Limit Order Book as a Market for Liquidity[1]

Thierry Foucault
HEC School of Management
1 rue de la Liberation
78351 Jouy en Josas, France
Tel: 33-1-39679411
foucault@hec.fr

Ohad Kadan
School of Business Administration
Hebrew University,
Jerusalem, 91905, Israel
Tel: 972-2-5883232
ohad@mscc.huji.ac.il

Eugene Kandel
School of Business Administration
and Department of Economics
Hebrew University,
Jerusalem, 91905, Israel
Tel: 972-2-5883137
mskandel@mscc.huji.ac.il

July 10, 2001

[1] We would like to thank David Easley, Frank de Jong, Larry Glosten, Larry Harris, Pete Kyle, Leslie Marx, Narayan Naik, Maureen O Hara, Christine Parlour, Patrik Sandas, Ilya Strebulaev, Avi Wohl, and seminar participants at Amsterdam, Emory, Illinois, Insead, Jerusalem, LBS, Tel Aviv, Wharton for their helpful comments and suggestions. Comments by participants at the WFA 2001 meeting and the Gerzensee Symposium 2001 have been helpful as well. The authors thank J. Nachmias Fund, and Kruger Foundation for ¯nancial support.

Abstract

We develop a dynamic model of an order-driven market populated by discretionary liquidity traders. These traders must trade, yet can choose the type of order and are fully strategic in their decision. Traders differ by their impatience: less patient traders demand liquidity, more patient traders provide it. Three equilibrium types are obtained - the type is determined by three parameters: the degree of impatience of the patient traders, which we interpret as the cost of execution delay in providing liquidity; their proportion in the population, which determines the degree of competition among the liquidity providers; and the tick size, which is the cost of the minimal price improvement. Despite its simplicity, the model generates a rich set of empirical predictions on the relation between market parameters, time to execution, and spreads. We argue that the economic intuition of this model is robust, thus its main results will remain in more general models.

1 Introduction

Limit and market orders constitute the core of any order-driven continuous trading system (such as the NYSE, London Stock Exchange, Euronext, Tokyo and Toronto Stock Exchanges, as well as all the ECNs).[1] A market order guarantees an immediate execution at the best price available at the moment of the order arrival at the exchange. In general, a market order represents demand for liquidity (immediacy of execution). With a limit order, a trader can improve his execution price relative to the market order price, but the execution is neither immediate, nor certain. A limit order represents supply of liquidity to future traders.[2]

The optimal order choice ultimately involves a tradeoff between the cost of a delayed execution and the cost of immediate execution, which (for small transactions) is determined by the size of the inside spread. Intuitively we expect patient traders to post limit orders and supply liquidity to impatient traders, who opt for market orders. In his seminal paper Demsetz (1968) stresses the limit orders as the source of liquidity, pointing out the trade off between longer execution time and better prices. He states (p.41):

"Waiting costs are relatively important for trading in organized markets, and would seem to dominate the determination of spreads."

He conjectures that more aggressive limit orders will be submitted to gain priority in execution and shorten the expected time-to-execution. Moreover, he anticipates that the active securities should have lower spreads because the competition from limit orders will be fiercer in light of shorter waiting times. In this paper we explore the interactions between traders' impatience, order placements strategies and waiting times in the context of a dynamic order-driven market.

Our model features buyers and sellers arriving sequentially. Each trader wants

[1] Domowitz (1993) shows that over 30 important financial markets in the world in the early 90's had some of order-driven market features in their design. The importance of order-driven markets around the world has been steadily increasing since.

[2] We ignore here marketable limit orders.

to buy or sell one unit of a security. We assume that these are liquidity traders, i.e. they will buy/sell regardless of price. However, they choose between market and limit orders so as to minimize their cost of trading. Upon arrival, the traders decide to place a market order or a limit order, conditional on the state of the book. If submitting a limit order the trader chooses a price and bears the opportunity cost of postponing the trade.

Under several simplifying assumptions we are able to develop a recursive method for calculating the order placements strategies and the expected time-to-execution for limit orders. In general, in equilibrium, patient traders provide liquidity to impatient traders. We identify 3 types of equilibria characterized by markedly different dynamics for the limit order book. These dynamics turn out to be very sensitive to the ratio of the proportion of patient traders to the proportion of impatient traders. Actually the larger is this ratio, the more intense is competition among liquidity suppliers. They are also influenced by the dispersion of waiting costs across traders. Some of our main findings can be summarized as follows.

- Limit orders time-to-execution are large when the proportion of patient traders is relatively large. This effect enhances competition among liquidity providers who submit more aggressive orders to shorten their time-to-execution. Hence markets with a relatively large proportion of patient traders feature smaller spreads.

- In order to speed up execution, traders frequently find optimal to undercut or outbid the best quotes by more than one tick. This happens when (i) the proportion of patient traders is relatively large, (ii) waiting costs are large or (iii) the tick size is small.

- A decrease in the tick size can result in *larger* expected spreads. Actually it gives the possibility to traders to quote less competitive prices by expanding the set of prices. If competition among liquidity providers is weak, they use the new prices and the average spread increases.

- A decrease in the order arrival rate can result in *smaller* expected spreads. Intuitively, such a decrease extends the expected time-to-execution for limit orders. This effect induces liquidity suppliers to place more aggressively priced limit orders when the inside spread is large.

In some limit order markets, designated market-makers are required to enter bid and ask quotes in the limit order book. This is the case, for instance, in the Paris Bourse for medium and small capitalization stocks.[3] We consider the effect of introducing this type of trader in our model. We show that the presence of a trader who monitors the market and occasionally submits limit orders, can significantly alter the equilibrium. His intervention forces patient traders to submit more aggressive offers in order to speed up execution and hence narrows the spreads. This result provides important guidance for market design.

Our results contribute to the growing literature on limit order markets. Most of the models in the theoretical literature are focused on the optimal bidding strategies for limit order traders (see e.g. Glosten (1994), Chakravarty and Holden (1995), Rock (1996), Seppi (1997), Biais, Martimort and Rochet (2000), Parlour and Seppi (2001)). These models do not analyze the choice between market and limit orders and are static. For this reason they do not describe the interactions between impatience, time-to-execution and order placement strategies as we do in this paper.

Parlour (1998) and Foucault (1999) study dynamic models. Parlour (1998) shows how the order placement decision is influenced by the depth available at the inside quotes. Foucault (1999) analyzes the impact of the risk of being picked off and the risk of non execution on traders' order placement strategies. In both models, limit order traders do not bear waiting cost. Hence time-to-execution does not influence

[3] In the Paris Bourse, the designated market-makers are required to post bid-and ask quotes for a minimum number of shares and their spread cannot exceed 5% of the stock price.

traders' bidding strategies in these models whereas it plays a central role in the present article.[4]

We are not aware of other theoretical papers in which prices and time-to-execution for limit orders are jointly determined in equilibrium. Time-to-execution, however, is an important dimension of market quality in limit order markets (see SEC 1997). Lo, McKinlay and Zhang (2001) estimate various econometric models for the time-to-execution of limit orders. Some of their findings are consistent with our results, e.g. the expected time-to-execution increases with the distance between the limit price and the mid-quote. Our model also generates new predictions that could be tested with data on actual time-to-execution for limit orders. For instance we show that the average time-to-execution (across all limit orders) depends on (i) the tick size, (ii) the order arrival rate and (iii) the proportion of patient traders.[5] Biais, Hillion and Spatt (1995) describe the interactions between the size of the inside spread and the order flow.[6] They observe that limit order traders quickly improve the inside spread when it is large. In our model the amount by which a limit order trader undercuts or outbids the best offers depends on (i) the inside spread, (ii) the proportion of patient traders and (iii) the order arrival rate. These findings provide guidance for empirical studies of limit order markets.[7]

The paper is organized as follows. Section 2 describes the model. Section 3 derives the equilibrium of the limit order market and provides examples. In Section

[4] A few authors suggest other approaches to modeling the limit order book. This includes Angel (1994), Domowitz and Wang (1994) and Harris (1995) who consider models with exogenous order flow. Using queuing theory, Domowitz and Wang (1994) analyze the stochastic properties of the book. Angel (1994) and Harris (1995) study how the optimal choice between market and limit orders varies according to different market conditions (e.g. the state of the book, the rate of order arrival...). We use more restrictive assumptions than these authors. But these assumptions enable us to endogenize the order flow and the time-to-execution for limit orders.

[5] Lo et al. (2001) report that there is a large variation in mean time-to-execution across stocks. According to our model, these variations can be explained by the fact that stocks differ with respect to trading activity or tick size.

[6] See also Benston, Irvine and Kandel (2001).

[7] Empirical analyses of limit order markets include Goldstein and Kavajecz (2000), Handa and Schwartz (1996), Harris and Hasbrouck (1996), Hollifield, Miller and Sandas (2001a,b), Kavajecz (1999) and Sandas (2000).

4 we explore the effect of a change in tick size and a change in traders' arrival rate on measures of market quality. Section 5 presents some extensions. Section 6 concludes. All proofs (except for Proposition 1) are in the Appendix.

2 Model

2.1 Timing and Market Structure

Consider a continuous market for a single security, organized as a limit order book without intermediaries. We assume that latent information about the security value determines the range of admissible prices, however the transaction price itself is determined by traders who submit market and limit orders.[8] Specifically, at price A outside investors stand ready to sell an unlimited amount of security, thus the supply at A is infinitely elastic. We also assume that there exists an infinite demand for shares at price B ($B < A$). Moreover, A and B are constant over time. These assumptions assure that all the prices in the limit order book are in the range $[B, A]$.[9] The goal of this model is to investigate the behavior of the limit order book and transaction prices within this interval. This behavior is determined by the supply and demand of liquidity, or in other words by optimal submission of market and limit orders.

This is an infinite horizon model with discrete time periods. At the beginning of every period a trader arrives at the market and observes the limit order book. Each trader must buy or sell one unit of the security. These liquidity traders have a discretion on which type of order to submit. Each trader can submit a market order to ensure an immediate trade at the best quote available at the time. Alternatively, he can submit a limit order, which improves the price, but delays the execution. We assume that traders' waiting costs are proportional to the time they have to wait until

[8] We discuss this modelling strategy below.

[9] A similar assumption is used in Seppi (1997) and Parlour and Seppi (2001).

completion of their transaction. Hence traders face a trade-off between the execution price and the time-to-execution when they choose between market and limit orders. In contrast with Admati and Pfleiderer (1988) or Parlour (1998), traders are not required to carry their desired transaction by a deadline.

All prices (but not waiting costs and traders' valuations) are placed on a discrete grid. The tick size, which is chosen by the exchange designer, is denoted by $\Delta > 0$. All the prices in the model are expressed in terms of integer multiples of Δ. We denote by a and b the best ask and bid quotes when a trader comes to the market. The *inside spread* at that time is $s := a - b$. Given the setup we know that $a \le A$, $b \ge B$, and $s \le K := A - B$.[10]

Both buyers and sellers can be of two types which differ by the size of their waiting costs. Type 1 traders (the patient type) incur an opportunity cost of d_1 for an execution delay of one period. Type 2 traders (the impatient type) incur a cost of d_2 ($0 \le d_1 < d_2$). The proportion of patient traders in the population is denoted by θ ($0 < \theta < 1$). Patient types can be thought as institutions building up positions, or other long-term investors. Arbitragers or brokers conducting agency trades are examples of impatient traders.

