 T \quad \text{for } j = 1, ..., N. \tag{4}$$

Thus, the cost of monitoring is assumed to be 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 cost become lower when β (resp. γ) decreases. This can be a result, for example, of automation of the monitoring process or the decision to follow a more specialized strategy (i.e., to specialize in a few markets).

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. Thus, it is natural to assume that traders choose their monitoring intensities after observing the fees set by the trading platform. Thus, we assume that the trading game unfolds as follows:

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

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

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 is used 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 book 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 wins this transaction is the probability that she inspects an empty book before the other market-makers. Given our assumptions, this probability is $p_i \equiv \frac{\lambda_i}{\lambda_1+...+\lambda_M} = \frac{\lambda_i}{\bar{\lambda}}$. Thus, the expected profit (gross of monitoring costs) from a completed transaction for market-maker i is

$$p_i \pi_m = \frac{\lambda_i}{\bar{\lambda}} \left(\frac{\Delta}{2} - c_m \right). \tag{5}$$

Similarly, the probability that market-taker j wins the transaction in a specific cycle is $q_j \equiv \frac{\mu_j}{\bar{\mu}}$, and the expected profit per cycle is

$$q_j \pi_t = \frac{\mu_j}{\bar{\mu}} \left(L - \frac{\Delta}{2} - c_t \right). \tag{6}$$

Finally, the profit from a completed transaction for the trading platform is $\bar{c}$ for sure.

The average time it takes the book to move from state E to state F is $\frac{1}{\lambda_1+...+\lambda_M} = \frac{1}{\bar{\lambda}}$. Similarly, the average time from state F to state E is $\frac{1}{\mu_1+...+\mu_N} = \frac{1}{\bar{\mu}}$. It follows that the average duration of a cycle is

$$D \equiv \frac{1}{\bar{\lambda}} + \frac{1}{\bar{\mu}} = \frac{\bar{\lambda} + \bar{\mu}}{\bar{\lambda} \cdot \bar{\mu}}. \tag{7}$$

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

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

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 the objective function of each player is to maximize his/her long-term (steady-state) payoff per unit of time. That is, market-maker i seeks to maximize

$$\Pi_{im} \equiv \lim_{T\to\infty} \frac{\Pi_i(T)}{T} = \lim_{T\to\infty} \frac{E_{\tilde{n}_T}(\sum_{k=1}^{\tilde{n}_T} p_i \pi_m)}{T} - \frac{1}{2}\beta \lambda_i^2. \tag{8}$$

A standard theorem from the theory of stochastic processes (see Ross (1996), p. 133) implies that Π_{im} is equal to the expected payoff for market maker i per make/take cycle divided by the expected duration of a cycle. Thus, using equations (5) and (7), we can rewrite the objective function of market-maker i (equation (8)) as

$$\Pi_{im} = \frac{p_i \pi_m}{D} - \frac{1}{2}\beta\lambda_i^2 = \frac{\frac{\lambda_i}{\bar{\lambda}}\left(\frac{\Delta}{2} - c_m\right)}{\frac{\bar{\lambda}+\bar{\mu}}{\bar{\lambda}\cdot\bar{\mu}}} - \frac{1}{2}\beta\lambda_i^2 = \frac{\lambda_i\bar{\mu}\left(\frac{\Delta}{2} - c_m\right)}{\bar{\lambda}+\bar{\mu}} - \frac{1}{2}\beta\lambda_i^2. \qquad (9)$$

Similarly, the objective function of market-taker j is to maximize his expected payoff per cycle divided by the expected duration of a cycle,

$$\Pi_{jt} = \frac{q_j \pi_t}{D} - \frac{1}{2}\gamma\mu_j^2 = \frac{\mu_j\bar{\lambda}\left(L - \frac{\Delta}{2} - c_t\right)}{\bar{\lambda}+\bar{\mu}} - \frac{1}{2}\gamma\mu_j^2. \qquad (10)$$

From (9) and (10), other things being equal, the expected profit (gross of monitoring costs) of a trader on one side (e.g., the market-making side) declines in the monitoring intensities chosen by the traders on the same side. For instance, $\frac{\partial \Pi_{im}}{\partial \lambda_j} < 0$ (for $j \neq i$). Intuitively, this effect reflects the fact that traders on the same side compete for the same trading opportunities. They are engaged in a race to be first to detect a trading opportunity when it appears.[17]

Traders' expected profit functions have another interesting property. It is readily checked that the marginal effect of an increase in the monitoring level of a trader on one side increases in the aggregate monitoring level of the traders on the other side. That is, $\frac{\partial^2 \Pi_{im}}{\partial \mu_j \partial \lambda_i} > 0$ and $\frac{\partial^2 \Pi_{it}}{\partial \lambda_j \partial \mu_i} > 0$. 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.

Using the same type of argument as above, we write the expected profit of the

[17] In reality, this aspect is a key reason for automating order submission. See "Tackling latency-the algorithmic arms race," IBM Global Business Services report.

trading platform as

$$\Pi_E \equiv \frac{c_m + c_t}{D} = \bar{c} \cdot Vol\left(\bar{\lambda}, \bar{\mu}\right), \tag{11}$$

where

$$Vol\left(\bar{\lambda}, \bar{\mu}\right) \equiv \frac{\bar{\lambda} \cdot \bar{\mu}}{\bar{\lambda} + \bar{\mu}}. \tag{12}$$

The variable $Vol\left(\bar{\lambda}, \bar{\mu}\right)$ measures the trading rate on the trading platform (one over the duration of a cycle), which is also the average trading volume per unit of time on the trading platform. Thus, the long run payoff of the platform per unit of time is simply the average number of shares traded per unit of time multiplied by the total trading fee earned by the platform on each transaction.

Observe that $\bar{\lambda}$ and $\bar{\mu}$ can be viewed as measure of "latent liquidity." Indeed, $\overline{T}_t = \frac{1}{\bar{\mu}}$ is the average time it takes for the market-taking side to seize to a competitive offer whereas $\overline{T}_m = \frac{1}{\bar{\lambda}}$ is the average time it takes for the market-making side to post an offer. We refer to the first duration as the time from an order to a trade and the second duration as the time from a trade to an order. These durations are of interest as they can be measured empirically with high frequency data. We denote the ratio of these two durations by $C \overset{def}{=} \frac{\overline{T}_t}{\overline{T}_m} = \frac{\bar{\lambda}}{\bar{\mu}}$. We refer to this ratio as the "*time structure of a cycle*".

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 corresponds to 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. Using (9) and (10), an equilibrium in this stage is a vector of monitoring intensities $(\lambda_1^*, \ldots, \lambda_M^*, \mu_1^*, \ldots, \mu_N^*)$ such that for all $i = 1, \ldots, M$ and $j = 1, \ldots, N$:

$$\lambda_i^* = \arg\max_{\lambda_i} \left[\frac{\lambda_i \left(\mu_1^* + \ldots + \mu_N^*\right)\left(\frac{\Delta}{2} - c_m\right)}{\lambda_1^* + \ldots + \lambda_i^* + \ldots + \lambda_M^* + \mu_1^* + \ldots \mu_N^*} - \frac{1}{2}\beta\lambda_i^2 \right] \tag{13}$$

$$\mu_j^* = \arg\max_{\mu_j} \left[\frac{\mu_j \left(\lambda_1^* + \ldots + \lambda_M^*\right)\left(L - \frac{\Delta}{2} - c_t\right)}{\lambda_1^* + \ldots + \lambda_M^* + \mu_1^* + \ldots \mu_j + \ldots \mu_N^*} - \frac{1}{2}\gamma\mu_j^2 \right]. \tag{14}$$

For tractability, we further restrict attention to symmetric equilibria, i.e. equilibria in which $\lambda_1^* = \lambda_2^* = ... = \lambda_M^*$ and $\mu_1^* = \mu_2^* = ... = \mu_N^*$.

