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

November 14, 2008

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 periodically to capture profit opportunities. Complementarities between the two sides generate multiplicity of equilibria in which either liquidity is high or liquidity vanishes. We show how a reduction in monitoring costs (e.g. due to algorithmic trading) affects the distribution of gains from trade, the monitoring intensities, and the trading rate. We also analyze the optimal fee structure of the trading platform matching market-makers and market-takers. We find that it optimally charges different fees on each side and may even subsidize one side, as observed in reality. The main determinants of the fee structure are the tick-size, the relative number of traders on each side, and the relative cost of monitoring for each side.

Keywords: Monitoring, Make/Take Spread, Liquidity Rebates, Algorithmic trading, Two-Sided Markets.

1 Introduction

Trading in securities 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 profit 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 etc...) 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 this way, this paper speaks to two different sets of related issues.

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

[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.

distribution of trading gains between liquidity makers, liquidity takers, and trading platforms.

Secondly, the model sheds light on pricing schedules set by trading platforms. Increasingly, these platforms charge different fees on market orders (orders "taking liquidity") and limit orders (orders "making liquidity"). The difference between these fees is called the *make/take spread* and is usually positive. That is, investors providing liquidity pay a lower fee than investors taking liquidity. For instance, Table 1 in Appendix A gives the fees charged on liquidity makers and liquidity takers for U.S. equity trading platforms, as of July 2008. All these platforms subsidize liquidity makers by paying a rebate (or charge zero fees) 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-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 interest of market participants in 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.

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

[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.

[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

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 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, 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

[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 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).

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 determination of the *fee structure* – the breakdown between the fees charged on market-makers and market-takers by the trading platform. When the total fee per trade is fixed, we find that there is a unique fee structure that maximizes the trading rate and thereby the trading platform's profit. 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

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

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. In particular, it can be optimal for the trading platform to subsidize market-makers when (a) the number of market-makers is relatively low; (b) market-makers' monitoring cost is relatively large; or (c) the tick size is low. 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 liquidity by these investors.

The model also implies that 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. Market-takers' trading surplus vanishes when their monitoring cost becomes infinitesimal.

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

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

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] It also implies that 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.

Our analysis also contributes to the burgeoning literature on two-sided markets (e.g., Rochet and Tirole (2006) and Rochet and Tirole (2003)). 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 (see their Definition 1 on page 648).[10] This is a feature of our model in which the end-users are market-makers and market-takers.

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. Section 5 concludes. The proofs are in Appendix B.

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

[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 algoritmic traders to get faster access to this platform.

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

[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

and N.

In reality, traders can often choose whether to post a quote or 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) or Global Electronic Trading company (GETCO), which engage in high frequency market-making. 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.[12] Electronic market-makers primarily use limit orders whereas the second type 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., Hollifield et al.(2006), Duffie et al. (2005), or Parlour et al.(2005)).[13]

As market-takers have a higher valuation than market-makers, they will 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 to the mathematical complexity of 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-

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

[13] 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.

takers refuse to trade at a larger price on the grid ($\frac{\Delta}{2} < L < \frac{3}{2}\Delta$) 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). 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).[14]

The trading platform charges trading fees each time a trade occurs. The fee (per share) paid by a market-maker is denoted c_m, whereas the fee paid by a market-taker is denoted c_t. Thus, per transaction, the platform earns a revenue $\bar{c} \equiv c_m + c_t$. 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.[15]

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 net of the fee earned by the trading platform are $L - \bar{c}$. We focus on the case in which $\bar{c} < L$ since otherwise traders on one side at least lose money on each trade, and thereby 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

[14] Empirically, several papers document a reduction in quoted depth aftera 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.

[15] In reality,the net order processing cost may be negative since trading platforms in financial markets get revenues from the sale of information. This leads to a so called inverted fee structure in which $\bar{c} = c_m + c_t < 0$. For instance, for some stocks, the International Securities Exchange (ISE) charges a fee of \$0.003 per share for orders taking liquidity and rebates \$0.0032 per share for limit orders (see Table 1).

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 (empty book). A limit order is not posted at a.
2. State F – Liquidity is plentiful (full book). A limit order for one share is posted at a.

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 places a limit order 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 with each other for seizing profit opportunities reflected in full books. Market-makers and market-takers provide liquidity to one another as profits can only be realized after a limit order has been 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

[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-marker. Hence, in all cases, monitoring the market for a security more intensively is costly, because it reduces the monitoring of other securities.

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

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

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

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

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 (γ) decreases. This would be a result, for example, of automation of the monitoring process.

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. Rest of the game. 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 the 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 first (before the other market-makers do). 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

$$p_i \pi_m = \frac{\lambda_i}{\bar{\lambda}} \left(\frac{\Delta}{2} - c_m \right). \tag{3}$$

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

$$q_j \pi_t = \frac{\mu_j}{\bar{\mu}} \left(L - \frac{\Delta}{2} - c_t \right). \tag{4}$$

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

$$D \equiv \frac{1}{\bar{\lambda}} + \frac{1}{\bar{\mu}} = \frac{\bar{\lambda} + \bar{\mu}}{\bar{\lambda} \cdot \bar{\mu}}. \tag{5}$$

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

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

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

$$\Pi_{im} \equiv \lim_{T\to\infty} \frac{\Pi_i(T)}{T} = \lim_{T\to\infty} \frac{E_{\tilde{n}_T}(\sum_{k=1}^{\tilde{n}_t} p_i\pi_m)}{T} - \frac{1}{2}\beta\lambda_i^2. \tag{6}$$