Limit orders are stored in the limit order book and are executed in sequence according to *price priority* (e.g. sell orders with the lowest offer are executed first). For tractability, we make the following simplifying assumptions about the market structure.

A.1: Each trader arrives only once, submits a market or a limit order and exits. Submitted orders cannot be cancelled or modified.

A.2: Traders who submit limit orders must narrow the spread by at least one tick.

[10] Notice that a, b, s, A, B, K and all other spreads and prices that follow are positive integers. This is so since we use integer multiples of the tick size, Δ, instead of dollar prices and dollar spreads. Furthermore the model does not require time subscripts on variables, thus they are omitted for brevity.

A.3: Buyers and sellers alternate with certainty, e.g. first a buyer arrives, then a seller, then a buyer, and so on. The first trader is a buyer with probability 0.5.

Assumption A.1 implies that traders in the model do not adopt active trading strategies which may involve repeated submissions and cancellations. These active strategies require market monitoring, which is costly (e.g. because liquidity traders' time is valuable). The second assumption implies that limit order traders cannot queue at the same price (note however that they queue at different prices since limit orders do not drop out of the book). With this assumption, the inside spread is the only state variable which influences traders' order placement strategies. This greatly simplifies the description and the characterization of traders' order placement strategies. This assumption is less restrictive than it may appear. In Section 6, we show that we can dispense with assumption A2 if patient traders' waiting cost is large enough. The third assumption facilitates the computation of traders' expected waiting time and is imperative to keep the model tractable (see Section 3.1. for a discussion).

Let p_b and p_s be the prices paid by buyers and sellers, respectively. In our model, as in Admati and Pfleiderer (1988) for instance, traders do not have the option not to trade. Thus their only decision is a choice of strategy resulting in a trade. A buyer can either pay the lowest ask a or submit a limit order which creates a new inside spread with size j. In a similar way, a seller can either receive the largest bid b or submit a limit order which creates a new inside spread with size j. This choice determines the execution price:

$$p_b = a - j;\ p_s = b + j \text{ with } j \in \{0, ..., s - 1\},$$

where $j = 0$ represents a market order. It is convenient to consider j (rather than p_b or p_s) as the trader's decision variable. For brevity, we say that a trader uses a "*j-limit order*" when he posts a limit order which creates a spread with size j. The expected time-to-execution of a j-limit order is denoted by $T(j)$. Since the waiting costs are assumed to be linear in waiting time, the expected waiting cost of a j-limit order is $d_iT(j)$, $i \in \{1,2\}$. As a market order entails immediate execution, we set $T(0) = 0$.

We assume that traders are risk neutral. The expected profit of trader i ($i \in \{1,2\}$) who submits a j-limit order is:

$$\Pi_i(j) = \begin{cases} V_b - p_b\Delta - d_iT(j) = (V_b - a\Delta) + j\Delta - d_iT(j) & \text{if trader } i \text{ is a buyer} \\ p_s\Delta - V_s - d_iT(j) = (b\Delta - V_s) + j\Delta - d_iT(j) & \text{if trader } i \text{ is a seller} \end{cases}$$

where V_b, V_s are buyers' and sellers' valuations, respectively. To justify this classification to buyers and sellers, we assume that $V_b >> A\Delta$, and $V_s << B\Delta$.[11] Expressions in parenthesis represent profits associated with market order submission. These profits are determined by the trader's valuation and the best quotes when he submits his market order. It is immediate that the optimal order placement strategy when the inside spread has size s solves the following optimization problem, for buyers and sellers alike:

$$\max_{j\in\{0,\ldots s-1\}} \pi_i(j) := j\Delta - d_iT(j). \tag{1}$$

We will show that $T(j)$ is non-decreasing in j, in equilibrium. Hence a better execution price (larger value of j) is obtained at the cost of a larger expected waiting time.

A strategy for a trader is a mapping that assigns a j-limit order, $j \in \{0, ..., s-1\}$, to every possible spread $s \in \{1, ..., K\}$. Thus, a strategy determines which order to submit given the size of the inside spread. At the beginning of the game we set: $a = A$ and $b = B$ hence $s = K$. Let $o_i(.)$ be the order placement strategy of a trader with type i. A trader's optimal strategy depends on future traders' actions since they determine his expected waiting time, $T(\cdot)$. Consequently a *subgame perfect equilibrium* of the trading game is a pair of strategies, $o_1^*(.)$ and $o_2^*(.)$, such that the order prescribed by each strategy for every possible inside spread solves Program (1)

[11] Traders' valuations for the security can consist of common and idiosyncratic components as in Foucault (1999) or Hollifield, Sandas and Miller (2001a,b).

when the expected waiting time $T(¢$ is computed using the fact that traders follow strategies $o_1^*(.)$ and $o_2^*(.)$.[12]

2.2 Discussion

It is worth stressing that we abstract from the effects of asymmetric information and information aggregation. This is a marked departure from the "canonical model" in theoretical microstructure literature, surveyed in Madhavan (2000), and requires some motivation.

In most market microstructure models, quotes are determined by agents who have no reason to trade, and either trade for speculative reasons, or make money providing liquidity. For these *value-motivated traders*, the risk of trading with a better informed agent is a concern and affects the optimal order placement strategies. In contrast, in our model, traders have a non-information motive for trading and are precommited to trade. The risk of adverse selection is not an issue for these *liquidity traders*. Rather, they determine their order placement strategy with a view at minimizing their transaction cost and balance the cost of waiting against the cost of obtaining immediacy in execution.[13] In order to focus on this trade-off in the simplest way, we propose a framework that allows for a simple dichotomy between "macro" information-based asset pricing and market "micro"structure. We assume that information-related considerations determine the price range, rather than the price itself. The equilibrium in the market for liquidity provision determines quotes inside this range. At this stage we do not model the determination of this range, but rather assume that it exists. For fixed income securities these boundaries are quite natural, given the existence of close substitutes. In case of equities we conjecture that this price range represents the consensus among all analysts/investors, yet is not

[12] The rules of the game, as well as all the parameters are assumed to be common knowledge among all the traders.

[13] Harris (1998) and Glosten (2000) also argue that optimal order placement strategies are different for liquidity traders and value-motivated traders.

subject to arbitrage (see Shleifer and Vishny 1997).

The trade-off between the cost of immediate execution and the cost of delayed execution may be relevant for value-motivated traders as well. However, it is very difficult to solve dynamic models with asymmetric information among traders who can strategically choose between market and limit orders. In fact we are not aware of such dynamic models.[14]

3 Equilibrium Patterns

In this section we characterize the equilibrium strategies for each type of trader. For given values of the parameters, the equilibrium is unique. We also calculate the stationary probability distribution of the inside spread in equilibrium. The dynamics of the order flow and the distribution of the inside spread depend on (i) the proportion of patient traders relative to the proportion of impatient traders and (ii) the difference in waiting costs between patient and impatient traders. This leads us to distinguish between three different types of equilibria. We provide examples which illustrate the attributes of each one of the three equilibrium types.

3.1 Expected Waiting Time

In order to characterize the equilibrium, we first analyze the behavior of the expected waiting time function $T(j)$. Suppose the trader arriving this period chooses a j-limit order. We denote by $\alpha_k(j)$ the probability that the trader arriving next period and observing an inside spread with size j chooses a k-limit order, $k \in \{0, 1, ..., j-1\}$.[15] Clearly $\alpha_k(j)$ depends on traders' strategies and

$$\sum_{k=0}^{j-1} \alpha_k(j) = 1, \ \forall j = 1, ..., K-1.$$

[14] Chakravarty and Holden (1995) consider a single period model in which informed traders can choose between a market and a limit orders. Glosten (1994) or Biais et al. (2000) consider limit order markets with asymmetric information but do not allow traders to choose between market and limit orders.

[15] Recall that $k = 0$ stands for a market order.

Assumption A.2 implies that a trader who faces a one tick spread submits a market order. Consequently, the time-to-execution for a 1-limit order is one period, i.e. $T(1) = 1$. Next, we establish a general recursive formula for the expected waiting time function. This formula links the expected waiting time function to traders' order placement strategies (described by the $\alpha s'$).

Lemma 1 *If $\alpha_0(j) > 0$, the expected waiting time for the execution of a j-limit order is given by the following recursive formula:*

$$T(j) = \frac{1}{\alpha_0(j)}\left[1 + \sum_{k=1}^{j-1} \alpha_k(j)T(k)\right] \quad \forall j = 2, ..., K - 1 \text{ and } T(1) = 1 \tag{2}$$

Two extreme cases are worth emphasizing. The first is when no trader submits a market order when he faces a spread with size j^*. In this case $\alpha_0(j^*) = 0$ and the expected waiting time of a j-limit order, with $j \geq j^*$, is infinite. Such limit orders will never be submitted in equilibrium, since they are dominated by a market order. Hence, in equilibrium, the expected waiting time of limit orders is always finite. This implies that limit orders execute with certainty.[16] The second case is when all traders submit a market order when they face a spread with size j^{**}. In this case $T(j^{**}) = 1$. It will become apparent that no spreads smaller than j^{**} and larger than j^* can be observed in equilibrium. In between, there is a variety of cases in which some traders find it optimal to submit limit orders, while others submit market orders.

Assumption A.3 is used to obtain the expected waiting function (Eq.(2)). The alternation of buyers and sellers yields a simple ordering of the queue of unfilled limit orders (the book): a j-limit order cannot be executed before j'-limit orders where $j' < j$. This is of course true when we consider two buy or two sell limit orders because of price priority. Without A.3, this would not be true however if the j-limit order and the j'-limit order are in opposite direction (a buy order and a sell order for instance). The ordering implied by A.3 explains why the expected waiting time

[16] However, execution may take place after a very long time. In fact, in any finite time interval, the execution probability of a j-limit order is strictly smaller than 1, if $T(j) > 1$.

has a simple recursive structure. Without this recursive structure, it becomes very difficult to compute the expected waiting time function and the model is (in general) intractable.

3.2 Equilibrium strategies

Although the trading game has an infinite horizon, the nodes with one-tick spread serve as end-nodes in the usual finite game trees, since everybody submit a market order. Thus we can solve the game by *backward induction.* To see this point, consider a trader who arrives in the market when the size of the inside spread is $s = 2$. The trader has two choices: either he submits a market order or a 1-limit order. The latter improves his execution price by one tick compared to a market order but results in a one period delay in execution. Choosing the best action for each type of trader, we determine $\alpha_k(2)$ (for $k = 0$ and $k = 1$). If $\alpha_0(2) = 0$, the expected waiting time for a 2-limit order is infinite. It follows that no spread larger than one tick can be observed in equilibrium. If $\alpha_0(2) > 0$,we compute $T(2)$ (using Eq.(2)). Then we proceed to $s = 3$ and so forth. This inductive approach is the key to most results in the paper.

Three results follow immediately. First, as this is a game of perfect information an equilibrium in pure strategies always exists. Second, since this is a one-play game for each trader, there are no Nash equilibria (in pure strategies) other than the sub-game perfect equilibria that we trace by backward induction. And third, the equilibrium is unique for any tie-breaking rule. We choose the following rule. If a trader is indifferent between a j_1-limit order and a j_2-limit order, with $j_1 < j_2$, he submits the limit order with the smallest spread (in this case the j_1-limit order).