Then, given a symmetric Nash equilibrium in the monitoring intensities, we solve the trading platform's problem by finding the fee structure (c_m^*, c_t^*) that maximizes the trading platform's expected profit (equation (11)). In most of the paper we assume that $\bar{c}$ is not a choice parameter for the platform to better focus the analysis on the fee structure. It is straightforward to endogenize $\bar{c}$, as shown in Section 4.1.

3 Equilibrium Monitoring Intensities in the Short Run

In this section we first study the equilibrium monitoring intensities for a given set of fees (c_m, c_t). 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 both sides.

To see this point, consider how the no-trade equilibrium arises. If a market-maker expects that market-takers do not monitor the quotes on the trading platform, then she expects no trade on the platform. Given that monitoring is costly, it is not worth inspecting the state of the platform, and so she sets $\lambda_i = 0$. Similarly, if a market-taker expects market-makers not to post quotes, then he has no incentive to monitor, setting $\mu_j = 0$. 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. That is, $\lambda_i^* = \mu_j^* = 0$ for all $i \in \{1, \ldots, M\}$ and $j \in' 1, \ldots, N\}$ is an equilibrium. The trading volume in this equilibrium is zero.*

The second equilibrium does involve monitoring and trade. To describe this equilibrium, let

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

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 symmetric equilibrium with trading. In this equilibrium, traders' monitoring intensities are given by*

$$\lambda_i^* = \left(\frac{M + (M-1)\,\Omega^*}{(1+\Omega^*)^2}\right)\left(\frac{\pi_m}{M\beta}\right) \qquad i = 1, \ldots, M \tag{15}$$

$$\mu_j^* = \left(\frac{\Omega^*\left((1+\Omega^*)\,N - 1\right)}{(1+\Omega^*)^2}\right)\left(\frac{\pi_t}{N\gamma}\right) \qquad j = 1, \ldots, N \tag{16}$$

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

$$\Omega^3 N + (N-1)\Omega^2 - (M-1)\,z\Omega - Mz = 0. \tag{17}$$

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

The next corollary describes the effect of a change in the number of market participants on monitoring intensities and trading volume more systematically.

Corollary 1 *In the unique equilibrium with trading, for fixed fees of the platform,*

1. *Market-makers' individual monitoring levels increase with the number of market-takers and vice versa, that is $\frac{\partial \lambda_i^*}{\partial N} > 0$ and $\frac{\partial \mu_j^*}{\partial M} > 0$ for all i and j.*

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

3. *Thus, the trading rate increases in the number of participants on either side ($\frac{\partial Vol\left(\bar{\lambda}^*, \bar{\mu}^*\right)}{\partial M} > 0$ and $\frac{\partial Vol\left(\bar{\lambda}^*, \bar{\mu}^*\right)}{\partial N} > 0$).*

Thus, an increase in the number of market participants on one side increases market monitoring on **both** sides. This is a consequence of the strategic complementarity between both sides that we noticed in the previous section. Consider for instance an increase in the number of market-makers. The immediate effect of this increase is to enlarge the aggregate monitoring of market-makers since they are more

numerous. But in turn, this effect is conducive to more monitoring by market-takers as they expect trading opportunities to be more frequent. As a consequence, market-takers monitor more in equilibrium ($\frac{\partial \bar{\mu}^*}{\partial M} > 0$). This increase feeds back positively on market-makers' incentive to monitor, and thereby amplifies the initial increase in market-makers' monitoring intensity. Eventually, the trading rate enlarges.

We cannot sign the effect of an increase in participation on one side on the monitoring levels chosen by participants on this side. Consider again an increase in the number of market-makers. This increase intensifies competition for trading opportunities between market-makers since their total monitoring enlarges. Thus, it lowers each market-maker's incentive to monitor. But on the other hand, this increase is conducive to more monitoring by market-takers, which fosters each market-makers' incentive to monitor. We cannot, in general, determine whether the first effect (competition) or the second effect (complementarity) dominates. Yet, an increase in the number of market participants on one side enlarges the total attention of *all* market participants and thereby the trading rate, as shown by Corollary 1.

The next corollary analyzes the effect of a change in the monitoring cost or the monitoring benefit (profit-per-cycle) of one side on traders' monitoring intensity and the trading rate, for a fixed tick size.[18]

Corollary 2 *In the unique equilibrium with trading, for a fixed tick size,*

1. *Market-makers and market-takers' monitoring intensities decrease in market-makers' monitoring cost* ($\frac{\partial \lambda_i^*}{\partial \beta} < 0$ *and* $\frac{\partial \mu_j^*}{\partial \beta} < 0$) *and increase in market-makers' profit per-cycle* ($\frac{\partial \lambda_i^*}{\partial \pi_m} > 0$ *and* $\frac{\partial \mu_j^*}{\partial \pi_m} > 0$).

2. *Market-makers and market-takers' monitoring intensities decrease in market-takers' monitoring cost* ($\frac{\partial \lambda_i^*}{\partial \gamma} < 0$ *and* $\frac{\partial \mu_j^*}{\partial \gamma} < 0$) *and increase in market-takers' profit per cycle* ($\frac{\partial \lambda_i^*}{\partial \pi_t} > 0$ *and* $\frac{\partial \mu_j^*}{\partial \pi_t} > 0$).

[18] We fix the tick size because a change in the tick size has opposite effects on the benefit per trade of market-makers and market-takers. Thus, we cannot unambiguously sign its effect on the trading rate.

3. *The trading rate decreases in the monitoring costs* ($\frac{\partial Vol(\bar{\lambda}^*,\bar{\mu}^*)}{\partial \beta} < 0$ *and* $\frac{\partial Vol(\bar{\lambda}^*,\bar{\mu}^*)}{\partial \gamma} < 0$) *and increases in profits per cycle* ($\frac{\partial Vol(\bar{\lambda}^*,\bar{\mu}^*)}{\partial \pi_m} > 0$ *and* $\frac{\partial Vol(\bar{\lambda}^*,\bar{\mu}^*)}{\partial \pi_t} > 0$).

The strategic complementarity in monitoring decisions of both sides is again key for this finding. For instance, consider a decrease in the monitoring cost for marke-makers. This decrease raises their monitoring levels, other things equal. But, in turn, it induces market-takers to monitor more intensively, even though their cost of monitoring has not changed and this effect reinforces the initial increase in market-makers' monitoring level. This multiplier effect shows that a small decrease in monitoring cost on one side can eventually trigger a large increase in trading volume as it raises the attention of both sides.

The corollary also implies that the trading platform must account for this complementarity in setting its fees. For instance, an increase in the fee on market-makers, c_m, decreases their benefit per trade, π_m. Thus, it directly decreases market-makers' monitoring intensity and it indirectly decreases market-takers' monitoring intensities, since monitoring of both sides are complements.[19] As a consequence, the initial effect of the decrease in fees on trading volume is amplified.

In equilibrium, there can be an imbalance in the aggregate attention of each side to the trading process, as shown by the next corollary.