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 (3) and (5), we can rewrite the objective function of market-maker i (equation (6)) as

$$\Pi_{im} = \frac{p_i\pi_m}{D} = \frac{\frac{\lambda_i}{\bar{\lambda}}\left(\frac{\Delta}{2} - c_m\right)}{\frac{\bar{\lambda}+\bar{\mu}}{\bar{\lambda}\cdot\bar{\mu}}} - \frac{1}{2}\beta\lambda_i^2 = \frac{\lambda_i\bar{\mu}\left(\frac{\Delta}{2} - c_m\right)}{\bar{\lambda}+\bar{\mu}} - \frac{1}{2}\beta\lambda_i^2. \tag{7}$$

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

$$\Pi_{jt} = \frac{q_j\pi_t}{D} - \frac{1}{2}\gamma\mu_j^2 = \frac{\mu_j\bar{\lambda}\left(L - \frac{\Delta}{2} - c_t\right)}{\bar{\lambda}+\bar{\mu}} - \frac{1}{2}\gamma\mu_j^2. \tag{8}$$

From (7) and (8), 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 $j \neq i$). Thus, traders' monitoring decisions on one side are substitutes. 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]

Conversely, the expected profit of a trader on one side increases in the monitoring intensities of the traders on the other side. For instance, $\frac{\partial\Pi_{im}}{\partial\mu_j} > 0$. That is, market-makers are more likely to check the state of the market frequently when they expect market-takers to inspect quotes frequently and vice versa. Thus, market-makers and market-takers' monitoring decisions reinforce each other.

[17] This aspect drove traders to automate the monitoring decision. See "Tackling latency-the algorithmic arms race," IBM Global Business Services report.

Using the same type of argument as for market-makers and market-takers, we write the objective function of the trading platform as

$$\Pi_E \equiv \frac{c_m + c_t}{D} = \bar{c} \cdot Vol\left(\bar{\lambda}, \bar{\mu}\right), \tag{9}$$

where

$$Vol\left(\bar{\lambda}, \bar{\mu}\right) \equiv \frac{\bar{\lambda} \cdot \bar{\mu}}{\bar{\lambda} + \bar{\mu}}. \tag{10}$$

The variable $Vol\left(\bar{\lambda}, \bar{\mu}\right)$ measures the trading rate on the trading platform (one over the duration of a cycle) and hence is 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 shares traded per unit of time multiplied by the trading fee charged by the platform.

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). 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 (7) and (8), 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 $j = 1, \ldots, N$

$$\lambda_i^* = \arg\max_{\lambda_i} \left[\frac{\lambda_i \left(\mu_1^* + \ldots + \mu_N^*\right) \left(\frac{\Delta}{2} - c_m\right)}{\lambda_1^* + \ldots + \lambda_i^* + \ldots + \lambda_M^* + \mu_1^* + \ldots \mu_N^*} - \frac{1}{2}\beta\lambda_i^2 \right] \tag{11}$$

$$\mu_j^* = \arg\max_{\mu_j} \left[\frac{\mu_j \left(\lambda_1^* + \ldots + \lambda_M^*\right) \left(L - \frac{\Delta}{2} - c_t\right)}{\lambda_1^* + \ldots + \lambda_M^* + \mu_1^* + \ldots \mu_j + \ldots \mu_N^*} - \frac{1}{2}\gamma\mu_j^2 \right]. \tag{12}$$

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 equation (9).

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). Thus, the comparative statics result obtained in this section describe the short run adjustments of monitoring intensities and trading rates to an change in the parameters $(M, N, \beta$ etc...). In the longer run, trading fees should adjust as well, as described in the next section.

For all parameters values, the model admits exactly two equilibria: (i) an equilibrium with no trading; and (ii) an equilibrium with trading. This multiplicity of equilibria is due to the complementarity in market-makers and market-takers' monitoring decisions.

The rationale behind the no-trade equilibrium is simple. 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 for her to inspect 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 $i = 1, \ldots, M$ and $j = 1, \ldots, N$ 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

$$\lambda_i^* = \left(\frac{M + (M-1)\,\Omega^*}{(1+\Omega^*)^2}\right)\left(\frac{\pi_m}{M\beta}\right) \qquad i = 1, \ldots, M \tag{13}$$

$$\mu_j^* = \left(\frac{\Omega^*\left((1+\Omega^*)\,N - 1\right)}{(1+\Omega^*)^2}\right)\left(\frac{\pi_t}{N\gamma}\right) \qquad j = 1, \ldots, N \tag{14}$$

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

$$\Omega^3 N + (N-1)\Omega^2 - (M-1)\,z\Omega - Mz = 0. \tag{15}$$

Moreover,

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

This unique equilibrium with trading has several interesting properties. First, the aggregate monitoring intensities of both sides, $\bar{\lambda}^*$ and $\bar{\mu}^*$, are positively related. Indeed,

$$\bar{\lambda}^* = \Omega^* \bar{\mu}^*. \tag{16}$$

Thus, a shock on the parameters affecting the total monitoring intensity of one side also affects the monitoring intensity of the other side in the same direction. This change in the monitoring intensity of the other side amplifies the initial effect of the shock.

Consider for instance an increase in the number of market-makers. Using equation (16), the effect on market-makers' total monitoring is:

$$\frac{\partial \bar{\lambda}^*}{\partial M} = \frac{\partial \Omega^*}{\partial M}\bar{\mu}^* + \Omega^* \frac{\partial \bar{\mu}^*}{\partial M} \tag{17}$$

The first term on the R.H.S is positive and captures the direct effect of an increase in the number of market-makers. Namely market-makers' total monitoring is larger because 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. This amplification effect is captured by the second term on the R.H.S of equation (17).

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 (*$\frac{\partial Vol(\bar{\lambda}^*, \bar{\mu}^*)}{\partial M} > 0$ *and* $\frac{\partial Vol(\bar{\lambda}^*, \bar{\mu}^*)}{\partial N} > 0$*).*

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.

Corollary 2 *In the unique equilibrium with trading, for fixed fees of the platform,*

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* $\frac{\partial \mu_j^*}{\partial \pi_t} > 0$).