We proceed by proving results that characterize the equilibrium. Traders submit limit orders only if they can cover their waiting cost. Since limit orders wait at least one period, there is a spread below which a trader strictly prefers to use market orders. We refer to this spread as being the trader's *"reservation spread"* and we denote it j_i^R for trader i ($i \in \{1, 2\}$). This the smallest spread trader i is willing to establish with a limit order, and still the associated expected profit is greater than zero (dominates a market order). In order to give a formal definition of the reservation spread, let $int(x)$ be the largest integer smaller than or equal to x. The reservation spread of trader i is:[17]

$$j_i^R := int(\frac{d_i}{\Delta}) + 1 \qquad i \in \{1, 2\} \tag{3}$$

Clearly, the reservation spread of a patient trader cannot exceed that of an impatient one, however the two can be equal. The latter case yields the first equilibrium type for all values of other parameters. We say that the two trader types are *indistinguishable* if they possess the same reservation spreads: $j^R := j_1^R = j_2^R$. Intuitively, traders are indistinguishable if the two waiting costs fall into the same cell on the grid: $[0, \Delta), [\Delta, 2\Delta), [2\Delta, 3\Delta), \ldots$.

Proposition 1 *Suppose traders' types are indistinguishable ($j_1^R = j_2^R = j^R$) then, in equilibrium all traders submit a market order if $s \leq j^R$ and submit a j^R-limit order if $s > j^R$.*

The proof of Proposition 1 is simple and intuitive hence we present it here instead of relegating it to the Appendix. Consider a trader who arrives in the market when the inside spread is $s > j^R$. If he submits a j-limit order with $j^R < j$ then the next trader submits a j^R-limit order given the specification of traders' strategies. This implies that $\alpha_0(j) = 0$ (i.e. the waiting time is infinite) for $j^R < j$. Therefore a j-limit order with $j^R < j$ cannot be optimal since it is never executed. If the trader submits a j^R-limit order, his order is cleared by the next trader. By definition of the reservation spread, this choice dominates a market order. This establishes that when the inside spread is larger than traders' reservation price, the optimal strategy is to submit a j^R-limit order. Finally consider a trader who arrives in the market when the spread is $s \leq j^R$. By definition of the reservation spread, the submission of a market

[17] A trader who submits a limit order waits a least one period before execution. Hence the smallest waiting cost for a trader with type i is d_i. It follows that the smallest spread trader i can establish is the smallest integer, j_i^R, such that $\pi_i(j_i^R) = j_i^R \Delta - d_i > 0$. This remark yields Eq.(3).

order is a dominant strategy for this trader. This completes the proof of Proposition 1.

The equilibrium with indistinguishable traders is characterized by an oscillating pattern. The first, as well as every odd-numbered trader afterwards, submits a limit order which creates a spread with size j^R. The second, and every even-numbered trader afterwards, submits a market order. The inside spread oscillates between K and j^R and transactions take place only when the spread is small. Trade prices are either $A - j^R$ if the first trader is a buyer, or $B + j^R$, if the first buyer is a seller. The outcome is competitive in the sense that limit order traders always quote their reservation spread, that is the spread such that they just cover their waiting cost.[18]

After characterizing the first type of equilibrium, we proceed by assuming that traders are *heterogeneous*: $j_1^R < j_2^R$. Given two spreads $j_1 < j_2$ we denote by $\langle j_1, j_2 \rangle$ the set: $\{j_1, j_1 + 1, j_1 + 2, ..., j_2\}$, i.e. the set of all possible spreads between j_1 and j_2 (inclusive). In particular, the range of all possible spreads is $\langle 1, K \rangle$.

Proposition 2 *Suppose traders are heterogeneous ($j_1^R < j_2^R$). In equilibrium there exists a cutoff spread $s_c \in \langle j_2^R, K \rangle$ such that:*

1. *Given a spread $s \in \langle 1, j_1^R \rangle$, patient and impatient traders submit a market order.*

2. *Given a spread $s \in \langle j_1^R + 1, s_c \rangle$, a patient trader submits a limit order and an impatient trader submits a market order.*

3. *Given a spread $s \in \langle s_c + 1, K \rangle$, patient and impatient traders submit a limit order.*

The proposition shows that when $j_1^R < j_2^R$, the state variable s (the inside spread) is partitioned into three regions: (i) $s \le j_1^R$, (ii) $j_1^R < s \le s_c$ and (iii) $s > s_c$. The

[18] Observe that the tick size determines the resolution of traders' categories. The larger is the tick size - the more traders with differing waiting costs are pooled together into the same equilibrium strategies. Conversely observe that d_1 and d_2 may be arbitrarily close and still fall into different cells of the grid if the tick size is sufficiently small.

reservation spread of the patient trader, j_1^R, represents the smallest spread observed in the market. At the other end s_c is the largest quoted spread in the market. Limit orders which create a larger spread have an infinite waiting time since no traders submit a market order when the inside spread is larger than s_c. Hence these limit orders are never submitted. This observation permits us to restrict our attention to cases where $s_c = K$, for brevity. This equality holds true when the cost of waiting for an impatient trader is sufficiently large.[19] Under this condition impatient traders always demand liquidity (submit market orders), while patient traders supply liquidity (submit limit orders) when the inside spread is larger than their reservation spread.

Proposition 3 *Suppose $s_c = K$. Any equilibrium exhibits the following structure: there exist q spreads, $n_1 < n_2 < ... < n_q$, with $n_1 = j_1^R$, $n_q = K$ and $2 \leq q \leq K$, such that the optimal order submission strategy is as follows:*

- *An impatient trader submits a market order, for any spread in $[1, K]$.*
- *A patient trader submits a market order when he faces a spread in $[1, n_1]$ and submits a n_h-limit order when he faces a spread in $[n_h + 1, n_{h+1}]$ for $h = 1, ..., q - 1$.*

Hence when a patient trader faces an inside spread with size $n_{h+1} > j_1^R$, he responds by submitting a limit order which improves upon the inside spread by $(n_{h+1} - n_h)$ ticks. This order establishes a new inside spread equal to n_h. When the inside spread is K, it takes a streak of $q - 1$ patient traders to bring the inside spread to the competitive level j_1^R. Hence q determines the maximal number of limit orders which can be observed in the book. We refer to q as *the length of the book*. A small length of the book means that patient traders quickly make good offers since it takes a few patient traders to bring the spread to the competitive level.

[19] For instance, $s_c = K$ if $j_2^R \geq K$. It is worth stressing that this condition is sufficient but not necessary. In Examples 2 and 3 below, j_2^R is much smaller than K but $s_c = K$.

Next we analyze the expected waiting time in equilibrium. Let $r := \frac{\theta}{1-\theta}$ be the ratio of the proportion of patient traders to the proportion of impatient traders. Intuitively, when this ratio is smaller (larger) than 1, liquidity is consumed more (less) quickly than it is supplied. As we show below, this ratio determines traders' bidding strategies and time-to-execution for limit orders.

Proposition 4 *The expected waiting time function in equilibrium is given by:*

$$T(n_1) = 1 \quad \textit{and} \quad T(n_h) = 1 + 2\sum_{k=2}^{h} r^{k-1} \quad \textit{for} \quad h = 2, ..., q - 1,$$

and

$$T(j) = T(n_h) \quad \forall j \in \langle n_{h-1} + 1, n_h \rangle.$$

Clearly the expected waiting time function (weakly) increases with j. Hence the larger is the distance between the price of a limit order and the mid-quote, the larger is the expected waiting time for the order. This result is consistent with the evidence in Lo, McKinley and Zhang (2001).

Another determinant of the expected waiting time is the proportion of patient traders relative to the proportion of impatient traders, r. The intuition is as follows. Notice that h determines the priority status of a limit order in the queue of unfilled limit orders. Actually an n_h-limit order can not be executed before $n_{h'}$-limit orders have been executed if $h' < h$ (when these orders are present in the book, of course). When r increases, the likelihood of a market order decreases. It follows that the expected waiting time for the h^{th} limit order in the queue enlarges. It turns out that the rate of increase in the waiting time from one limit order to the next in the queue of limit orders depends on r as well. Actually when $r > 1 (r < 1)$ the marginal expected waiting time $T(n_h) - T(n_{h-1})$ is non-decreasing (non-increasing) in h. In this case, we say that $T(\cdot)$ is "convex" ("concave") in h. The next corollary summarizes these remarks.

Corollary 1 *The expected waiting time of the h^{th} limit order in the queue of limit orders increases with r, the ratio of the proportion of patient traders to the proportion of impatient traders. The expected waiting time function is "convex" when $r > 1$, and "concave" when $r < 1$.*

We show below that these properties of the expected waiting time function influence traders' bidding strategies. In the next proposition we express the spreads on the equilibrium path, i.e. $n_1, n_2, .., n_q$, in terms of the exogenous parameters. Define $\Psi_h := n_h - n_{h-1}$ for $h \geq 2$ as *the spread improvement,* when the inside spread has a size equal to n_h. The spread improvement is the number of ticks by which a trader narrows the spread when he submits a limit order. The larger is the spread improvement, the more aggressive is the limit order.

Proposition 5 *The set of equilibrium spreads is given by:*

$$\begin{aligned} n_1 &= j_1^R;\ n_q = K, \\ n_h &= n_1 + \sum_{k=2}^{h} \Psi_k \quad h = 2, ..., q - 1; \end{aligned}$$

where

$$\Psi_h = int(2r^{h-1}\frac{d_1}{\Delta}) + 1$$

and the length of the book, q is the smallest integer such that:

$$j_1^R + \sum_{k=2}^{q} \Psi_k \geq K. \tag{4}$$

The previous proposition shows that whenever, $2d_1r^{h-1} \geq \Delta$, a limit order trader finds optimal to undercut or to outbid the best prices by more than one tick ($\Psi_h > 1$). Biais, Hillion and Spatt (1995) observe that liquidity suppliers frequently improve upon the best quotes by several ticks. Our result identifies four determinants for the spread improvement which could be considered in future empirical investigation.

These determinants are: (i) the proportion of patient traders, r, (ii) the per period waiting cost, d_1 (iii) the tick size, Δ, and (iv) the inside spread. We analyze each of these determinants in turn.

When r increases, the time-to-execution for a given position in the queue of limit orders becomes larger. Hence, *other things equal*, liquidity suppliers bear larger waiting costs ($d_1 T$). Traders react by submitting more agressive orders to preempt good positions in the queue of limit orders and thereby reduce their time-to-execution. The same effect operates when d_1 increases. In this case, traders bear larger waiting costs because the per-period waiting cost is larger. The smaller is the tick size, the smaller is the cost of improving upon the best bid and ask prices. Thus a smaller tick results in larger spread improvements in terms of ticks.

The spread improvement, Ψ_h, increases (decreases) with h when $r > 1$ ($r < 1$). This means that when $r > 1$ the spread improvement increases with the size of the inside spread, while the opposite is true when $r < 1$. The intuition is as follows. Consider the $(h - 1)^{th}$ trader in the queue of unfilled limit orders . This trader's time to execution is $T(n_{h-1})$ instead of $T(n_h)$ for the trader behind him in the queue. Hence the difference in expected waiting cost between the h^{th} and the $(h - 1)^{th}$ positions in the queue of limit orders is equal to $(T(n_h) - T(n_{h-1}))d_1$. Intuitively, this should be the "price" of acquiring the $(h - 1)^{th}$ position instead of the h^{th}position in the queue. The dollar spread improvement plays the role of this price and, for this reason, it is approximately equal to $(T(n_h) - T(n_{h-1}))d_1$.[20] This shows that the shape of the waiting time function determines the relationship between the spread improvement and the inside spread. When $r > 1$, the waiting time function is convex in h. Hence liquidity suppliers offer larger spread improvements when the spread is large. When $r < 1$, the waiting time function is concave and liquidity suppliers offer larger spread improvements when the spread is small.