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 side have identical monitoring intensities.*

Thus, in equilibrium, there is an *excess of attention* by the market-making side (resp. market-taking side) when $\frac{z(2M-1)}{(2N-1)} > 1$ ($\frac{z(2M-1)}{(2N-1)} < 1$). 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. If instead $\frac{\pi_m}{\beta} = \frac{\pi_t}{\gamma}$ and $M > N$, the market-making side inspects

[19] This indirect effect does not arise in Rochet and Tirole's (2003) model of two-sided markets.

the market more frequently simply because this side has more participants. Figure 1 provides a graphical illustration of Corollary 3.

Insert Figure 1 about here.

The possibility of an excess of attention of one side relative to the other has interesting implications for (i) the make-take spread and (ii) the empirical relationship between the make-take spread and the time structure of a cycle ($\frac{\overline{T}_t}{\overline{T}_m} = \frac{\bar{\lambda}}{\bar{\mu}}$). We discuss these implications Sections 4 and 5 below.

In general we do not have a closed-form solution for traders' monitoring levels because we cannot solve for Ω^* in equation (17). However, there are two polar cases in which we can do so.

The first case is the "bilateral monopoly case" in which the market features one market-maker and one market-taker ($M = 1$ and $N = 1$). In this case, the solution to equation (17) is $\Omega^* = z^{\frac{1}{3}}$. Thus, using equations (15) and (16), we obtain the monitoring intensities of the market-making side and the market-taking side:

$$\lambda_1^* = \left(1 + z^{\frac{1}{3}}\right)^{-2} \cdot \left(\frac{\pi_m}{\beta}\right), \tag{18}$$

$$\mu_1^* = \left(1 + z^{-\frac{1}{3}}\right)^{-2} \cdot \left(\frac{\pi_t}{\gamma}\right). \tag{19}$$

The second case is the polar case in which the number of participants on both sides is very large but the ratio q of the number of market-makers to the number of market-takers is fixed.

Lemma 1 *:Consider the case in which $M = qN$. When the number of market participants goes to infinity, traders' individual monitoring levels in equilibrium are given by:*

$$\lambda_i^\infty \equiv \lim_{M\to\infty} \lambda_i^* = \frac{1}{1+\Omega^\infty}\left(\frac{\pi_m}{\beta}\right) \quad i = 1, 2, 3, ...$$

$$\mu_j^\infty \equiv \lim_{M\to\infty} \mu_j^* = \frac{\Omega^\infty}{1+\Omega^\infty}\left(\frac{\pi_t}{\gamma}\right) \quad j = 1, 2, 3, ...$$

with $\Omega^\infty = (zq)^{\frac{1}{2}}$.

Thus, traders' individual monitoring levels remain finite even though the number of participants goes to infinity. Consequently, traders' aggregate monitoring levels and the trading rate explode in this case. Thus, we shall focus on the case in which the number of participants is very large but finite. We call this case "the large market case." In this case, we can obtain approximations of the aggregate monitoring level on each side and the resulting trading rate, as shown by the next lemma.

Lemma 2 *:Consider the case in which $M = qN$ and let define*

$$\bar{\lambda}^{\infty}(M) \equiv \frac{\pi_m}{\beta}\frac{M}{1+\Omega^{\infty}} - \frac{\pi_m}{\beta}\frac{\Omega^{\infty}\left(q+2+\Omega^{\infty}\right)}{2\left(1+\Omega^{\infty}\right)^3}, \tag{20}$$

$$\overline{\mu}^{\infty}(M) \equiv \left(\Omega^{\infty}\right)^{-1}\cdot\bar{\lambda}^{\infty}(M), \tag{21}$$

$$Vol^{\infty}(M) \equiv \frac{\bar{\lambda}^{\infty}(M)}{1+\Omega^{\infty}}. \tag{22}$$

Then, (i) $\lim_{M\to\infty}\left(\bar{\lambda}^{*}(M) - \bar{\lambda}^{\infty}(M)\right) = 0$, *(ii)* $\lim_{M\to\infty}\left(\overline{\mu}^{*}(M) - \overline{\mu}^{\infty}(M)\right) = 0$, *and (iii)* $\lim_{M\to\infty}\left(Vol\left(\bar{\lambda}^{*},\bar{\mu}^{*}\right) - Vol^{\infty}(M)\right) = 0$.

Thus, when the number of market participants becomes large, we can approximate the trading rate and traders' aggregate monitoring levels by $Vol^{\infty}(M)$, $\bar{\lambda}^{\infty}(M)$, and $\overline{\mu}^{\infty}(M)$. Numerical simulations indicate that these approximations become good very quickly (that is, they hold even for small values of M and N).

4 The Determinants of the Make/Take Spread

Now, we study the determination of its fees by the trading platform. In most of the analysis, we fix exogenously the total fee charged by the trading platform, $\bar{c}$, as we are mainly interested in the determinants of the platform's fee structure, (c_m, c_t). We refer to $c_m - c_t$ as being 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. As explained in the introduction, in reality, the make-take spread is in general negative ($c_m < c_t$). Our goal is to understand how the exogenous parameters of the model (the tick size, the monitoring costs, and the number of participants) affect the make-take spread.

It is worth stressing at the outset that price discreteness explains why the fee structure matters. Indeed, it prevents market-makers from neutralizing a change in fees by an adjustment in their offers. For instance, market-makers cannot pass-through a decrease in their trading fee by quoting a more attractive price because their quotes must be on the grid. Thus, the breakdown of the total trading fee alters the balance of attention between the market-making side and the market-taking side and thereby it affects the trading rate.

As explained in Section 2.3, the expected profit of the trading platform per unit of time is $\Pi_E = (c_m + c_t)\, Vol(\bar{\lambda}^*, \bar{\mu}^*)$. Trading fees affect traders' monitoring decisions and thereby the trading rate. For instance, consider an increase in the fee charged on market-makers, c_m. This increase reduces their expected profit per trade (π_m) and thereby their monitoring intensity in equilibrium (Corollary 2). As a consequence, market-takers' monitoring intensities decrease as well and the trading rate becomes smaller (Corollary 2).

For a given total fee $\bar{c}$, the objective function of the trading platform is

$$\max_{c_m, c_t} \quad (c_m + c_t)\, Vol(\bar{\lambda}^*, \bar{\mu}^*), \tag{23}$$

$$s.t \quad : \quad c_m + c_t = \bar{c} \tag{24}$$

Thus, the problem of the trading platform is to find the fee structure (c_m^*, c_t^*) that maximizes it trading rate, $Vol(\bar{\lambda}^*, \bar{\mu}^*) = \frac{\bar{\lambda}^* \cdot \bar{\mu}^*}{\bar{\lambda}^* + \bar{\mu}^*}$. The first order conditions for this optimization problem impose that:

$$\frac{\partial Vol(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_m} = \frac{\partial Vol(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_t}. \tag{25}$$

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

$$\begin{aligned} \eta_{mm} &\equiv \frac{\partial \log(\bar{\lambda}^*)}{\partial c_m} \quad \text{and} \quad \eta_{mt} \equiv \frac{\partial \log(\bar{\mu}^*)}{\partial c_m}, \\ \eta_{tm} &\equiv \frac{\partial \log(\bar{\lambda}^*)}{\partial c_t} \quad \text{and} \quad \eta_{tt} \equiv \frac{\partial \log(\bar{\mu}^*)}{\partial c_t}. \end{aligned} \tag{26}$$

Variables η_{mm} and η_{tm} measure the elasticities of the total monitoring level of the market-making side to the fee charged on the market-makers on the one hand and the

fee charged on market-takers on the other hand. Variables η_{tm} and η_{tt} measure the elasticities of the total monitoring level of the market-taking side to the fees. Using equation (25), we obtain the following result.