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* $\frac{\partial Vol(\bar{\lambda}^*, \bar{\mu}^*)}{\partial \pi_t} > 0$).

As explained previously, this corollary shows that a change in the cost-benefit of monitoring for one side affects the monitoring levels of both sides. For instance, an increase in the fee on market-makers, c_m, decreases the benefit per trade for market-makers, 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.[18]

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 checks 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 checks the market more frequently simply because this side has more participants.

Excess of attention by one side translates in imbalance in the duration of each phase in a liquidity cycle. For instance, if there is an excess of attention by market-makers, the average time to move from state E (scarce liquidity) to state F (full

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

liquidity) is shorter than the average time to move from state F to state E since $\frac{1}{\bar{\lambda}^*} < \frac{1}{\bar{\mu}^*}$. Interestingly, this imbalance in the reaction times to a given state of the market may enable traders to learn about changes in aggregate monitoring by one side.

4 The Determinants of the Make/Take Spread

Now, we study the determination of fees by the trading platform. As explained in Section 2.3, the objective function of the trading platform is to choose the fees (c_m, c_t) that maximize its expected profit per unit of time. That is, it chooses (c_m^*, c_t^*) such that:

$$(c_m^*, c_t^*) \in \arg\max_{c_m, c_t} \; (c_m + c_t)\, Vol(\bar{\lambda}^*, \bar{\mu}^*). \tag{18}$$

Trading fees affect traders' monitoring decisions because they change the profitability of monitoring. 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). The trading platform chooses its fees, taking into account the impact of its pricing decision on traders' monitoring intensities and the trading rate. As larger fees reduce the demand for trading, the trading platform faces the standard price-quantity trade-off for a monopolist.

We solve for the optimal fees in two steps. In the first step we fix the total fees $\bar{c} = c_m + c_t$, and ask how this total should be allocated between market makers fees (c_m) and market takers fees (c_s). In the second stage, we solve for the optimal $\bar{c}$, establishing the complete fee schedule.

To start, consider an exogenously given total fees $\bar{c}$. In this case, the problem of the trading platform is to find the fee structure (c_m, c_t) that maximizes it trading

rate. That is, it solves:

$$Max_{c_m,c_t}\ Vol(\bar{\lambda}^*,\bar{\mu}^*) = \frac{\bar{\lambda}^* \cdot \bar{\mu}^*}{\bar{\lambda}^* + \bar{\mu}^*} \tag{19}$$

$$s.t:\ c_m + c_t = \bar{c}. \tag{20}$$

The first order conditions to this problem impose that:

$$\frac{\partial Vol(\bar{\lambda}^*,\bar{\mu}^*)}{\partial c_m} = \frac{\partial Vol(\bar{\lambda}^*,\bar{\mu}^*)}{\partial c_t}. \tag{21}$$

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

$$\begin{aligned} \eta_{mm} &\equiv \frac{\partial \log(\bar{\lambda}^*)}{\partial c_m} \text{ and } \eta_{mt} \equiv \frac{\partial \log(\bar{\mu}^*)}{\partial c_m}, \\ \eta_{tm} &\equiv \frac{\partial \log(\bar{\lambda}^*)}{\partial c_t} \text{ and } \eta_{tt} \equiv \frac{\partial \log(\bar{\mu}^*)}{\partial c_t}. \end{aligned} \tag{22}$$

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. Variables η_{tm} and η_{tt} measure the elasticities of the total monitoring level of the market-making side to the fee charged on the market-takers. Using equation (21), we obtain the following result.

Proposition 3 *For each level $\bar{c}$ of the total fee charged by the platform, there is a unique allocation of this fee between the market-making side and the market-taking side that maximizes the trading rate. The optimal fee structure is obtained for c_m^* and c_t^* such that:*

$$\begin{aligned} c_m^* &= \left(\frac{h}{h+1}\right)\bar{c}, \\ c_t^* &= \bar{c} - c_m^* = \left(\frac{1}{h+1}\right)\bar{c}, \end{aligned} \tag{23}$$

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 themselves on the fees through Ω^*, π_m, and π_t. Thus, the optimal fee structure is *implicitly* defined by

equation (23). In general, we cannot obtain a closed-form solution for the optimal fee structure. In the next section, to develop intuition, we consider in details the particular case in which $M = N = 1$ in which we can characterize in closed-form the optimal fee structure. Then, in Section 4.2, we show that the insights obtained in this case are robust in the general case and we also analyze the effect of changing the number of market participants on the optimal fee structure.

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 if the market-making side pays a larger fee than the market-taking side. In general, there is no reason to expect a flat fee structure to be optimal (i.e., $h = 1/2$ in equation (23)). We first provide an example that illustrates this point.

Numerical Example: We set the parameter values at $M = N = 1$; $L = 1$; $\bar{c} = 0.1$; $\Delta = 1$; $\gamma = 2$ and $\beta = 1$. Suppose that the platform sets a flat fee of $c_m = c_t = 0.05$. In this case, Proposition 2 implies that $\lambda^* \approx 0.088$ and $\mu^* \approx 0.06$. As a consequence the trading rate is 0.0389 trades per unit of time. Calculations also reveal that in this case the sensitivity of trading volume to the market-making fee (i.e., $\frac{\partial Vol(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_m}$) is -0.98 and the elasticity of trading volume to the market-taking fee (i.e., $\frac{\partial Vol(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_t}$) is -1.23. Thus, by slightly raising its fee on market-makers and by decreasing its fee on market-takers by the same amount, the platform can raise the trading rate without changing its total revenue of \$0.1 per trade.