[20] In fact observe that $\Psi_h \Delta \simeq 2r^{h-1}d_1 = (T(n_h) - T(n_{h-1}))d_1$. The dollar spread improvement is only approximately equal to the difference in waiting cost because the set of prices is discrete.

Notice that when spread improvements are larger than 1 tick, the traders do not make use of all the possible prices in equilibrium. This implies that the limit order book features “holes”, i.e. cases in which the distance between two consecutive ask or bid prices is larger than one tick.[21]

The last part of the previous proposition (Eq.(4)) implies that that the length of the book decreases when spread improvements get larger. Actually, limit order traders improve on the best quotes by a larger number of ticks so that a smaller number of prices on the grid are used. This means that more competitive outcomes are expected when the length of the book is small. This is the case in particular when $r \geq 1$ because (a) spread improvements are large and (b) liquidity is not consumed too quickly (which leaves time for the inside spread to narrow). For this reason we call the equilibrium when $r \geq 1$ a *High Competition (HC) Equilibrium* and the equilibrium when $r < 1$, a *Low Competition (LC) Equilibrium.* Using this terminology, we classify all equilibria in three categories described in Table 1.

Table 1 - Three equilibrium patterns

Equilibrium pattern	Description	Specification
Oscillating	Indistinguishable Traders Spreads oscillate between K and j^R.	$j_1^R = j_2^R;\ \forall r$
HC	Heterogeneous traders High level of competition among liquidity providers “Convex” time function	$j_1^R < j_2^R\ ;\ r \geq 1$
LC	Heterogeneous traders Low level of competition among liquidity providers “Concave” time function	$j_1^R < j_2^R\ ;\ r < 1$

In the next sections, we show that (i) the stationary probability distribution of spreads and (ii) the impact of a change in the tick size are strikingly different in HC and LC equilibria.

[21] Holes in the limit order book is a phenomenon documented by several empirical studies: Biais, Hillion and Spatt (1995) - Paris Bourse; Goldstein and Kavajecz (2000) - NYSE; Hollifield, Miller, and Sandas (2001a) - Stockholm; Benston, Irvine, and Kandel (2001) - Toronto; and Kandel, Lauterbach, and Tkach (2000) - Tel Aviv.

3.3 Examples

We illustrate the three equilibrium patterns by numerical examples. The tick size is $\Delta = \$0.125$. The lower price bound of the book is set to $B\Delta = \$20$, and the upper bound is set to $A\Delta = \$22.5$. Thus, the maximal spread is $K = 20$ ($K\Delta = \$2.5$). The parameters that differ across the examples are presented in Table 2.

Table 2: Three Examples

	Example 1 (Oscillating)	**Example 2 (HC)**	**Example 3 (LC)**
d_1	0.15	0.10	0.10
d_2	0.20	0.25	0.25
θ	Any value	0.55	0.45

Table 3 presents the equilibrium strategy for patient (type 1) and impatient (type 2) traders in each example. Each entry in the table presents the optimal limit order (in terms of ticks) given the current spread (0 stands for a market order).[22]

Table 3 - Equilibrium strategies

[22] The equilibrium strategies in Examples 2 and 3 follow from the formulae given in Proposition 5.

Current	Example 1		Example 2		Example 3	
Spread	**Type 1**	**Type 2**	**Type 1**	**Type 2**	**Type 1**	**Type 2**
1	0	0	0	0	0	0
2	0	0	1	0	1	0
3	2	2	1	0	1	0
4	2	2	3	0	3	0
5	2	2	3	0	3	0
6	2	2	3	0	5	0
7	2	2	6	0	6	0
8	2	2	6	0	7	0
9	2	2	6	0	8	0
10	2	2	9	0	9	0
11	2	2	9	0	10	0
12	2	2	9	0	11	0
13	2	2	9	0	12	0
14	2	2	13	0	13	0
15	2	2	13	0	14	0
16	2	2	13	0	15	0
17	2	2	13	0	16	0
18	2	2	13	0	17	0
19	2	2	18	0	18	0
20	2	2	18	0	19	0

Order Placement Strategies

Table 3 reveals the qualitative differences between the three equilibrium types. In Example 1, $j_1^R = j_2^R = 2$, thus patient and impatient traders are indistinguishable. The inside spread oscillates between the maximal spread of 20 ticks and the reservation spread of 2 ticks. In Example 2 and 3, the traders are heterogeneous since $j_1^R = 1$ and $j_2^R = 3$. In Example 2, the inside spreads on the equilibrium path are (in terms of ticks): $\{1, 3, 6, 9, 13, 18, 20\}$. Spreads of other sizes will not be observed.[23] In Example 3, the inside spreads on the equilibrium path are (in terms of ticks): $\{1, 3, 5, 6, 7, ..., 20\}$. In these two examples, transactions can take place at spreads which are strictly larger than patient traders' reservation spreads. However, traders place much more aggressive limit orders in Example 2 where $r > 1$. In fact spread improvements are larger than one tick for all spreads on the equilibrium path in this

[23] Table 3 specifies actions for spreads on and off the equilibrium path. This is necessary for a full specification of the equilibrium strategy.

case. In contrast, in Example 3, spread improvements are equal to one tick in most cases. Hence the market will appear more competitive in Example 2 ($r > 1$) than in Example 3 ($r < 1$).

Expected Waiting Time

The expected waiting time function in Examples 2 and 3 is illustrated in Figure 1.This figure presents the expected waiting time of a limit order as a function of the spread it creates. In both examples the expected waiting time increases when we move from one reached spread to the next one, while it is constant over the spreads which are not posted in equilibrium. The expected waiting time is smaller at any spread in Example 3. This explains the differences in bidding strategies in Examples 2 and 3. When $r < 1$, limit order traders are less aggressive because they expect a faster execution.

Book Dynamics

Figure 2 illustrates the book resulting from 40 rounds of simulation making independent draws from the distribution of traders' type. We use the same realizations for Examples 2 and 3 and look at the dynamics of the limit order book.

As is apparent from Figure 2, the inside spread converges more quickly towards small levels in Example 2 than in Example 3. Since the type realizations in both books are identical, this observation is only due to the fact that patient traders use more aggressive limit orders, in order to speed up execution, in Example 2. If the type realizations were not held constant, there would be a second force acting in the same direction. When r is larger than 1, the liquidity offered by the book is *consumed less rapidly* than when r is smaller than 1. This means that the likelihood of a market order arriving while the spread is large is smaller when $r > 1$. This effect would reinforce the fact that spreads tend to be smaller in Example 2. We prove this point more formally in the next section by deriving the probability distribution of the inside spread.

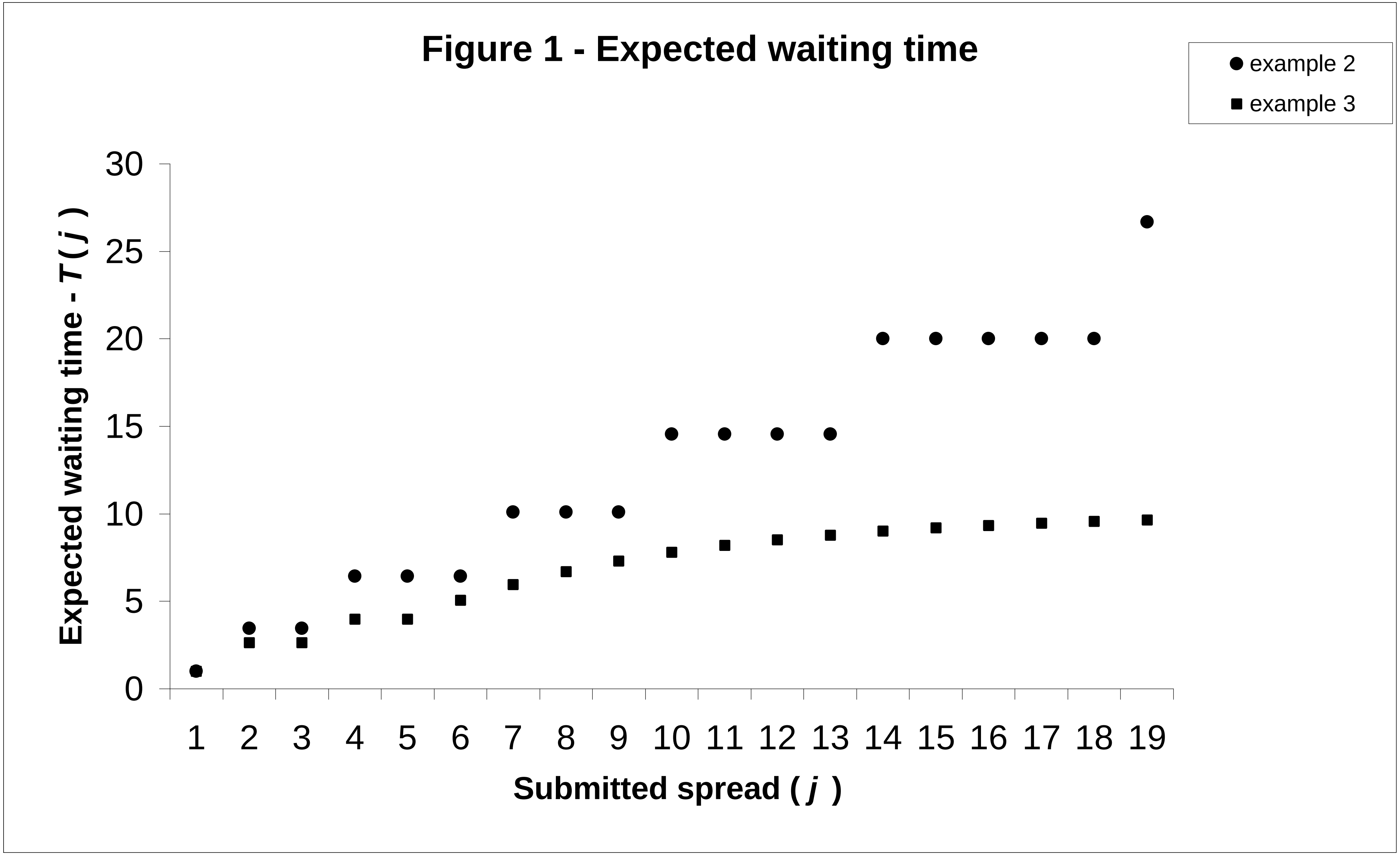
Figure 1 - Expected waiting time
example 2
example 3
Expected waiting time - T(j)
Submitted spread (j)
0
5
10
15
20
25
30
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

Figure 2 - Book Simulation (same realizations of type arrivals for two examples)

Example 2 - Intense competition among liquidity suppliers (*r* = 1.222)

Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Trader	B2	S1	B1	S2	B2	S2	B1	S1	B2	S1	B2	S1	B1	S1	B2	S1	B1	S1	B1	S2	B1	S1	B1	S1	B2	S2	B1	S2	B1	S1	B1	S2	B2	S1	B2	S2	B2	S1	B1	S2
22 1/2	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s
22 3/8																																								
22 1/4		s	s	s																																				
22 1/8																																								
22																																								
21 7/8								s		s		s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s		s	s	s
21 3/4																																								
21 5/8																																								
21 1/2														s		s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s		s						
21 3/8																																								
21 1/4																		s				s		s						s										
21 1/8																	b	b	b		b	b	b	b	b		b		b	b	b									
21																																								
20 7/8																																								
20 3/4													b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b				b	
20 5/8			b																																					
20 1/2																																								
20 3/8																																								
20 1/4							b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b
20 1/8																																								
20	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b

Example 3 - Low level of competition among liquidity suppliers (*r* = 0.818)

Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Trader	B2	S1	B1	S2	B2	S2	B1	S1	B2	S1	B2	S1	B1	S1	B2	S1	B1	S1	B1	S2	B1	S1	B1	S1	B2	S2	B1	S2	B1	S1	B1	S2	B2	S1	B2	S2	B2	S1	B1	S2
22 1/2	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s
22 3/8		s	s	s				s		s		s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s
22 1/4														s		s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s
22 1/8																		s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s
22																						s	s	s	s	s	s	s	s	s	s	s	s	s	s	s		s	s	s
21 7/8																								s						s	s	s		s						
21 3/4																																								
21 5/8																																								
21 1/2																																								
21 3/8																																								
21 1/4																																								
21 1/8																																								
21																																								
20 7/8																																								
20 3/4																															b									
20 5/8																							b	b	b		b		b	b	b	b	b	b	b				b	
20 1/2																			b		b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b
20 3/8																	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b
20 1/4													b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b
20 1/8			b				b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b
20	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b

Legend:

B1 - Patient buyer, B2 - Impatient buyer, S1 - Patient seller, S2 - Impatient seller

b - a buyers limit order, s - a sellers limit order.