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

$$\begin{aligned} c_m^* &= \left(\frac{h}{h+1}\right)\bar{c}, \\ c_t^* &= \bar{c} - c_m^* = \left(\frac{1}{h+1}\right)\bar{c}, \end{aligned} \tag{27}$$

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

The elasticities of monitoring levels to a change in fees depend on the fees through Ω^*, π_m, and π_t. Thus, the optimal fee structure is *implicitly* defined by equation (27). Yet, the previous result shows that in general, the flat fee structure is not optimal, except if $h = 1$.

To develop insights on the determinants of the make-take spread, we now consider two special cases (i) the large market and (ii) the bilateral monopoly. We show that the effects of the monitoring costs (γ and β) and (ii) the tick size are identical in both cases. Moreover, in the large market case, we can study the effects of varying the ratio of the number of market-makers to the number of market-takers (q) on the make-take spread. We will then show through numerical simulations that the insights obtained in the two polar cases are robust in the general case.

4.1 The Large Market

We first consider the case in which the number of participants is large and such that $\frac{M}{N} = q$. Using Lemma 2, we can solve for the optimal fee structure of the platform. We obtain the following result.

Proposition 3 *: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:*

$$c_m^* = \frac{1}{2}\left(\Delta - \frac{2(L-\bar{c})}{(1+(qr)^{\frac{1}{3}})}\right) \quad \textit{and } c_t^* = \bar{c}^* - c_m^*. \tag{28}$$

For these fees,

$$\pi_m^* = \frac{L-\bar{c}}{(1+(qr)^{\frac{1}{3}})} \quad \textit{and} \quad \pi_t^* = \frac{L-\bar{c}}{(1+(qr)^{-\frac{1}{3}})}, \tag{29}$$

and the equilibrium monitoring intensities are:

$$\lambda_i^\infty = \frac{L-\bar{c}}{\beta\left(1+(qr)^{\frac{1}{3}}\right)^2} \quad \textit{and} \quad \mu_i^\infty = \frac{L-\bar{c}}{\gamma\left(1+(qr)^{-\frac{1}{3}}\right)^2}. \tag{30}$$

We now discuss how the tick size, the monitoring costs and the ratio of market participants on both sides determine the optimal fee structure of the platform. Let $\overline{\Delta}(q,r) \overset{def}{=} 2(L-\bar{c})(1+(qr)^{\frac{1}{3}})^{-1}+\bar{c}$. Using equation (28), it is immediate that the make/take spread is zero if and only if $\Delta = \overline{\Delta}$. If $\Delta > \overline{\Delta}$, the make-take spread is positive and if $\Delta < \overline{\Delta}$, the make-take spread is negative, as shown on Figure 2.[20]

Insert Figure 2 about here

Furthermore, the model has implications for the sources of variations in the make-take spread, as shown by the next corollary.

Corollary 4 *: 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.*

These findings follow from the same general principle. That is, when a parameter changes so that the level of attention of one side rises relative to the level of attention of the other side then the trading platform raises its fee on the side whose attention increases. In other words, the trading platform uses its fee to equilibrate the level of attention of the market-making and the market-taking side.

For instance, consider an increase in the tick size. This increase reinforces market-makers' incentive to monitor the market since, other things equal, they get a larger

[20] The model also implies that in some cases it might be optimal to subsidize one side. Indeed, equation (28) implies that the fee charged on market-makers (resp. market-takers) can be negative (a subsidy) if the tick size is small (resp. large) enough. In this case, one may wonder whether it is not optimal for a market-maker to undercut his competitors by posting an offer at $a-\Delta$ when an offer is already standing at a. Given the optimal fees charged by the platform, this is never optimal however since this yields a profit of $a-\Delta-v_0-c_m^* = \frac{L-\bar{c}}{1+r^{1/3}} - \Delta \leq 0$ since $L \leq \Delta$.

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 amount of attention 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.

Let $\Phi^*(r,q) \equiv \frac{\pi_m^*}{\pi_m^*+\pi_t^*}$ be the fraction of the *net* gains from trade $(L-\bar{c})$ obtained by the market-maker in a given transaction. Using equation (29), we obtain that in equilibrium:

$$\Phi^*(r,q) \equiv \frac{\pi_m^*}{\pi_m^*+\pi_t^*} = \frac{1}{1+(qr)^{\frac{1}{3}}}. \tag{31}$$

We deduce the following result.

Corollary 5 *:*

1. *Market-makers get a smaller fraction of the net gains from trade when (i) their monitoring cost becomes relatively smaller($\frac{\partial \Phi^*}{\partial r} < 0$) or when (ii) the size of the market-making sector relative to the market-taking sector enlarges ($\frac{\partial \Phi^*}{\partial q} < 0$).*

2. *When the relative size of the market-making sector goes to infinity ($q \to \infty$) or market-makers' monitoring cost goes to zero ($\beta \to 0$), market-makers' gains from trade go to zero ($\Phi^* \to 0$).*

3. *When the relative size of the market-making sector goes to zero ($q \to 0$) or market-takers' monitoring cost goes to zero ($\beta \to 0$), market-takers' gains from trade go to zero ($\Phi^* \to 1$).*

As pointed out in the introduction, algorithmic trading reduces the cost of monitoring but not necessarily at the same speed for both sides. In this case, Corollary 5 shows that the development of algorithmic trading results in a counter-intuitive redistribution of trading profits per trade. In the short-run (that is, for fixed fees of the platform), a decline in the monitoring cost leaves the division of the gains per trade unchanged. But, in the long run, fees adjust and the division of the gains from trade is shifted in favor of the side whose monitoring cost declines the least.

The previous findings about the optimal fee structure hold for any level $\bar{c}$. As $\bar{c}$ enlarges, market-makers and market-takers watch the market less closely. As a result, the trading rate decreases. Thus, in choosing its total fee, the trading platform faces the standard price-quantity trade-off for a monopolist.

Corollary 6 *For r and q being fixed, when the trading platform optimally chooses its fee structure, the trading rate is (i) inversely related to traders' monitoring cost, (ii) positively related to the size of gains from trade $(L - \bar{c})$ and (iii) independent from the tick size. Moreover, it is maximal for $\bar{c} = L/2$.*

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.

4.2 The Bilateral Monopoly ($M = N = 1$)

Using the expressions for monitoring levels on each side (equations (18) and (19)), we can solve for the optimal fee structure of the platform. We obtain the following result.

Proposition 4 *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:*

$$c_m^* = \frac{1}{2}\left(\Delta - \frac{2(L-\bar{c})}{(1+r^{\frac{1}{4}})}\right) \quad \textit{and } c_t^* = \bar{c}^* - c_m^*. \tag{32}$$

For these fees,

$$\pi_m^* = \frac{L - \bar{c}}{(1 + r^{\frac{1}{4}})} \quad \textit{and} \quad \pi_t^* = \frac{L - \bar{c}}{(1 + r^{-\frac{1}{4}})}, \tag{33}$$

and the equilibrium monitoring intensities are:

$$\lambda_1^* = \frac{L - \bar{c}}{\beta \left(1 + r^{\frac{1}{4}}\right)^3} \quad \textit{and} \quad \mu_1^* = \frac{L - \bar{c}}{\gamma \left(1 + r^{-\frac{1}{4}}\right)^3}. \tag{34}$$

Clearly, this result is qualitatively similar to Proposition 3. In particular, it is readily checked that our findings regarding the effects of the tick size, the relative size of the market-making side and the relative monitoring cost of market-takers (Corollary 4) still hold in this case.

5 Implications

We now discuss some empirical implications of the model.