Thus, for these numerical values, charging the same fee on both sides is not optimal for the platform. The optimal fee structure is in fact such that $c_m \approx 0.088$ and $c_t \approx 0.011$, so that the market-making side is charged more than the market-taking side (see Proposition 4 below). For this fee structure, the trading rate is 0.039 trades per unit of time (an increase of 0.5%). The monitoring intensity of the market-making side is $\lambda^* \approx 0.085$ and the monitoring intensity of the market-taking side is 0.072. At the optimal fee structure, there is still excess attention by market-makers (because they have lower cost of monitoring) but less than in the flat fee structure.■

4.1 The Case $M = N = 1$

When $M = N = 1$, the solution to equation (15) is $\Omega^* = z^{\frac{1}{3}}$. Thus, using equations (13) and (14), we obtain the monitoring intensities of the market-making side and the market-taking side:

$$\lambda_1^* = \frac{1}{\left(1 + z^{\frac{1}{3}}\right)^2} \cdot \left(\frac{\pi_m}{\beta}\right) \tag{24}$$

$$\mu_1^* = \frac{1}{\left(1 + z^{-\frac{1}{3}}\right)^2} \cdot \left(\frac{\pi_t}{\gamma}\right) \tag{25}$$

Using these expressions, we can solve for the optimal total fee charged by the trading platform ($\bar{c}^*$) and the optimal breakdown of this fee between the market-making and the market taking sides. We obtain the following result.

Proposition 4 *Let $r \equiv \frac{\gamma}{\beta}$. 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:*

$$c_m^* = \frac{1}{2}\left(\Delta - \frac{2(L - \bar{c})}{(1 + r^{\frac{1}{4}})}\right) \quad \textit{and } c_t^* = \bar{c}^* - c_m^*. \tag{26}$$

For these fees, $z = r^{\frac{3}{4}}$,

$$\pi_m^* = \frac{L - \bar{c}}{(1 + r^{\frac{1}{4}})} \quad \textit{and} \quad \pi_t^* = \frac{L - \bar{c}}{(1 + r^{-\frac{1}{4}})}, \tag{27}$$

and the equilibrium monitoring intensities are:

$$\lambda_1^* = \frac{L - \bar{c}}{\beta\left(1 + r^{\frac{1}{4}}\right)^3} \quad \textit{and} \quad \mu_1^* = \frac{L - \bar{c}}{\gamma\left(1 + r^{-\frac{1}{4}}\right)^3}. \tag{28}$$

Moreover, if the trading platform is a for-profit monopolist, it sets $\bar{c}$ at $\bar{c}^* = \frac{L}{2}$.

The first part of the proposition (Equation (26)) gives the optimal allocation of the total trading fee between the market-making side and the market-taking side. It is worth stressing that this allocation is the same for any level $\bar{c}$ of the fee charged by the trading platform as long as it is less than L (otherwise trading volume is nil). As

explained previously, this allocation maximizes the trading rate for a fixed revenue $\bar{c}$ per trade for the trading platform.

This means that the results regarding the allocation of the trading charge between the market-making and the market-taking sides are not driven by the trading platform's market power. Rather, price discreteness is key. Indeed, it prevents market-makers from fully neutralizing a change in fees by an adjustment in their offers. For instance, market-makers cannot fully pass-through a decrease in their trading fee by quoting a more attractive price because their quotes must be on the grid. Thus, the trading platform can use trading fees to control traders' monitoring intensities and achieve higher trading rates, for any fixed level of the total trading fee.

Of course, the level of the fee chosen by the trading platform depends on its monopolistic position. It is half the total gains from trade (L) for all parameter values. The trading platform cannot extract all the gains from trade because as $\bar{c}$ enlarges, market-makers and market-takers watch the market less closely. As a result, the demand for trading decreases.

Let $\overline{\Delta} \overset{def}{=} 2(L-\bar{c})(1+r^{\frac{1}{4}})^{-1}+\bar{c}$. Using equation ((26)), it is immediate that the make/take spread increases in the tick size and is zero if and only if $\Delta = \overline{\Delta}$. Thus, to maximize the trading rate, market-makers should pay a larger fraction of the fee charged by the trading platform when the tick size enlarges. Intuitively, when the tick size is large, a market-maker has a high incentive to inspect the state of the market since it gets a high fraction of the gains from trade. But, by symmetry, this means that a market-taker's incentive to inspect the state of the market is small. As a consequence, trading opportunities are missed because the market-taker does not inspect frequently the state of the market. In this case, there is excess attention by market-makers. Thus, it is optimal for the platform to allocate a larger fraction of the total fee to the market-maker so as to better balance the monitoring intensities of both sides. If the tick size is small, the symmetric argument applies: market-makers do not inspect the market sufficiently frequently. Thus, in absence of monetary inducements, the market appears illiquid to market-takers, and again trading opportunities are lost.

Interestingly, when the tick size is small enough, the trading platform optimally subsidizes market-makers (i.e., $c_m^* < 0$).[19] As pointed out in the introduction, this implication of the model matches the fact that the practice of subsidizing limit orders in U.S. markets coincide with a sharp reduction in the tick size of these markets. Moreover, the tick size on the trading platforms that pioneered this practice in the U.S (trading platforms such Archipelago or Island) was initially much smaller than the incumbent markets.[20]

The model also implies that the make-take spread increases in the ratio of market-takers to market-makers' monitoring cost ($r = \frac{\gamma}{\beta}$). Thus, other things equal, market-makers contribute relatively more to the total fee per transaction when their monitoring costs per unit of time decreases in relative terms. Thus, the make/take spread observed in reality may reflect the fact that monitoring costs is relatively higher for the market-making side. One possible reason is that institutional investors who need to take a position in a list of stocks 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 the market-taking side.

The monitoring costs ratio also determines how the gains from trade are split between each side in equilibrium. To see this, let $\Phi^* \equiv \frac{\pi_m^*}{\pi_m^* + \pi_t^*}$ be the fraction of the *net* gains from trade ($L - \bar{c}$) obtained by the market-maker in each transaction. Using equation (27) in Proposition 4, we obtain that in equilibrium:

$$\Phi^* = \frac{1}{1 + r^{\frac{1}{4}}}. \tag{29}$$

We deduce the following result.