3.4 Distribution of Spreads

We have so far established the structure of equilibrium strategies. Our next step is to derive the probability distribution of spreads induced by these strategies. In this way, we show that small spreads are more frequent in markets where $r > 1$. This formalizes the intuition that competition in these markets is more intense.We also use the distribution of spreads in order to calculate measures of market quality in the next section.

When $j_1^R = j_2^R = j^R$ the spread oscillates between K and j^R. Thus, the ex-ante probability of each one of these two spreads is 0.5. Now consider the case in which traders are heterogeneous, i.e. $j_1^R < j_2^R$. From Proposition 3 we know that an equilibrium can be described by q spreads: $n_1 < n_2 < ... < n_q$. A patient trader submits a n_{h-1}-limit order when the inside spread has size n_h $(h = 2, ..., q)$ and a market order when he faces a spread of size n_1. An impatient trader always submits a market order (we maintain the assumption that $s_c = K$). Thus, if the inside spread has size n_h $(h = 2, ..., q - 1)$ the probability that its size becomes n_{h-1} in the next period is θ, and the probability that its size becomes n_{h+1} in the next period is $1 - \theta$. If the size of the inside spread is n_1 all the traders submit market orders and its size becomes n_2 with certainty. If the size of the inside spread is K then it remains unchanged with probability $1 - \theta$ (a market order) or it decreases to n_{q-1} with probability θ (a limit order). Hence the inside spread is a finite Markov chain with $q \geq 2$ states. The $q \times q$ transition matrix of this Markov chain, denoted by W, is:

$$W = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ \theta & 0 & 1-\theta & \cdots & 0 & 0 \\ 0 & \theta & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1-\theta \\ 0 & 0 & 0 & \cdots & \theta & 1-\theta \end{pmatrix}$$

The j^{th} entry in the h^{th} row of this matrix gives the probability that the size of the inside spread becomes n_j conditional on the inside spread having size n_h $(h, j = 1, ..., q)$. A stationary distribution of this Markov chain, may be regarded as the long term probability distribution of the inside spreads.[24] We denote the stationary probabilities by $u_1, ...u_q$, where u_h is the probability of an inside spread with size n_h.

Lemma 2 *The Markov chain given by W has a unique stationary distribution. The stationary probabilities are given by:*

$$u_1 = \frac{\theta^{q-1}}{\theta^{q-1} + \sum_{i=2}^{q} \theta^{q-i}(1-\theta)^{i-2}}, \tag{5}$$

$$u_h = \frac{\theta^{q-h}(1-\theta)^{h-2}}{\theta^{q-1} + \sum_{i=2}^{q} \theta^{q-i}(1-\theta)^{i-2}} \qquad h = 2, ..., q \tag{6}$$

Using this result, Figure 3 depicts the stationary distribution in Examples 2 and 3. The distribution of spreads is skewed toward higher spreads in Example 3 (where $r < 1$). In contrast, it skewed toward lower spreads in Example 2 (where $r > 1$). This observation is easily explained by considering the expressions for the stationary probabilities. For $h, h' \in \{2, 3, ..., q\}$ with $h > h'$, the previous lemma implies that

$$\begin{aligned} \frac{u_h}{u_{h'}} &= r^{h'-h}, \\ \frac{u_h}{u_1} &= \frac{1}{r^{h-1}(1-\theta)}, \end{aligned}$$

which yields the following proposition.

Proposition 6 *For a given tick size and given values of the waiting costs*

1. *If $r < 1$ (LC equilibrium) , $u_h > u_{h'}$ for $1 \leq h' < h \leq q$. This means that the distribution of spreads is skewed towards higher spreads when $r < 1$.*

2. *If $r > 1$ (HC equilibrium) , $u_h < u_{h'}$ for $2 \leq h' < h \leq q$.[25] This means that the distribution of spreads is skewed towards lower spreads when $r > 1$.*

[24] See Feller (1968).

[25] The inequality, $u_h < u_{h'}$, does not necessarily hold for $h' = 1$; even if $r > 1$. Actually the smallest inside spread can only be reached from higher spreads while other spreads can be reached from both directions ($n_q = K$ can be reached either from n_{q-1} or from n_q itself). This implies that the probability of observing the smallest possible spread is relatively small for all values of r.

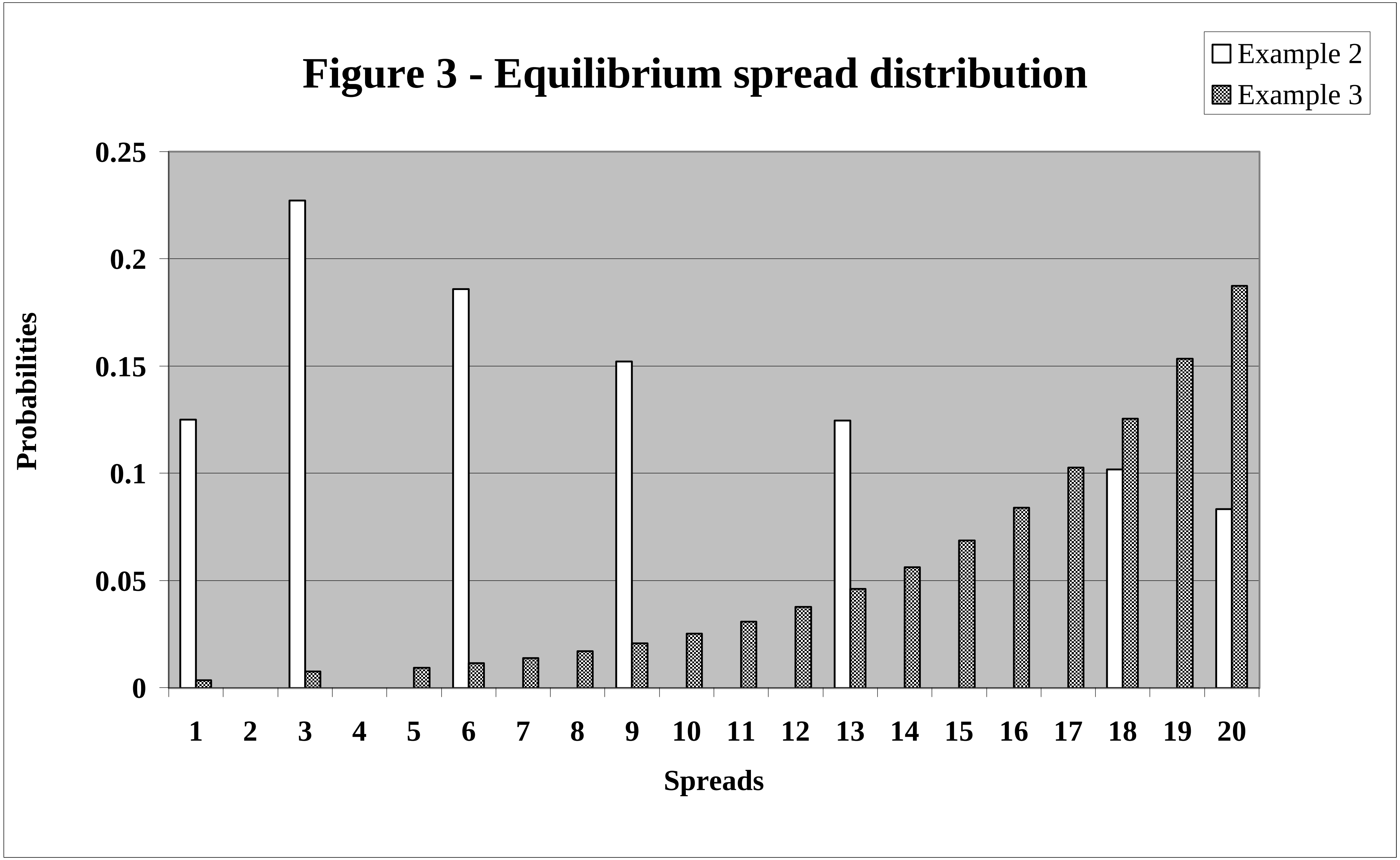
Figure 3 - Equilibrium spread distribution
Example 2
Example 3
Probabilities
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Spreads

Proposition 6 establishes that small spreads are relatively more (less) frequent than large spreads in markets with intense (low) competition between liquidity providers (we explained the intuition in the previous section). This results in a smaller average spread in markets where $r > 1$ compared to markets where $r < 1$. For instance, the expected spread in Example 2 is 8.4 ticks and the standard deviation is 6.2 ticks. In Example 3 the expected spread is 16.04 ticks and the standard deviation is 3.9 ticks. The higher standard deviation in Example 2 reflects the fact that the limit order book in this case features larger holes than in Example 3.

4 Market Quality, Tick Size and Arrival rate

In this section we explore the effect of a change in the tick size or in traders' arrival rate on measures of market performance (the average spread and the average waiting cost). For brevity we restrict our attention to the cases in which traders have different reservation spreads, i.e. $j_1^R < j_2^R$. Furthermore, we maintain our assumption that the parameters are such that $s_c = K$, so that impatient traders always choose market orders.

4.1 Measuring Market Performance

We would like to compare various equilibria in terms of market performance. In order to do so we introduce two measures, which take into account the benefits/costs accruing to different types of market participants. Our first measure is the expected dollar spread given by:

$$ES := \Delta \sum_{h=1}^{q} u_h n_h$$

This is one of the standard measures of market performance. The smaller is the expected dollar spread, the more distant are transaction prices from the "physical boundaries" A and B. Thus, smaller bid-ask spreads are associated with higher profits to liquidity demanders (the impatient traders), since their market orders meet more advantageous prices. Thus, we consider $-ES$ as a measure for the welfare of impatient traders who submit market orders.