Duration Clustering. We pointed out that market-makers' and market-takers' monitoring decisions are complements. Thus, an exogenous shock that positively affects the aggregate monitoring level of one side also raises the aggregate monitoring level of the other side, as shown by Corollaries 1 and 2. This naturally leads to a positive correlation between (i) the average duration from a trade to a quote ($\overline{T}_m = \frac{1}{\lambda}$) and (ii) the average duration from a quote to a trade ($\overline{T}_t = \frac{1}{\mu}$).

For instance, consider an increase in the number of market-takers. In equilibrium, it leads to both a decrease in the reaction time of the market-taking side (as they monitor more) and the reaction time of the market-making side (as more monitoring by market-takers encourages more monitoring by market-makers). Thus, both $\overline{T}_m$ and $\overline{T}_t$ fall. As a consequence, the duration between trades ($\overline{T}_m + \overline{T}_t$) falls as well (these claims directly follow from Corollary 1).

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)). Our model suggests that clustering in time between trades could reflect the positive correlation between $\overline{T}_m$ and $\overline{T}_t$. More generally, it suggests to explain clustering by

the complementarity of liquidity suppliers' and liquidity demanders' monitoring decisions. Thus, 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) lead to a positive correlation between the various components ($\overline{T}_m$ and $\overline{T}_t$) of the total duration of a cycle.

Time Structure of a Cycle. The model has also interesting application for what we call the "time structure of a cycle", that is the ratio:

$$C \overset{def}{=} \frac{\overline{T}_t}{\overline{T}_m} = \frac{\overline{\lambda}}{\overline{\mu}}.$$

In equilibrium, this ratio is equal to Ω^* (Proposition 2). For the discussion, we focus again on the large market case but our predictions hold in other cases as well. We obtain the following result.

Corollary 7 *: In equilibrium, for fixed fees of the trading platform, the time structure of a cycle in the large market is:*

$$C(r, q, c_m, c_t) = \frac{\overline{T}_t}{\overline{T}_m} = (\frac{\pi_m rq}{\pi_t})^{1/2}. \tag{35}$$

Thus, the time from a quote to a trade relative to the time from a trade to a quote becomes relatively larger when (i) the relative size of the market-making side enlarges, (ii) the relative monitoring cost of the market-taking side enlarges, (iii) the fee charged on market-makers decreases and (iv) the fee charged on market-takers increases.

The two first implications (those regarding the effect of q and r on C) also hold when fees are set at the optimal level. Indeed, using Proposition 3 and equation (35), we obtain that:

$$C(r, q, c_m^*, c_t^*)) = (\frac{\pi_m^* rq}{\pi_t^*})^{1/2} = (rq)^{2/3}. \tag{36}$$

The optimal make-take spread is also positively related to r and q (see Corollary 4). Thus, if fees are set optimally, the model also implies a positive correlation between the make-take spread and the ratio of the durations of the two phases in a cycle, $\frac{\overline{T}_t}{\overline{T}_m}$. 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 ratio of durations.

Tick size and Make-Take Spread. The model also implies a positive association between the make-take spread and the tick size. Interestingly, the proliferation of negative make-take spreads on U.S. equity trading platforms (and even rebates paid to liquidity suppliers) coincide with a reduction in the tick size on these platforms. Moreover, this practice was introduced by ECNs such as Archipelago or Island in the 90s which, at this time, were operating on much finer grids than their competitors (Nasdaq and NYSE).[21] Last, since January 2007, the tick size has been reduced for some options in U.S. option markets (so called "penny pilot program"). Interestingly this reduction is associated with the adoption of make/take fees by a some trading platforms (e.g., NYSE Arca Options and the Boston Options Exchange) for the options that trade on pennies.

The model suggests two other reasons for the low make-take spreads that are observed in reality (see Figure 2). 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 higher. 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.

[21] 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' main competitor at the time of their study.

Volume and Algorithmic Trading. The model also implies that an improvement in monitoring technology, such as algorithmic trading, can lead to a burst in trading volume. Indeed, as shown by Corollary 2, a decrease in the monitoring costs of both the market-making side and the market-taking side translates into an increase in trading volume. The result also holds when the fee structure is endogenous (as shown by Corollary 6).

This result is interesting as analysts relate recent surges in trading volume to the development of algorithmic trading. 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 model suggests that this surge happens because algorithmic trading accelerates the speed at which liquidity demanders and liquidity suppliers respond to each other.

Bid-Ask Spread and Algorithmic Trading. Quoted bid-ask spreads are often used as a measure of liquidity. To compute the bid-ask spread in our model, assume that there is a large number of shares are 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 and therefore they cannot intervene profitably at price a.

This assumption does not change traders' optimal behavior since market-takers only trade at a. Thus, the (half) bid-ask spread (the best offer less v_0) is either a (in state F) or $a + \Delta$ (in state E). During a cycle, the market is in state F for an average duration $\overline{T}_t$ and in state E for an average duration $\overline{T}_m$. Thus, the average half bid-ask spread (denoted ES) is:

$$ES = \theta a + (1 - \theta)(a + \Delta) = a + (1 - \theta)\Delta. \tag{37}$$

with

$$\theta \overset{def}{=} \frac{\overline{T}_t}{\overline{T}_t + \overline{T}_m} = (1 + \frac{\overline{\mu}}{\overline{\lambda}})^{-1} \tag{38}$$

Thus, the average bid-ask spread decreases when θ increases, that is when the ratio $\frac{\overline{\lambda}}{\overline{\mu}}$ enlarges. In equilibrium, $\frac{\overline{\lambda}}{\overline{\mu}}$ increases in the relative monitoring cost ratio, $r = \frac{\gamma}{\beta}$.

For instance, in the large market, $\frac{\overline{\lambda}}{\overline{\mu}} = (\frac{\pi_m rq}{\pi_t})^{1/2}$. This relationship shows that the effect of algorithmic trading on the bid-ask spread depends on whether the associated reduction in monitoring cost is faster for the market-taking side or the market-making side.

In the former case, r decreases and therefore $\frac{\overline{\lambda}}{\overline{\mu}}$ decreases as well. Accordingly, the bid-ask spread enlarges. Intuitively, the reduction in the cost of monitoring for market-takers accelerates the speed at which liquidity is consumed without being matched by a commensurate increase in the provision of liquidity. As a consequence the bid-ask spread widens. If instead, algorithmic trading leads to a faster reduction in monitoring costs for market-makers then it should result in smaller bid-ask spreads. The same result holds if fees are set optimally since in this case $\frac{\overline{\lambda}}{\overline{\mu}} = (rq)^{2/3}$

Hendershott et al. (2008) find empirically that, on average, algorithmic trading has triggered a reduction in the average effective bid-ask spreads for stocks listed on the NYSE. This suggests that, in reality, monitoring costs for market-makers have fallen more quickly than for market-takers.

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.

Our main findings are as follows:

1. Monitoring decisions by market-makers and market-takers are complements. Thus, there is a coordination problem in the decisions of both sides that can result in high or low levels of trading activity.

2. An increase in the number of participants on one side or a decrease in the monitoring cost of one side result in more attention by both sides and a higher trading rate.

3. For a fixed trading fee earned by the platform, there is an allocation of this fee between market-makers and market-takers that maximizes the trading rate. This allocation is such that there is a make/take spread: the fee charged on market-makers is different from the fee charged on market-takers.

4. The make/take spread enlarges with (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.

5. When fees are set optimally, market-makers (resp. market-takers) get a smaller fraction of the gains from trade when (i) their number enlarges or (ii) their monitoring costs decreases.