Corollary 4 *In the case $M = N = 1$.*

[19] In this case, one may wonder whether it is not optimal for a market-maker to undercut its 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 $a - \Delta - v_0 - c_m^* = \frac{L}{1+r^{1/4}} - \Delta < 0$ since $L < \frac{3}{2}\Delta$.

[20] 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.

1. *The market-maker (resp. market-takers) gets a smaller fraction of the net gains from trade as the ratio of her monitoring cost to market-taker's monitoring cost declines ($\frac{\partial \Phi^*}{\partial r} < 0$).*

2. *When the market-maker's monitoring cost goes to zero ($\beta \to 0$), her fraction of net gains from trade goes to zero ($\Phi \to 0$).*

3. *When the market-taker's monitoring cost goes to zero ($\gamma \to 0$), his fraction of net gains from trade goes 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 4 shows that the development of algorithmic trading results in a counter-intuitive redistribution of trading profits. Namely, in the long run, the side whose monitoring cost declines the most appropriates a smaller fraction of the gains from trade. Indeed, in the short-run (that is, for fixed fees of the platform), a decline in the monitoring cost benefits to each side. In the long run, however, trading fees adjust. Through this adjustment, the benefit of a lower monitoring cost for one side is in fact passed to the other side.

Corollary 5 *Assume $M = N = 1$. In equilibrium, when the trading platform optimally allocates its total fee between market-makers and market-takers, the trading rate is*

$$Vol^*(\beta, \gamma, L) = \frac{(L - \bar{c})}{\left(\beta^{\frac{1}{4}} + \gamma^{\frac{1}{4}}\right)^4}.$$

Thus, the trading rate is inversely related to traders' monitoring cost and positively related to the size of gains from trade. It is independent from the tick size.

Thus, the model implies that algorithmic trading (a reduction in monitoring costs) enlarges the trading rate.

4.2 The effect of the number of participants on the trading fees

We now turn to the analysis of the optimal fees for an arbitrary number of participants on each side. First, we consider the particular case in which (i) the number of participants on both sides are identical and (ii) all traders bear the same monitoring cost. In this case, we say that the market-making side and the market-taking sides are symmetric.

Proposition 5 *When $N = M$ and $\gamma = \beta$, it is optimal for the trading platform to break its total fee $\bar{c}$ between market-makers and market-takers as follows:*

$$c_m^* = \frac{\Delta}{2} - \frac{(L - \bar{c})}{2} \tag{30}$$

$$c_t^* = \bar{c} + \frac{L - \bar{c}}{2} - \frac{\Delta}{2}. \tag{31}$$

The associated monitoring levels in equilibrium are:

$$\lambda_k^* = \mu_k^* = \frac{(2N - 1)(L - \bar{c})}{8N} \qquad k = 1, ..., N$$

Moreover, in this case, the trading platform optimally charge a total fee $\bar{c}^* = \frac{L}{2}$.

Thus, in this case, the fees charged on each side are independent of the number of participants. Moreover, as in the case analyzed in the previous section, the make-take spread $(c_m^* - c_t^*)$ increases in the tick size. In fact, the pricing strategy of the platform consists in allocating the total fee between the two sides in such a way that the benefit of monitoring are equalized $(\pi_m^* = \pi_t^*)$. In this way, both sides have equal attention to the trading process $(\lambda_k^* = \mu_k^*)$.

This solution is quite natural since both sides are symmetric. When there is an imbalance in the number of participants on each side, we have not been able to derive the fees chosen by the trading platform in closed-form. However, we can prove the following result.

Proposition 6 *Let $r \equiv \frac{\gamma}{\beta}$. For a fixed number of participants on the market-taking side, the fee charged on market-takers (resp. market-makers) increases (resp. de-*

creases) in the number of market-takers. That is:

$$\frac{\partial c_m^*(N,M,r)}{\partial N} < 0 \text{ and } \frac{\partial c_m^*(N,M,r)}{\partial M} > 0 \quad (32)$$

$$\frac{\partial c_t^*(N,M,r)}{\partial N} > 0 \text{ and } \frac{\partial c_t^*(N,M,r)}{\partial N} < 0 \quad (33)$$

When $N = M$ and $\frac{\gamma}{\beta} = 1$ the trading chooses its fee structure so as to equalize the benefit of monitoring for the market-making side and the market-making side. When $M > N$, this solution is not optimal anymore since it results in excessive attention by the market-making side. In this situation, as implied by Corollary 6, it is optimal for the trading platform to raise its fee on the market-making side and decrease its fee on the market-taking side so as to rebalance the attention of both sides.

Figures 1 and 2 in the appendix illustrate the previous result for specific parameter values, namely $\Delta = 1, L = 1.5, \beta = \gamma = 1$. The total fee is set at $\bar{c} = 0.75$. In the baseline case $N = M = 10$. Figure 1 shows how the fees charged on market-takers and market-makers change as the number of market-takers increases from $N = 10$ to $N = 1000$. As implied by Proposition 6, the make-take spread becomes smaller as the number of market-takers enlarges. When N is large enough, it is optimal for the trading platform to subsidize the side that is relatively "short" in participants (here market-makers). Thus, as the number of market-takers increases, the fraction of the gains from trade captured by market-takers get smaller.

Figure 2 shows how the trading rate varies with the number of market-takers. It also compares the trading rate when the trading platform optimally allocates its total fee between the two sides and when it follows the simpler, but suboptimal, policy of splitting equally its fee between the market-makers and the market-takers.

5 Conclusion

This paper considers a model in which traders must monitor the market to seize trading opportunities. One group of traders ("market-makers") specialize in posting quotes while another group of traders ("market-takers") specialize 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.