Many studies exclusively focus on the bid-ask spread as a measure of market quality. The suppliers of liquidity, who are perhaps considered to be more professional traders or intermediaries, are frequently ignored. In our setting, however, we have no reason to ignore the liquidity providers, since they are traders just like the others.[26] Accordingly, our second measure of market quality is the cost of providing liquidity. The ex-ante expected cost of waiting for the traders posting limit orders (the patient traders) is:

$$EC := d_1 ET,$$

where $ET = \frac{\sum_{h=1}^{q-1} u_h T(n_h)}{1-u_q}$. We refer to ET as the ex-ante expected waiting time. It can be interpreted as the *average waiting time* of limit orders placed by patient traders.[27]

The welfare of the patient traders is determined by two factors: they would like to minimize their price concessions (or equivalently maximize the spread) in limit orders and they would like to minimize their expected waiting cost. Thus, $(ES - EC)$ measures the welfare of the patient traders. Observe that the expected spread is a transfer payment, while the cost of waiting is a dead-weight loss. This dead-weight loss is minimized when (a) only patient traders provide liquidity and (b) patient traders post their reservation spreads so that the expected waiting time is one period. When $j_1^R < j_2^R$, the division of roles is efficient: patient traders provide liquidity to impatient traders. However, the patient traders post spreads above their reservation spread for strategic reasons. This is a source of inefficiency since this behavior implies that the expected waiting time is strictly larger than one period.

[26] Glosten (2000) also argues that the welfare of all groups of traders must be taken into account in evaluations of specific market designs.

[27] The stationary probabilities are divided by the probability that the inside spread is less than K because no patient trader submits K-limit orders.

4.2 Tick Size and Market Quality

The tick size has been reduced in many limit order markets in the recent years.[28] It has often been argued that such a decrease would reduce the average dollar spread and would enhance market quality.[29] In this section we analyze the impact a reduction in the tick size on our measures of market performance. Our main result is that a decrease in the tick size does not necessarily improve the quality of a limit order market. In particular it can result in larger average dollar spreads.

We proceed as follows. Let us fix K, d_1, d_2 and θ and suppose that given a tick size equal to Δ we obtain an equilibrium with spreads: $1 \leq n_1 < n_2 < ... < n_q = K$. Let $\eta > 1$ be an integer, and let $\tilde{\Delta} = \Delta/\eta$ be the new tick size. We set $\tilde{K} = K\eta$ so that the dollar value of the largest spread does not change: $\tilde{K}\,\tilde{\Delta} = K\Delta$ (the change in tick size does not affect the monetary value of the range in which traders choose their prices). Now, for the tick size $\tilde{\Delta}$, we get a new equilibrium characterized by the spreads: $1 \leq \tilde{n}_1 < \tilde{n}_2 < ... < \tilde{n}_{\tilde{q}} = \tilde{K}$, where $\tilde{q} = \tilde{q}(\eta)$ is the length of the book in the new equilibrium.[30] We compare the two equilibria in the next lemma.

Lemma 3 *A decrease in tick size:*

- *increases or leaves unchanged the length of the book* ($\tilde{q} \geq q$),
- *decreases or leaves unchanged the monetary value of the smallest* q *spreads (i.e.* $\tilde{n}_h\tilde{\Delta} \leq n_h\Delta$ *for* $h = 1, .., q$*).*

On the one hand, a decrease in the tick size expands the set of prices which can be chosen by the traders in the range $[A, B]$. If limit order traders do not place aggressive orders, they will make use of the additional prices. This effect increases the length of the book. On the other hand, the decrease in the tick size shifts downward traders'

[28] For instance, the Toronto Stock Exchange in 1996 or the NYSE in 1997.

[29] See Harris (1997) for a review of the arguments in favor or against the reduction in the tick size.

[30] It can be checked that if $s_c = K$ when the tick size is Δ then this is still the case when the tick size is smaller. This means that impatient traders keep using only market orders when the tick size is $\tilde{\Delta}$.

reservation spread and results in larger spread improvements in terms of ticks (recall that Ψ_h is inversely related to the tick size). This effect reduces the monetary value of the smallest spreads in the book. These two effects have *opposite* impacts on the average spread. The first effect increases the average spread whereas the second effect decreases the average spread. As shown below, which effect is dominant mainly depends on the intensity of the competition among liquidity providers (r).

The ex-ante expected waiting time for limit orders depends on the length of the book. For this reason, a change in the tick size can also modify the ex-ante expected waiting costs for limit order traders. In the next propositions, we use the following notation: $\tau(q, r) := 1 + 2\sum_{h=1}^{q-1} r^h$. Observe that $\tau(q, r)$ increases with r.

Proposition 7 *A decrease in the tick size does not affect the length of the book if and only if $d_1\tau(q, r) \geq K\Delta$. In this case, the decrease in the tick size*

1. *Decreases the expected bid-ask spread.*

2. *Does not change the ex-ante expected waiting cost.*

The condition $d_1\tau(q, r) \geq K\Delta$ requires r to be sufficiently large since $\tau(q, r)$ increases with r. When r is large, liquidity providers submit aggressive limit orders. For this reason, they do not make use of the new prices created by the reduction in the tick size. It follows that the reduction in the tick size does not affect the length of the book in this case. When the change in the tick size leaves unchanged the length of the book, it has no effect on the probability distribution of spreads (see Lemma 2). However the dollar value of each inside spread posted in equilibrium is smaller (Lemma 3). For this reason the decrease in the tick size narrows the expected spread.

To illustrate the previous result, consider Example 2 where $r = 1.22$. The values of the different parameters in this example are such that the condition $d_1\tau(q, r) \geq K\Delta$ is satisfied. Hence the results of Proposition 7 applies. For instance, Table 4 presents the inside spreads posted in equilibrium before and after reducing the tick size from

$\$\frac{1}{8}$ to $\$\frac{1}{16}$ ($\eta = 2$ and $\tilde{K} = 40$). As expected, the length of the book does not change ($\tilde{q} = q = 7$). The dollar spreads decrease and the expected spread narrows from \$1.05 to \$1.01.

Table 4 - The impact of a decrease in tick size on equilibrium spreads in Example 2

	$\boldsymbol{\Delta} = \$\mathbf{0.125}$		$\tilde{\boldsymbol{\Delta}} = \$\mathbf{0.0625}$	
h	Spread (ticks)	Spread (\$)	Spread (ticks)	Spread (\$)
1	1	0.125	2	0.125
2	3	0.375	6	0.375
3	6	0.75	11	0.6875
4	9	1.125	17	1.0625
5	13	1.625	25	1.5625
6	18	2.25	34	2.125
7	20	2.5	40	2.5

Now we consider the case in which the length of the book enlarges when the tick size is reduced.

Proposition 8 *Suppose that $d_1\tau(q,r) < \Delta\left[K - \frac{q}{\eta}\right]$. Then a decrease in the tick size from Δ to $\tilde{\Delta} = \Delta/\eta$*

1. *Increases the length of the book ($\tilde{q}(\eta) > q$).*

2. *Increases the ex-ante expected waiting cost ($EC < \tilde{EC}$).*

The condition $d_1\tau(q,r) < \Delta\left[K - \frac{q}{\eta}\right]$ requires r to be sufficiently small since $\tau(q,r)$ increases in r. Intuitively when r is small, patient traders do not submit aggressive limit orders. This means that the new prices created by the reduction in the tick size are used by patient traders. This effect enlarges the length of the book and creates $\tilde{q} - q$ new large spreads. Hence there are more limit orders in the book and the average waiting time (or cost) increases. The new spreads are in the range $\left[K, \tilde{K}\right]$. Thus they contribute to widen the expected spread. At the same time the decrease in the tick size drives the first q inside spreads smaller in monetary terms. Therefore the impact of a decrease in the tick size on the expected spread is ambiguous. Cases in which the expected spread enlarges when the tick size is reduced do exist. For instance consider a reduction in the tick size from $\$\frac{1}{8}$ to $\$\frac{1}{16}$ (i.e. $\eta = 2$) in Example 3 where $r = 0.818$. The values of the parameters are such that the condition $d_1\tau(q, r) < \Delta\left[K - \frac{q}{\eta}\right]$ is satisfied. The decrease in the tick size causes the length of the book to grow dramatically from $q = 18$ to $\tilde{q} = 32$. The expected spread and the expected waiting time rise from \$2.0 to \$2.2 and from 8.86 to 9.87 periods, respectively.

For the values of r such that $d_1\tau(q, r) < \Delta K$, there always exists a value of η large enough such that the condition in Proposition 8 holds true. Consequently for these values of r, the ex-ante expected waiting cost starts increasing when the tick size becomes too small. In these cases, the tick size which maximizes welfare (i.e. minimizes the ex-ante waiting cost) is always strictly positive. A policy implication is that exchanges and regulators should consider the impact of the tick size on average waiting costs for liquidity suppliers and not only on spreads.

4.3 Arrival Rate and Market Quality

In presence of waiting costs, the order arrival rate is a determinant of traders' bidding strategy. Demsetz (1968), p.41 points out that: "*The fundamental force working to reduce the spread is the time rate of transactions. The greater the frequency of transacting, the lower will be the cost of waiting in a trading queue of a specified length, and, therefore the lower will be the spreads that traders are willing to submit to preempt positions in the trading queue.*" In this section, we argue that the impact of a decrease in the order arrival rate on market quality is ambiguous. Actually a decrease in the order arrival rate induces liquidity suppliers to submit more aggressively priced limit orders when the inside spread is large. For this reason, counter-intuitively, a decrease in the order arrival rate does not necessarily increase the expected spread. We provide an example supporting this claim.

Let t be the average length of a period in *calendar time* and let δ_i be trader i's waiting cost per unit of calendar time. In this case the per period waiting cost is

$$d_i = \delta_i t.$$

The larger is the order arrival rate (the smaller t), the smaller is the per period waiting cost, for a given level of impatience (δ_i). Hence variations in the order arrival rate are tantamount to variations in the per period waiting cost. We consider the effect of an increase in traders' waiting cost (a decrease in traders' arrival rate) using the characterization of the equilibrium provided in Section 3.[31] Let q be the original length of the book and $\tilde{q}$ be the length of the book after the decrease in traders' arrival rate.

Proposition 9 *A decrease in traders' arrival rate (an increase in d_1) :*

1. *decreases or leaves unchanged the length of the book ($\tilde{q} \cdot q$),*
2. *decreases or leaves unchanged the ex-ante expected waiting time.*
3. *increases or leaves unchanged the $\tilde{q}$ smallest inside spreads posted in the book.*

For given bidding strategies, an increase in d_1 enlarges liquidity providers' ex-ante expected waiting cost. In order to counteract this effect, liquidity providers react by submitting more aggressive orders (spread improvements get larger). For this reason the length of the book tends to decrease when traders' arrival rate decreases. As a consequence the ex-ante expected waiting time (which is expressed *in number of periods*) becomes smaller. Hence the net impact of a decrease in traders' arrival rate on the ex-ante expected waiting cost (EC) is ambiguous.

As patient traders' waiting cost increases, they require a larger compensation to submit limit orders. For instance, their reservation spread increases. This explains the last part of the proposition. However limit order traders are more aggressive

[31] If the condition $s_c = K$ holds true for a given level of per period waiting cost for an impatient trader, it also holds true for any larger level. Hence the characterization of the equilibrium given in Section 3 when traders are heterogeneous remains valid when the order arrival rate decreases.

when spreads are large so that the inside spread adjusts more quickly to small levels. It follows that an increase in liquidity providers' waiting costs can indeed result in a smaller average spread.

We illustrate the previous discussion with the following example. Suppose that in Example 3, the per period waiting cost increases from $d_1 = 0.1$ to $d_1 = 0.249$. In this case, calculations show that the length of the book becomes $\tilde{q} = 9$ instead of $q = 18$. The expected spread narrows and is equal to \$15.85 instead of \$16.03. The average waiting time decreases as well (6.10 periods instead of 8.86 periods). However, the ex-ante expected waiting cost enlarges and is equal to \$1.52 instead of \$0.886.