7 Appendix

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

Proof of Proposition 2: From (13), the first order condition for market-maker i is:

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

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

$$\frac{\bar{\mu}\left(\left(\bar{\mu}+\bar{\lambda}\right)M-\bar{\lambda}\right)}{\left(\bar{\lambda}+\bar{\mu}\right)^2}\frac{\pi_m}{\beta}=\bar{\lambda}. \tag{39}$$

Similarly, for market-takers we obtain,

$$\frac{\bar{\lambda}\left(\left(\bar{\mu}+\bar{\lambda}\right)N-\bar{\mu}\right)}{\left(\bar{\lambda}+\bar{\mu}\right)^2}\frac{\pi_t}{\gamma}=\bar{\mu}. \tag{40}$$

Let $\Omega \equiv \frac{\bar{\lambda}}{\bar{\mu}}$. Dividing (39) and (40) by $\bar{\mu}^2$ we have,

$$\frac{M + (M-1)\,\Omega}{(1+\Omega)^2}\frac{\pi_m}{\beta} = \bar{\lambda}. \tag{41}$$

$$\frac{\Omega\,((1+\Omega)\,N-1)}{(1+\Omega)^2}\frac{\pi_t}{\gamma} = \bar{\mu} \tag{42}$$

Dividing these two equations gives,

$$\frac{(M + (M-1)\,\Omega)}{\Omega^2\,((1+\Omega)\,N-1)}z = 1, \tag{43}$$

or equivalently,

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

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

$$\Omega = g(\Omega, M, N, z). \tag{44}$$

with

$$g(\Omega, M, N, z) = \frac{(M-1)z}{\Omega N} + \frac{Mz}{N\Omega^2} - \frac{N-1}{N}. \tag{45}$$

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, (44) 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 (41) and (42). Traders' individual monitoring levels then follow since, in a symmetric equilibrium, $\lambda_i = \bar{\lambda}/M$ and $\mu_j = \bar{\mu}/N$ for all i, j.

■

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

$$\Omega^* = g(\Omega^*, M, N, z), \tag{46}$$

where $g(.)$ is defined in equation (45). It is immediate that $g(.)$ increases in M, decreases in N, and increases in z. As $g(.)$ decreases in Ω, we have

$$\frac{\partial \Omega^*}{\partial M} > 0, \tag{47}$$

$$\frac{\partial \Omega^*}{\partial N} < 0. \tag{48}$$

Now, using Equations (47) and (15), we conclude that:

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

Similarly, using equations (48) and (16), we deduce that

$$\frac{\partial \mu_j^*}{\partial M} > 0. \tag{49}$$

This proves the first part of Corollary 1. We also have

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

Thus, using equations (47) and (48), we conclude that $\frac{\bar{\lambda}^*}{\bar{\mu}^*}$ increases in M and decreases in N. Equation (49) 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, which proves the second part of Corollary 1. The last part of the corollary follows from the second part and the fact that the trading rate increase in traders' monitoring intensities. ■

Proof of Corollary 2: We first 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 $\frac{\partial \lambda_j^*}{\partial \beta} < 0$. Other claims in the corollary are proved in the same way. ■

Proof of Corollary 3: Using equation (17), 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 1, Ω^* increases in z. Hence, $\bar{\lambda}^* > \bar{\mu}^*$ iff $z > \frac{2N-1}{2M-1}$. ■

Proof of Lemma 1: Recall that Ω^* is the unique positive solution to the cubic equation

$$\Omega^3 N + (N-1)\Omega^2 - (M-1)\, z\Omega - Mz = 0. \tag{50}$$

and

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

Thus, using Equation (50),

$$\begin{aligned} z &= \frac{\Omega^{*3} N + (N-1)\Omega^{*2}}{(M-1)\,\Omega^* + M} = \frac{\Omega^{*3}\frac{M}{q} + (\frac{M}{q}-1)\Omega^{*2}}{(M-1)\,\Omega^* + M} \\ &= \frac{\Omega^{*3}\frac{1}{q} + \frac{(\frac{M}{q}-1)}{M}\Omega^{*2}}{\frac{(M-1)}{M}\Omega^* + 1} \underset{M\to\infty}{\longrightarrow} \frac{\Omega^{*2}}{q}. \end{aligned}$$

That is, when M and N becomes very large, Ω^* converges to a finite number, which we will denote by Ω^∞ given by

$$\Omega^\infty = (zq)^{\frac{1}{2}}. \tag{51}$$

Thus, using the expression for the monitoring intensity of a market-maker (equation (15)), we deduce that:

$$\begin{aligned} \lambda_i^\infty &\equiv \lim_{M\to\infty} \lambda_i^* = \left(\frac{M + (M-1)\,\Omega^*}{M\,(1+\Omega^*)^2}\right)\left(\frac{\pi_m}{\beta}\right) && (52) \\ &= \frac{1}{1+\Omega^\infty}\left(\frac{\pi_m}{\beta}\right), && (53) \\ &= \frac{1}{1+(zq)^{\frac{1}{2}}}\frac{\pi_m}{\beta} \quad \text{for} \quad i = 1,..,M. \end{aligned}$$

Similarly, for a market-taker:

$$\begin{aligned} \mu_j^\infty &\equiv \lim_{M\to\infty} \mu_j^* = \lim_{M\to\infty}\left(\frac{\Omega^*\left((1+\Omega^*)\frac{M}{q} - 1\right)}{\frac{M}{q}\,(1+\Omega^*)^2}\right)\left(\frac{\pi_t}{\gamma}\right) && (54) \\ &= \frac{1}{1+(zq)^{-\frac{1}{2}}}\left(\frac{\pi_t}{\gamma}\right). \quad \text{for} \quad j = 1,..,N. \blacksquare \end{aligned}$$

Proof of Lemma 2: After some algebra, we obtain:

$$\bar{\lambda}^*(M) - \frac{\pi_m}{\beta}\frac{M}{1+\Omega^\infty} = \frac{\pi_m}{\beta}\left(\frac{M(\Omega^\infty - \Omega^*)}{(1+\Omega^*)(1+\Omega^\infty)} - \frac{\Omega^*}{(1+\Omega^*)^2}\right). \tag{55}$$

Moreover it can be shown that:

$$\lim_{M\to\infty} M(\Omega^\infty - \Omega^*) = \frac{\Omega^\infty(\Omega^\infty - q)}{2(\Omega^\infty + 1)}, \tag{56}$$

We skip the proof of this claim for brevity. Thus, using equation (56), we obtain:

$$\begin{aligned}
\lim_{M\to\infty}\left(\bar{\lambda}^*(M) - \frac{\pi_m}{\beta}\frac{M}{1+\Omega^\infty}\right) &= \frac{\pi_m}{\beta}\lim_{M\to\infty}\left(\frac{M(\Omega^\infty - \Omega^*)}{(1+\Omega^*)(1+\Omega^\infty)} - \frac{\Omega^*}{(1+\Omega^*)^2}\right) \\
&= \frac{\pi_m}{\beta}\left(\frac{\Omega^\infty(\Omega^\infty - q)}{2(1+\Omega^\infty)^3} - \frac{\Omega^\infty}{(1+\Omega^\infty)^2}\right) \\
&= \frac{\pi_m}{\beta}\frac{\Omega^\infty}{(1+\Omega^\infty)^2}\left(\frac{\Omega^\infty - q - 2 - 2\Omega^\infty}{2(1+\Omega^\infty)}\right) \\
&= -\frac{\pi_m}{\beta}\frac{\Omega^\infty(q+2+\Omega^\infty)}{2(1+\Omega^\infty)^3}.
\end{aligned}$$