6 Appendices

6.1 Appendix A

	Tape A Make Fee	Tape A Take Fee	Tape B Make Fee	Tape B Take Fee	Tape C Make Rebate	Tape C Take Fee
AMEX	-30	30	-30	30	-30	30
	-25	25	-25	25	-25	25
BATS	-24	25	-30	25	-24	25
CBSX	-26	29	-26	29	-26	29
CHX	26	29	-32	29	-26	29
EDGX	-25	30	-25	30	-25	30
	-29	26	-29	26	-29	26
EDGA	0	30	0	30	0	30
		0		0		0
IES	-35	30	-17	15	-32	30
	-35				-35	
LavaFlow	-24	26	-24	26	-24	26
Nasdaq	-20	30	-20	30	-20	30
	-28	29	-31	29	-28	29
NSX	-26	30	-30	30	-26	30
		25		25		25
NYSE	0	8				
NYSE Arca	-25	30	-20	30	-20	26
	-28	29		28	-26	24.5
PHLX	-22	30	-2	30	-22	30
Track	-22	23	-22	23	-22	23
Tradebook	-20	25	-20	25	-20	25

Table 1: This table gives the fees per share (in cents for 100 shares) charged by U.S. equity trading platforms for executed limit orders (Make Fee, columns 2, 4 and 6) and market orders (Take Fee, columns 3, 5 and 7), as of July 2008. A minus sign indicates that the fee is a rebate. Source: Traders Magazine, July 2008.

6.2 Appendix B

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

Proof of Proposition 2: From (11), 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 $i=1,\ldots M$, we obtain

$$\frac{\bar{\mu}\left(\left(\bar{\mu}+\bar{\lambda}\right)M-\bar{\lambda}\right)}{\left(\bar{\lambda}+\bar{\mu}\right)^2}\frac{\pi_m}{\beta}=\bar{\lambda}. \tag{34}$$

Similarly, for market-takers we obtain,

$$\frac{\bar{\lambda}\left(\left(\bar{\mu}+\bar{\lambda}\right)N-\bar{\mu}\right)}{\left(\bar{\lambda}+\bar{\mu}\right)^2}\frac{\pi_t}{\gamma}=\bar{\mu}. \tag{35}$$

Let $\Omega \equiv \frac{\bar{\lambda}}{\bar{\mu}}$. Dividing (34) and (35) by $\bar{\mu}^2$ we have,

$$\frac{M+(M-1)\,\Omega}{(1+\Omega)^2}\frac{\pi_m}{\beta} = \bar{\lambda}. \tag{36}$$

$$\frac{\Omega\left((1+\Omega)\,N-1\right)}{(1+\Omega)^2}\frac{\pi_t}{\gamma} = \bar{\mu} \tag{37}$$

Dividing these two equations gives,

$$\frac{(M+(M-1)\,\Omega)}{\Omega^2\left((1+\Omega)\,N-1\right)}z=1, \tag{38}$$

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

$$\Omega = g(\Omega, M, N, z). \tag{39}$$

with

$$g(\Omega, M, N, z)=\frac{(M-1)z}{\Omega N}+\frac{Mz}{N\Omega^2}-\frac{N-1}{N}. \tag{40}$$

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, (39) 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 (36) and (37). 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:

$$\Omega^* = g(\Omega^*, M, N, z), \tag{41}$$

where $g(.)$ is defined in equation (40). It is immediate that $g(.)$ increases in M, decreases in N, and increases in z. As $g(.)$ decreases in Ω, we have

$$\frac{\partial \Omega^*}{\partial M} > 0, \tag{42}$$

$$\frac{\partial \Omega^*}{\partial N} < 0. \tag{43}$$

Now, using Equations (42) and (13), we conclude that:

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

Similarly, using equations (43) and (14), we deduce that

$$\frac{\partial \mu_j^*}{\partial M} > 0. \tag{44}$$

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

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

Thus, using equations (42) and (43), we conclude that $\frac{\bar{\lambda}^*}{\bar{\mu}^*}$ increases in M and decreases in N. Equation (44) 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 (15), 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 Proposition 3: We have

$$\begin{aligned} \frac{\partial Vol(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_m} &= -Vol(\bar{\lambda}^*, \bar{\mu}^*)^2 \left(\frac{\partial \overline{\lambda}}{\partial c_m} \frac{1}{\overline{\lambda}^2} + \frac{\partial \overline{\mu}}{\partial c_m} \frac{1}{\overline{\mu}^2}\right) \\ &= -\frac{Vol(\bar{\lambda}^*, \bar{\mu}^*)^2}{c_m} \left(\frac{\eta_{mm}}{\overline{\lambda}} + \frac{\eta_{mt}}{\overline{\mu}}\right). \end{aligned} \tag{45}$$

and

$$\frac{\partial Vol(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_t} = -\frac{Vol(\bar{\lambda}^*, \bar{\mu}^*)^2}{c_t} \left(\frac{\eta_{tm}}{\overline{\lambda}} + \frac{\eta_{tt}}{\overline{\mu}}\right). \tag{46}$$

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 (45) and (46), 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 4: Consider the case $M = N = 1$

Step 1: Optimal fee structure for each level of $\bar{c}$. 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_{mm} and π_{mt}. Thus, instead of using c_m and c_t as the decision variables of the platform, we can use π_{mm} and π_{mt}. It turns out that this is easier. Thus, for a fixed $\bar{c}$, we rewrite the platform problem as:

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

From equations (36) and (37),

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

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:

$$Max_{\pi_m,z} \frac{\lambda_1^*}{1+z^{\frac{1}{3}}}\bar{c} \tag{47}$$

$$s.t \quad \pi_m\left(1+\frac{\gamma}{\beta z}\right) = L - \bar{c}. \tag{48}$$

$$\text{and } \lambda_1^* \;=\; \frac{L-\bar{c}}{\beta\left(1+z^{\frac{1}{3}}\right)^2\left(1+\frac{\gamma}{\beta z}\right)} \tag{49}$$

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

$$z = \left(\frac{\gamma}{\beta}\right)^{\frac{3}{4}} = r^{\frac{3}{4}}. \tag{50}$$

Using the constraint (48), we have,

$$\pi_m^* = \frac{L-\bar{c}}{1+r^{\frac{1}{4}}}. \tag{51}$$

It follows that,

$$\pi_t^* = L - \bar{c} - \pi_m = \frac{L - \bar{c}}{1 + r^{-\frac{1}{4}}}. \tag{52}$$

Then, plugging (50), (51), and (52) into equations (24) and (25), we obtain the required expressions for λ_1^* and μ_1^*.