5 Extensions

Qeuing at the inside spread

We have assumed that traders cannot queue, i.e. can not place limit orders at the existing inside quote. In reality, such quotes are allowed. Then time priority determines the sequence in which limit orders placed at the same price are executed. In the next proposition, we identify a condition on the parameters such that the equilibrium we have described in Section 3 is unchanged when traders are allowed to queue at the best quotes. We just focus on the case in which traders are heterogeneous for brevity.

Proposition 10 *When time priority is enforced and traders are heterogeneous, the equilibria when traders are allowed to queue at the inside spread and when they are not are identical if* $\frac{d_1}{\Delta} \geq \frac{1}{2}$.

The intuition is as follows. Suppose that traders use the trading strategies described in Section 3 and give them the freedom to queue at the best quotes. Under the condition of the proposition, traders prefer to submit limit orders improving upon the inside spread rather than queuing. Hence traders' strategies form an equilibrium even though traders have the possibility to queue. Not surprisingly queuing is not optimal if (i) liquidity providers' waiting cost is large, or (ii) the tick size is small. Queuing increases the expected waiting time substantially, so that undercutting is always optimal when patient traders' waiting cost is sufficiently large. When the tick size is small, liquidty providers can seize time priority at a low cost since they need to undercut by a small amount only.[32]

Multiple trader types.

Introduction of multiple trader types (ordered by their patience level - reservation spread) does not change the model qualitatively. In particular the model still exhibits sensitivity to the proportion of relatively patient traders in the population (see Kadan 2001). The presentation of the model is more complex, however. In particular, more than three distinct types of equilibria do appear.

Commonality in liquidity.

Recent literature (see Chordia, Roll, and Subrahmanyam (2001), and Huberman and Halka (2001)) identifies common elements in liquidity across stocks. The conventional models have difficulty explaining this phenomenon, since private information arrivals are unlikely to be correlated across stocks, and one does not expect strong correlation in dealer inventory levels either. Market liquidity in our model is determined by the proportion of patient traders, trader arrival rates, and the tick size. The first two parameters vary continuously over time, and may very well have market-wide components.[33] In such a case our model predicts commonality in liquidity across stocks that is consistent with the empirical findings. The model generates many predictions that allow to test this conjecture.

Professional liquidity provider.

The analysis of Section 3 reveals that patient traders' strategic behavior can result in transactions taking place when spreads are wide (relative to liquidity providers'

[32] It is worth stressing that the condition given in Proposition 10 is satisfied in all the numerical examples we gave in the paper. Furthermore this condition is sufficient for queuing at the inside quote to be sub-optimal but not necessary.

[33] These would be consistent with the popular notions such as "active market", "jittery investors" and others.

reservation spreads), especially in markets where the proportion of patient traders is relatively small ($r < 1$). Hence the book offers profit opportunities that invite submission of limit orders by a professional trader (the "intermediary") monitoring the market. We briefly discuss the impact of such an intermediary on traders' behavior in our model.

We assume that the intermediary is risk neutral and has no cost of waiting.[34] When he intervenes, he submits two limit orders that improve on the current spread by one tick from each side. Thus, if the inside spread has size s, its new size becomes $s - 2$ after the intermediary's intervention. In this way the intermediary earns $(s - 2)\Delta$ when subsequent market orders clear his limit orders. We also assume that there is a spread s_0 below which the intermediary does not intervene. This reflects the fact that he incurs (per share) trading costs or monitoring costs. Since these costs are non-negative, we assume $s_0 \geq 3$. We refer to the range $\langle s_0, K\rangle$ as being the intermediary's intervention zone. When the inside spread is in this intervention zone, the intermediary will intervene with probability β (the intervention rate). To simplify we take the intervention rate as exogenous. An intervention rate less than 100% can be due to the fact that the intermediary cannot monitor continuously the market (e.g. he is active in several markets).

Let $T^{\beta}(j)$ be the expected waiting for a j-limit order when the intermediary's intervention rate is β. The formula given in Lemma 1 generalizes as follows

$$T^{\beta}(j) = \begin{cases} \dfrac{\left[1+\sum_{k=1}^{j-1}\alpha_k(j)T(k)\right]}{\alpha_0(j)} & for \quad j = 1, ..., s_0 - 1 \\[2ex] \dfrac{(1+\beta\left[T^{\beta}(j-1)+T^{\beta}(j-2)\right]+(1-\beta)\sum_{k=1}^{j-1}\alpha_k(j)T^{\beta}(k))}{\alpha_0(j)(1-\beta)} & for \quad j = s_0, ..., K \end{cases} \tag{7}$$

Using Eq. (7) and proceeding recursively, as in Section 3, we can calculate traders' optimal placement strategies for each intervention rate. Interestingly even a small intervention rate creates a large change in the behavior of liquidity suppliers, especially when the proportion of patient traders is relatively small ($r < 1$). We demonstrate

[34] This assumption is not crucial but simplifies the analysis.

this using Example 3. Recall that in this example we set $d_1 = 0.25$, $d_2 = 0.1$ and $\theta = 0.45$ ($r = 0.82$). Table 5 presents the optimal strategies of a patient trader for various intervention rates[35]. The intervention zone is set to $\langle 3, 20\rangle$, i.e. the costs of the intermediary enable him to intervene whenever the spread is at least $s_0 = 3$ ticks.

Table 5

Patient traders' optimal strategies for various intervention rates

Current Spread	$\beta = 0.00$	$\beta = 0.10$	$\beta = 0.15$	$\beta = 0.25$
1	0	0	0	0
2	1	1	1	1
3	1	1	1	1
4	3	1	1	1
5	3	4	1	1
6	5	4	5	1
7	6	4	5	1
8	7	4	5	1
9	8	8	5	1
10	9	8	5	1
11	10	8	5	1
12	11	8	5	1
13	12	8	5	1
14	13	8	5	1
15	14	14	5	1
16	15	14	5	1
17	16	14	16	1
18	17	14	16	1
19	18	14	16	1
20	19	14	16	1

Clearly patient traders are more aggressive in presence of the intermediary ($\beta > 0$) than when he does not intervene ($\beta = 0$). Intuitively the threat of intervention by the intermediary increases liquidity providers' expected waiting time, other things equal. In turn they submit more aggressive limit orders in order to shorten execution times. When $\beta = 0.25$, the competitive pressure created by the intermediary is so strong

[35] Recall that the optimal strategy of impatient traders when $\beta = 0$ is to submit market orders only. It remains so when $\beta > 0$.

that the patient traders submit limit orders that prevent the intermediary profitably submitting orders (the inside spread is less than s_0).

This simple illustration demonstrates that adding a designated liquidity provider to the pure order-driven market forces liquidity suppliers to submit more aggressive limit orders. This translates into lower trading costs for liquidity demanders. For some intervention rate (e.g. $\beta = 0.25$), the presence of the designated liquidity provider also reduces the ex-ante expected waiting cost for liquidity suppliers. Some exchanges using limit order markets (e.g. the Paris Bourse or the Frankfurt Stock Exchange) encourage the intervention of designated liquidity providers in less liquid stocks. The previous analysis provides a rationale for this policy.

6 Conclusions

We model the limit order book as a market for liquidity provision and consumption. In contrast with the extant literature we consider the optimal order placement decision of liquidity traders who incur waiting costs. Furthermore we endogenize the expected waiting time of limit order traders.

The proportion of patient traders relative to the proportion of impatient traders turns out to be a main determinant of the dynamics of the book. Actually it determines the intensity of competition between liquidity providers and the speed at which liquidity is consumed. In markets with a relatively large proportion of patient traders, traders submit aggressively priced limit orders in order to reduce their execution time. Furthermore market orders are less frequent. The combination of these two effects imply that the probability distribution of spreads is skewed towards small spreads in these markets.

We also find that a decrease in the tick size may enlarge the average inside spread in markets where the proportion of patient traders is relatively small. Actually in this case, patient traders do not place very aggressive orders. A finer price grid gives them the possibility to place even less aggressive orders by expanding the set of eligible prices. A decrease in traders' arrival rate induces liquidity providers to place limit orders which are more aggressively priced in order to get faster execution. For this reason, counter-intuively, lower trading activity does not necessarily result in smaller average spreads.

Finally, we show that a designated market-maker increases competitive pressures among liquidity providers. In this way his presence can drastically improve the provision of liquidity (lower spreads and smaller average waiting costs) when the proportion of liquidity providers is small.

References

[1] Angel, J., (1994), Limit versus Market Order, working paper, Georgetown University.

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

[3] Benston, G, P. Irvine, and E. Kandel (2001), Liquidity Beyond the Inside Spread: Measuring and Using Information in the Limit Order Book, working paper.

[4] Biais, B., Hillion, P. and C. Spatt (1995), An Empirical Analysis of the Limit Order Book and the Order Flow in the Paris Bourse, Journal of Finance 50, 1655-1689.

[5] Biais, B, Martimort, D. and J.C. Rochet (2000), Competing Mechanisms in a Common Value Environment, Econometrica, 68, 799-837.

[6] Chakravarty, S. and C. Holden (1995), An Integrated Model of Market and Limit Orders, Journal of Financial Intermediation, 4, 213-241.

[7] Chordia, T., Roll, R. and A. Subrahmanyam (2001), Commonality in Liquidity, working paper, Anderson Graduate School of Management.

[8] Demsetz H., (1968), The Costs of Transacting, Quarterly Journal of Economics, 82, 33-53.

[9] Domowitz I., (1993) A Taxonomy of Automated Trade Execution Systems, Journal of International Money and Finance, 12, 607-631.

[10] Domowitz I., and A. Wang, (1994) Auctions as Algorithms, Journal of Economic Dynamics and Control, 18, 29-60.

[11] Feller W., (1968) An Introduction to Probability Theory and its Applications, 3^{rd} Edition, John Wiley & Sons.

[12] Foucault T., (1999), Order Flow Composition and Trading Costs in a Dynamic Limit Order Market, Journal of Financial Markets, 2, 99-134.

[13] Glosten, L., (1994), Is the Electronic Order Book Inevitable, Journal of Finance 49, 1127–1161.

[14] Glosten L., (2000), Competition, Design of Exchanges, and Welfare, Working Paper, Columbia University.

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

[16] Handa, P. and R. Schwartz (1996), Limit Order Trading, Journal of Finance 51, 1835–1861.

[17] Harris, L. (1997), Decimalization: A Review of the Arguments and Evidence, working paper, USC.

[18] Harris, L. (1998), Optimal Dynamic Order Submission Strategies in some Stylized Trading Problems, Financial Markets, Institutions and Instruments, Vol 7, 2.

[19] Harris, L. and J. Hasbrouck (1996), Market versus Limit orders : the Superdot evidence on Order Submission Strategy, Journal of Financial and Quantitative Analysis 31, 213-231.

[20] Hollifield, B., Miller, A. and P. Sandas (2001a), Empirical Analysis of Limit Order Markets, working paper, Carnegie-Mellon University.

[21] Hollifield, B., Miller, A. and P. Sandas (2001b), Liquidity Supply and Demand: Empirical Evidence from the Vancouver Stock Exchange, mimeo, Carnegie-Mellon University.

[22] Huberman, G. and D. Halka (2001), Systematic Liquidity, working paper, Columbia Business School.