Thus, we have shown that

$$\lim_{M\to\infty}\left(\bar{\lambda}^*(M) - \bar{\lambda}^\infty(M)\right) = 0,$$

as required. The other claims of the proposition are then immediate since: $\overline{\mu}^*(M) = (\Omega^*)^{-1}\bar{\lambda}^*(M)$. ■

Proof of Lemma 3: We have

$$\begin{aligned}
\frac{\partial Vol(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_m} &= -Vol(\bar{\lambda}^*, \bar{\mu}^*)^2\left(\frac{\partial\overline{\lambda}}{\partial c_m}\frac{1}{\overline{\lambda}^2} + \frac{\partial\overline{\mu}}{\partial c_m}\frac{1}{\overline{\mu}^2}\right) \\
&= -\frac{Vol(\bar{\lambda}^*, \bar{\mu}^*)^2}{c_m}\left(\frac{\eta_{mm}}{\overline{\lambda}} + \frac{\eta_{mt}}{\overline{\mu}}\right).
\end{aligned} \tag{57}$$

and

$$\frac{\partial Vol(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_t} = -\frac{Vol(\bar{\lambda}^*, \bar{\mu}^*)^2}{c_t}\left(\frac{\eta_{tm}}{\overline{\lambda}} + \frac{\eta_{tt}}{\overline{\mu}}\right). \tag{58}$$

The optimal fee structure is such that:

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

Thus, using equations (57) and (58), we deduce that:

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

Using this equation, the proposition is then straightforward. ■

Proof of Proposition 3

For a fixed tick size, 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 problem as:

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

Now, for M large, we can approximate $Vol(\bar{\lambda}^*, \bar{\mu}^*)$ by (see Lemma 2):

$$Vol^{\infty}(M) \equiv \frac{\bar{\lambda}^{\infty}(M)}{1+\Omega^{\infty}} = \frac{\pi_m}{\beta}\frac{M}{(1+\Omega^{\infty})^2} - \frac{\pi_m}{\beta}\frac{\Omega^{\infty}(q+2+\Omega^{\infty})}{2(1+\Omega^{\infty})^4}$$

Thus, in the large market, we rewrite the trading platform's problem as:

$$\begin{aligned} & Max_{\pi_m,\pi_t} \bar{c}\left(\frac{\pi_m}{\beta}\frac{M}{(1+\Omega^{\infty})^2} - \frac{\pi_m}{\beta}\frac{\Omega^{\infty}(q+2+\Omega^{\infty})}{2(1+\Omega^{\infty})^4}\right) \\ & s.t \;\; \pi_t + \pi_m = L - \bar{c}. \end{aligned}$$

Let $K = -\frac{\pi_m}{\beta}\frac{\Omega^{\infty}(q+2+\Omega^{\infty})}{2(1+\Omega^{\infty})^4}$. The first order condition with respect to π_t is

$$-\frac{1}{(1+\Omega^{\infty})^2\beta} - \frac{L-\bar{c}-\pi_t}{\beta}\frac{2}{(1+\Omega^{\infty})^3}\frac{d\Omega^{\infty}}{d\pi_t} + (\frac{\partial K}{\partial \pi_t})\frac{1}{M} = 0, \tag{59}$$

Thus, as $\frac{\partial K}{\partial \pi_t}$ does not depend on M, when M goes to infinity, the first order condition imposes:

$$1 + \frac{2(L-\bar{c}-\pi_t)}{1+\Omega^{\infty}}\frac{d\Omega^{\infty}}{d\pi_t} = 0. \tag{60}$$

Now, recall that $\Omega^{\infty} = (zq)^{\frac{1}{2}}$. Hence:

$$\begin{aligned} \frac{d\Omega^{\infty}}{d\pi_t} &= \frac{d\Omega^{\infty}}{dz}\frac{dz}{d\pi_t} = \frac{1}{2}q(zq)^{-0.5}\frac{d}{d\pi_t}\left(\frac{L-\bar{c}-\pi_t}{\pi_t}\frac{\gamma}{\beta}\right) \\ &= -\frac{q}{2\Omega^{\infty}}\frac{L-\bar{c}}{\pi_t^2}\frac{\gamma}{\beta}. \end{aligned} \tag{61}$$

Thus, we can rewrite (60) as

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

Or,

$$1 - \frac{zq}{(1 + \Omega^{\infty})\,\Omega^{\infty}} \frac{L - \bar{c}}{\pi_t} = 0, \tag{62}$$

which simplifies to

$$\frac{\pi_t}{L - \bar{c}} = \frac{\Omega^{\infty}}{1 + \Omega^{\infty}}. \tag{63}$$

Denote

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

Then equation (63) imposes:

$$w = \frac{\Omega^{\infty}}{1 + \Omega^{\infty}} = \frac{1}{1 + (zq)^{-0.5}}. \tag{64}$$

Now observe that:

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

Thus, we can rewrite equation (64) as

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

The solution(s) to this equation provides the optimal value of w and thus the optimal fees for the trading platform (since these fees fix the sharing of the gains from trade between market-makers and market-takers). It is immediate that the previous equation as a unique solution:

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

It is easily checked that for $w > w^*$, the R.H.S of equation (62) is strictly positive. Thus, trading volume first increases as w increases from 0 to w^* and then decreases. This means that w^* is the unique global maximum of the trading platform's optimization problem.

Now, using the fact that $w = \frac{\pi_t}{L-\bar{c}}$, we can easily derive the expressions for the optimal fees and π_t^* and π_m^*. It is easily checked that for the optimal fees, we have

$$\Omega^\infty = (rq)^{1/3}.$$

Thus, using Lemma 1, traders' monitoring frequencies given the fees set by the platform are

$$\begin{aligned}
\lambda_i^\infty &= \frac{1}{1+\Omega^\infty}\left(\frac{\pi_m}{\beta}\right) = \frac{L-\bar{c}}{\beta\left(1+(rq)^{\frac{1}{3}}\right)^2}, \\
\mu_j^\infty &= \frac{\Omega^\infty}{1+\Omega^\infty}\left(\frac{\pi_t}{\gamma}\right) = \frac{L-\bar{c}}{\gamma\left(1+(rq)^{-\frac{1}{3}}\right)^2}.
\end{aligned}$$

Proof of Corollary 5 Immediate from equation (31). ■

Proof of Corollary 6 *To be written.*

. ■

Proof of Proposition 4: As in the proof of Proposition 3, we can use π_{mm} and π_{mt}. Thus, for a fixed $\bar{c}$, when $M = N = 1$, the platform problem is:

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

From equations (41) and (42),

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

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:

$$Max_{\pi_m,z} \frac{\lambda_1^*}{1+z^{\frac{1}{3}}}\bar{c} \tag{65}$$

$$s.t \quad \pi_m\left(1+\frac{\gamma}{\beta z}\right) = L - \bar{c}. \tag{66}$$

$$\text{and } \lambda_1^* = \frac{L-\bar{c}}{\beta\left(1+z^{\frac{1}{3}}\right)^2\left(1+\frac{\gamma}{\beta z}\right)} \tag{67}$$

This problem is equivalent to finding z that minimizes

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

The FOC 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

$$z=\left(\frac{\gamma}{\beta}\right)^{\frac{3}{4}}=r^{\frac{3}{4}}. \tag{68}$$

Using the constraint (66), we have,

$$\pi_m^*=\frac{L-\bar{c}}{1+r^{\frac{1}{4}}}. \tag{69}$$

It follows that,

$$\pi_t^*=L-\bar{c}-\pi_m=\frac{L-\bar{c}}{1+r^{-\frac{1}{4}}}. \tag{70}$$

Then, plugging (68), (69), and (70) into equations (18) and (19), we obtain the required expressions for λ_1^* and μ_1^*. ■

Proof of Corollary 7: In the large market, in equilibrium, we have $\Omega^*=\Omega^\infty=(zq)^{1/2}$ (see the proof of Lemma 1). The proposition follows. ■

References

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

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

[3] 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.