Step 2: Optimal $\bar{c}$. Now we determine the optimal level of its fee by the trading platform. Using the expressions for λ_1^* and μ_1^* given in equation (28), the expected profit of the trading platform is:

$$\Pi_E = \frac{\lambda_1^* \mu_1^*}{\lambda_1^* + \mu_1^*} \bar{c} = \frac{(L - \bar{c})\bar{c}}{\gamma(1 + r^{-\frac{1}{4}}) + \beta(1 + r^{\frac{1}{4}})}.$$

Thus, it is immediate that the optimal level of $\bar{c}$ for the trading platform is $\bar{c} = L/2$. ■

Proof of Corollary 4 Immediate from equation (29). ■

Proof of Corollary 5 Immediate from the definition of the trading rate and the expressions for λ_1^* and μ_1^* given in equation (28). ■

Proof of Proposition 5: To establish the proposition, we must show that the fees given in Proposition 5 solves equation (23). We first observe that in general:

$$\begin{aligned}
\eta_{mm} &= c_m \left(-\frac{1}{\pi_m} + \frac{\partial \Omega^*}{\partial c_m} \left(\frac{M-1}{M + (M-1)\Omega^*} - \frac{2}{(1+\Omega^*)} \right) \right) \\
\eta_{mt} &= c_m \left(\frac{\partial \Omega^*}{\partial c_m} \left(\frac{N + 2N\Omega^* - 1}{\Omega^*((1+\Omega^*)N - 1)} - \frac{2}{(1+\Omega^*)} \right) \right) \\
\eta_{tt} &= c_t \left(-\frac{1}{\pi_t} + \frac{\partial \Omega^*}{\partial c_t} \left(\frac{N + 2N\Omega^* - 1}{\Omega^*((1+\Omega^*)N - 1)} - \frac{2}{(1+\Omega^*)} \right) \right) \\
\eta_{tm} &= c_m \left(\frac{\partial \Omega^*}{\partial c_t} \left(\frac{M-1}{M + (M-1)\Omega^*} - \frac{2}{(1+\Omega^*)} \right) \right)
\end{aligned}$$

In Proposition 5, the fees charged on market-makers and market-takers is such that $\pi_m = \pi_t$ and therefore $z = 1$ since $\gamma = \beta$. Moreover, since $N = M$, we have $\Omega^* = 1$ for these fees. Thus, the fees given in the proposition are such that $\bar{\lambda}^* = \bar{\mu}^*$. But then, using the previous expressions, we obtain:

$$\bar{\mu}^* \eta_{mm} + \bar{\lambda}^* \eta_{mt} = -\frac{\bar{\mu}^* c_m^*}{\pi_m}$$

and

$$\bar{\mu}^* \eta_{mm} + \bar{\lambda}^* \eta_{mt} = -\frac{\bar{\mu}^* c_t^*}{\pi_t}$$

But since $\pi_m = \pi_t$, we deduce that:

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

Thus, the fees given in Proposition 5 satisfy Equation (23). Thus, they are optimal for the trading platforms when market-makers and market-takers are symmetric. ■

Proof of Proposition 6: *to be written.*

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

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

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

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

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

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

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

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

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

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

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

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

Figure 1: Make-Take Spread and Number of Market-Takers

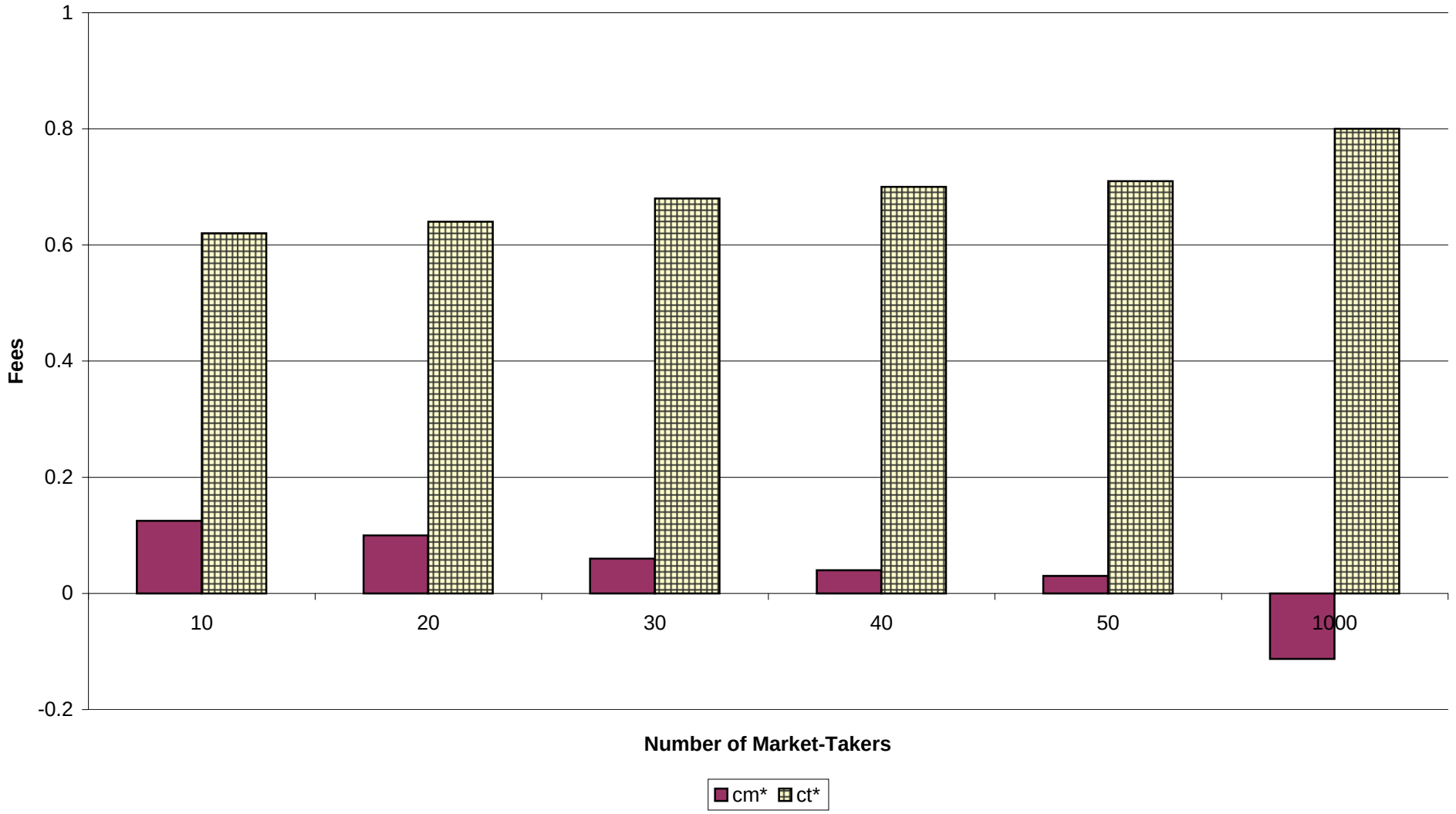
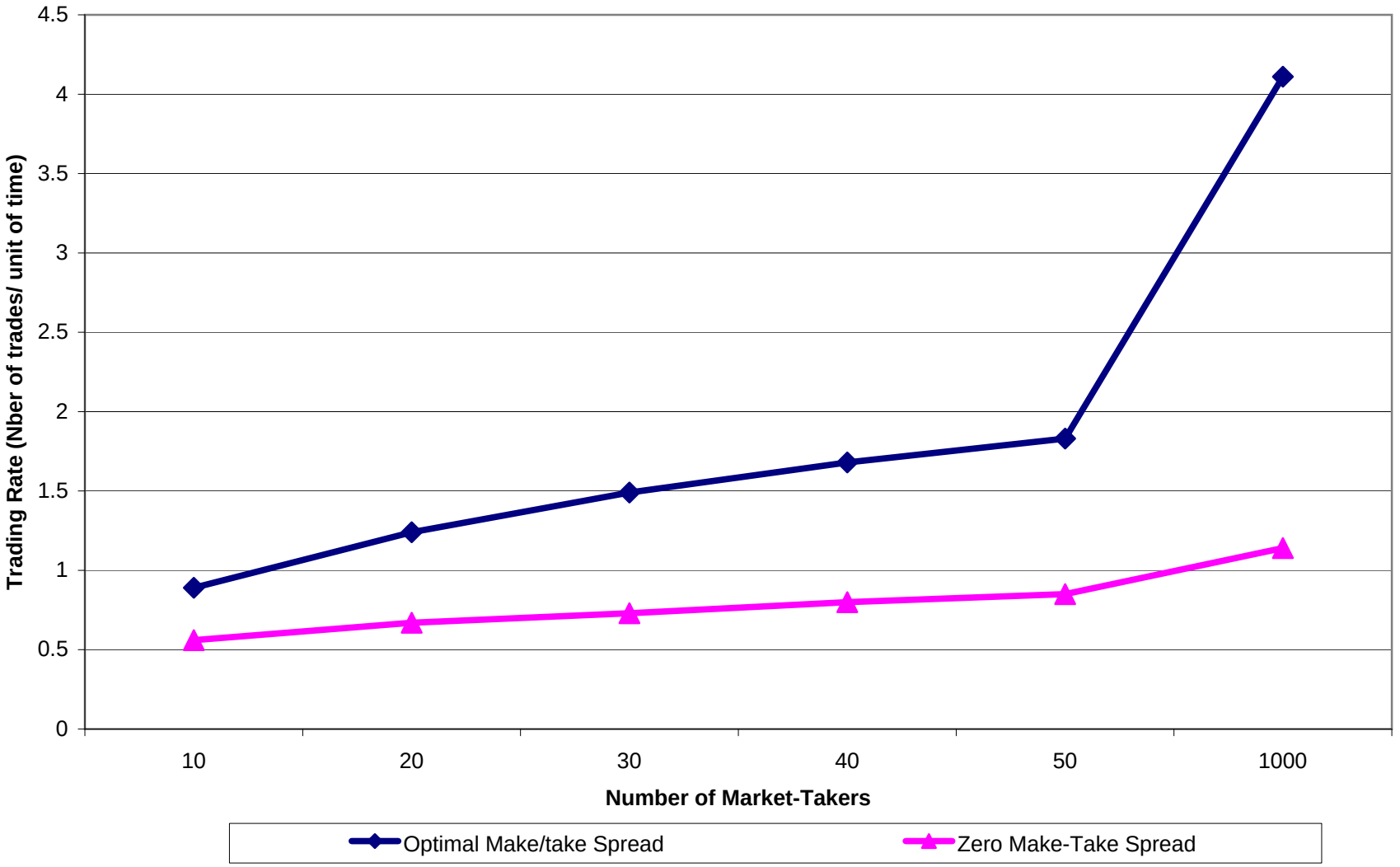


Figure 2


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
Trading Rate (Nber of trades/ unit of time)
10
20
30
40
50
1000
Number of Market-Takers
Optimal Make/take Spread
Zero Make-Take Spread