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

[5] 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.

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

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

[8] 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.

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

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

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

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

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

[14] Goldstein and Kavajecz (2000)

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

[16] Hendershott, T., Jones, C. and Menkveld, A. (2008) "Does algorithmic trading improve liquidity," mimeo, U.C. Berkeley.

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

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

[19] Liu, W. (2007) "Monitoring and Limit Order Submission Risks", Forthcoming *Journal of Financial Markets*.

[20] Pagano, M. (1989), "Trading Volume and Asset Liquidity", *Quarterly Journal of Economics*, 104, 255-276.

[21] Parlour, C. (1998), "Price Dynamics in Limit Order Markets," *Review of Financial Studies*, 11, 789-816.

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

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

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

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

[26] 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.

[27] Seppi (1997)

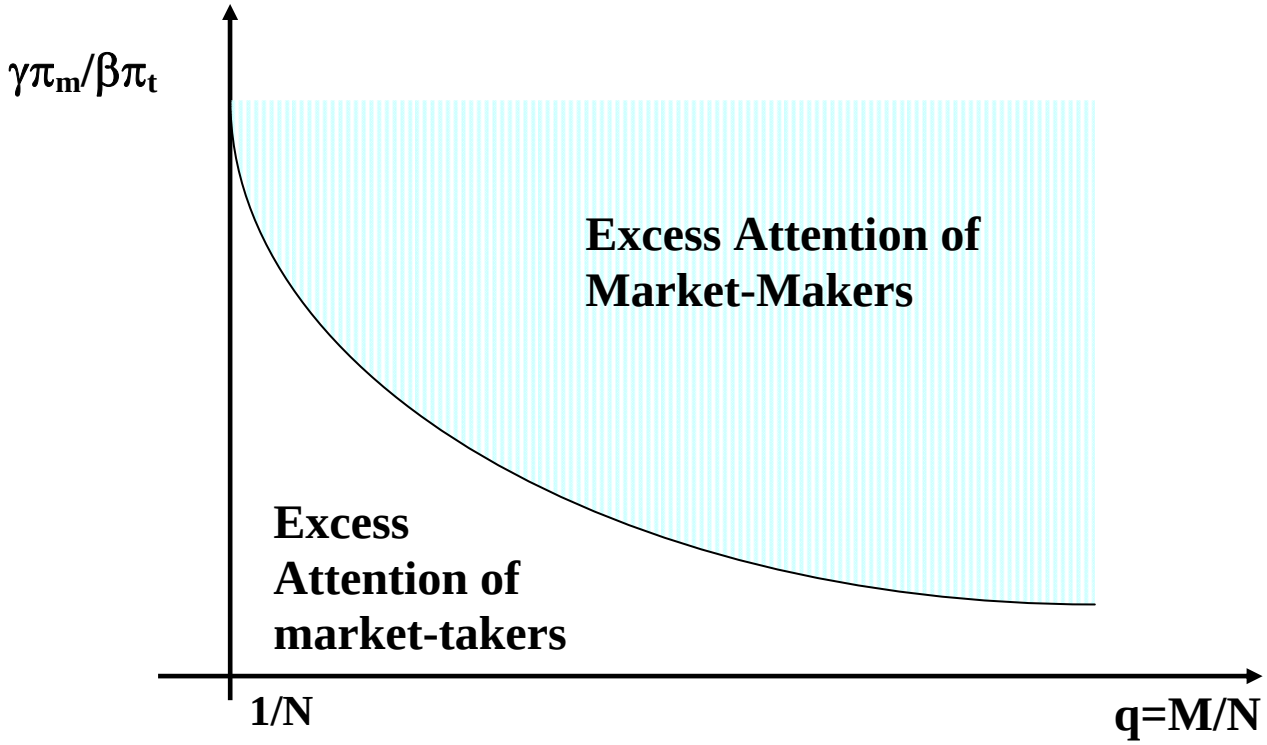
$\gamma\pi_m/\beta\pi_t$
Excess Attention of Market-Makers
Excess Attention of market-takers
1/N
q=M/N

Figure 1

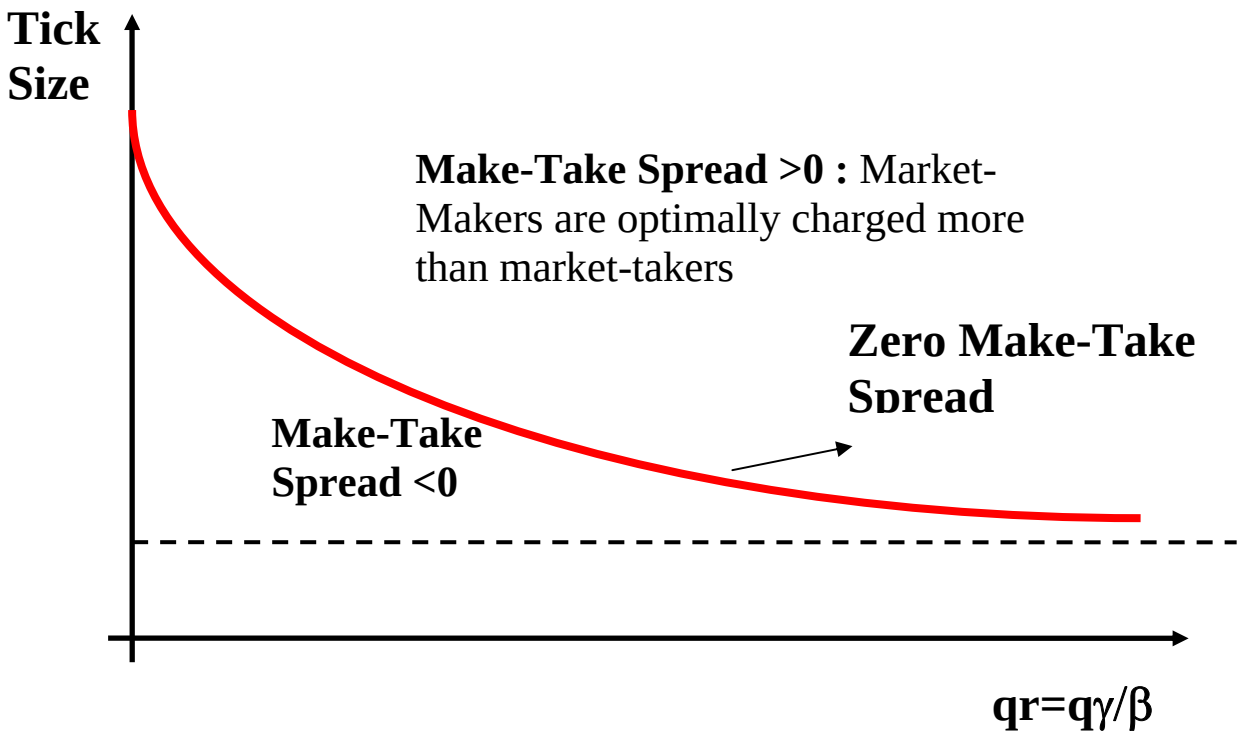
Tick Size
Make-Take Spread >0 : Market-Makers are optimally charged more than market-takers
Zero Make-Take Spread
Make-Take Spread <0
qr=qγ/β

Figure 2