[23] Kadan, O., (2001), Strategic Models of Bilateral Trade With Applications to Order Driven Markets, Ph.D. Thesis, Hebrew University.

[24] Kandel E., Lauterbach B., and I. Tkach (2000), Quality of Execution in Tel Aviv Stock Exchange, Working Paper, Hebrew University.

[25] Kavajecz, K. (1999), A specialist's Quoted Depth and the Limit Order Book, Journal of Finance, 54, 747-771.

[26] Lo, A., C. McKinlay, and Zhang, (2001), Econometric Models of Limit Order Executions, Journal of Financial Economics, forthcoming.

[27] Madhavan, A., (2000), Market Microstructure and Price Formation, Journal of Financial Markets, 205–258.

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

[29] Parlour, C., and D. Seppi, (2001), Liquidity-Based Competition for Order Flow, Working Paper, Carnegie Mellon University.

[30] Rock, K., (1996), The Specialist's Order Book and Price Anomalies, working paper.

[31] Sandas, P., 2000, Adverse Selection and Competitive Market Making: Evidence from A Limit Order Market, forthcoming, Review of Financial Studies.

[32] U.S. Securities and Exchange Commission (SEC), 1997 *Report on the Practice of Preferencing.*

[33] Seppi, D., (1997), Liquidity Provision with Limit Orders and a Strategic Specialist, Review of Financial Studies 10, 103-150.

[34] Shleifer, A. and R. Vishny (1997), Limits to Arbitrage, Journal of Finance, 52(1), 35-55.

7 Appendix

Proof of Lemma 1

Suppose a trader (say a buyer) has submitted a j-limit order. Suppose $j > 1$. The next trader (a seller) must choose among j options. With probability $\alpha_0(j)$, he submits a market order that clears the buyer's limit order. With probability $\alpha_k(j)$, the seller submits a k-limit order. In this case the seller has to wait $T(k)$ periods until his order is cleared. From that moment the original buyer has to wait another $T(j)$ periods. This follows from Assumption A.3. To see this point, suppose that $j = 2$ for instance. The seller submits a 1-limit order. Then a buyer arrives who clears the seller's order and the original buyer is back to the initial situation (the inside spread is $j = 2$ and the buyer has priority). Consequently, the original buyer's expected waiting time, $T(j)$, is:

$$T(j) = \alpha_0(j) + \sum_{k=1}^{j-1} \alpha_k(j)\left[1 + T(k) + T(j)\right]. \tag{8}$$

Solving for $T(j)$ and using the fact that $\sum_{k=0}^{j-1} \alpha_k(j) = 1$ yield Eq.(2). The same argument applies to a seller. ■

Proof of Proposition 2

The proof of this proposition relies on two lemmas that we prove in turn.

Lemma 4 *Suppose that facing a spread of size s, trader i submits a j-limit order with $0 \leq j < s$. In this case facing a spread of size $s+1$, he either submits a s-limit order or a j-limit order.*

Proof. By assumption trader i submits a j-limit order when he faces a spread with size s. Thus:

$$\begin{aligned} \pi_i(j) &> \pi_i(k) \quad k = 0, ...j - 1, \\ \pi_i(j) &\geq \pi_i(k) \quad k = j+1, .., s - 1. \end{aligned}$$

Now, suppose that trader i faces a spread of size $s+1$. If $\pi_i(s) \leq \pi_i(j)$ then trader i will submit a j-limit order since $\pi_i(j) > \pi_i(k)$ for all $k = 0, ..., j-1, j+1, ..., s$. If $\pi_i(s) > \pi_i(j)$ then trader i submits a s-limit order since $\pi_i(s) > \pi_i(k)$ for all $k = 0, ..., s-1$. ■

Lemma 5 *Suppose that facing a spread of size s, an impatient trader submits a j-limit order with $j \geq 1$. In this case facing a spread of size s, a patient trader submits a limit order as well.*

Proof. Suppose on the contrary that a patient trader submits a market order when the inside spread has a size equal to s. It follows that:

$$0 \leq \pi_1(0) - \pi_1(j) = -j\Delta + T(j)d_1 \leq -j\Delta + T(j)d_2 = \pi_2(0) - \pi_2(j), \ \forall j \geq 1.$$

But this means that an impatient trader prefers a market order to a $j-$limit order - a contradiction. ■

When the inside spread is equal to one tick, all the traders submit a market order, whatever their type. Now suppose that a patient trader faces a spread of two ticks and $j_1^R \geq 2$ (i.e. $d_1 > \Delta$). If he submits a 1-limit order he obtains:

$$\pi_1(1) = \Delta - d_1 < 0.$$

Therefore he prefers a market order. From Lemma 5 it follows that an impatient trader also prefers a market order when he faces a spread of two ticks. This implies that $T(1) = T(2) = 1$. By induction it follows that facing any spread in $\langle 1, j_1^R \rangle$, all the traders submit market orders, whatever their type and that $T(1) = T(2) = ... = T(j_1^R) = 1$.

Now suppose a patient trader faces a spread with size $j_1^R + 1$. Lemma 4 implies that he may either submit a j_1^R- limit order or a market order. He obtains a larger payoff with a j_1^R-limit order since

$$\pi_1(j_1^R) = j_1^R\Delta - T(j_1^R)d_1 = j_1^R\Delta - d_1 > 0.$$

From Lemma 4 it follows now that the patient type submits limit orders for all spreads $s \in \langle j_1^R + 1, K\rangle$. As for the impatient type there are two cases:

Case 1: The impatient type submits a market order for each $s \in \langle j_1^R + 1, K\rangle$ in which case we set $s_c = K$.

Case 2: There are spreads in $\langle 1, K\rangle$ for which the impatient type submits limit orders. In this case let s_c be the smallest spread that an impatient trader creates with a limit order. By definition of s_c, the impatient trader submits a market order when he faces a spread $s \in \langle 1, s_c\rangle$ and a s_c-limit order when he faces a spread with size $s_c + 1$. Lemma 5 implies that the patient type also submits a limit order when he faces a spread with size $s_c + 1$. Then, from repeated application of Lemma 4, it follows that both patient and impatient traders submit a limit order when they face a spread in $\langle s_c + 1, K\rangle$. Finally it cannot be optimal for an impatient trader to submit a limit order which creates a spread smaller than his reservation spread. This implies $s_c \geq j_2^R$. ■

Proof of Proposition 3

Since we assume that $s_c = K$ the impatient type always submits market orders. From Proposition 2, a patient trader submits a market order when he faces a spread in $\langle 1, j_1^R\rangle$ and a j_1^R-limit order when he faces a spread with size $j_1^R + 1$. Repeated application of Lemma 4 shows the existence of spreads $n_1 < n_2 < ... < n_q$ such that facing a spread in $\langle n_h + 1, n_{h+1}\rangle$ the patient trader submits a n_h-limit order for $h = 1, ..., q - 1$. Clearly, $n_1 = j_1^R$ and $n_q = K$. ■

Proof of Proposition 4

When they observe a spread with size j_1^R, all the traders submit a market order. Therefore $T(n_1) = T(j_1^R) = 1$. Let $h \in \{2, 3, ..., q\}$. Suppose that the posted spread is $s \in \langle n_{h-1} + 1, n_h\rangle$. When he observes this spread, a patient trader submits a n_{h-1}-limit order and an impatient trader submits a market order. Therefore $\alpha_0(s) = 1 - \theta$

and $\alpha_{n_{h-1}}(s) = \theta$. It follows from Lemma 1 that

$$T(s) = \frac{1}{1 - \theta}\left[1 + \theta T(n_{h-1})\right], \forall s \in \langle n_{h-1} + 1, n_h \rangle. \tag{9}$$

Hence $T(\cdot)$ is constant for all $s \in \langle n_{h-1} + 1, n_h \rangle$. Using Eq.(9) and the fact that $T(n_1) = 1$, we obtain

$$T(n_{h+1}) - T(n_h) = r(T(n_h) - T(n_{h-1})) \quad for \quad h \geq 2, \tag{10}$$

and

$$T(n_2) - T(n_1) = 2r > 0.$$

The claim follows now by repetitive application of Eq.(10) and the fact that $T(n_1) = 1$. ■

Proof of Corollary 1 Immediate using the expression for the waiting time function. ■

Proof of Proposition 5

Since $n_h = n_{h-1} + \Psi_h$, we immediately get that $n_h = n_1 + \sum_{k=2}^{k=h} \Psi_k$. Furthermore since $n_q = K$, it must be the case that q is the smallest integer such that $n_1 + \sum_{k=2}^{k=q} \Psi_k \geq K$. Proposition 3 implies that facing a spread of n_{h+1} the patient type prefers a n_h-limit order over a n_{h-1}-limit order for $h = 2, ..., q$. Hence:

$$n_h\Delta - T(n_h)d_1 > n_{h-1}\Delta - T(n_{h-1})d_1.$$

Rearranging and using Proposition 4 yields for $h = 2, .., q$:

$$\Psi_h := n_h - n_{h-1} > (T(n_h) - T(n_{h-1}))\frac{d_1}{\Delta} = 2r^{h-1}\frac{d_1}{\Delta}. \tag{11}$$

Again from Proposition 3, facing a spread of n_h a patient trader (a) strictly prefers a n_{h-1}-limit order over a limit order which creates a spread with size $n_h - 1$ or (b) $n_{h-1} = n_h - 1$. Therefore

$$n_{h-1}\Delta - T(n_{h-1})d_1 \geq (n_h - 1)\Delta - T(n_h - 1)d_1.$$

Rearranging and using the fact that $T(n_h - 1) \leq T(n_h)$ we obtain for $h = 2, ...q$:

$$\Psi_h = n_h - n_{h-1} \leq (T(n_h) - T(n_{h-1}))\frac{d_1}{\Delta} + 1 = 2r^{h-1}\frac{d_1}{\Delta} + 1. \tag{12}$$

Combining (11) and (12) we have for $h = 2, ..., q$:

$$\Psi_h = int(2r^{h-1}\frac{d_1}{\Delta}) + 1. \tag{13}$$

■

Proof of Lemma 2

We first show that the Markov chain given by W is (a) irreducible and (b) a-periodic.

The Markov chain is irreducible. Observe that given any two states j_1, j_2 with $1 \leq j_1$ and $j_2 \leq q$ there is a positive probability that the chain will move from j_1 to j_2 after a sufficiently large (though finite) number of transitions. This implies that any two states in the chain communicate, hence the chain is irreducible.

The Markov chain is a-periodic. Notice that $W_{q,q} = 1 - \theta > 0$. This means that when the chain is in state q, there is a probability equal to $(1 - \theta)^m$ that it will stay in this state for the next m transitions, $\forall m \geq 1$. Since state q communicates with all the other states of the chain, it follows that no state has a period greater than 1. Thus the chain is a-periodic.

These properties imply that the Makov chain is ergodic. Being ergodic, the induced Markov chain yields a unique stationary distribution of spreads (see Feller 1968). Let $u = (u_1, ..., u_q)$ denote the row vector of stationary probabilities. The stationary probability distribution is obtained by solving $q + 1$ linear equations given by:

$$uW = u \text{ and } u\varepsilon = 1, \tag{14}$$

where ε stands for the unit column vector. It is straightforward to verify that the probabilities given by Eq.(5) and Eq.(6) are a solution of this system of equations. ■

Proof of Lemma 3

We first show by induction on h that $\tilde{n}_h\tilde{\Delta} \le n_h\Delta$ for all $h = 1, .., q$. We start with $h = 1$. Applying Eq.(3) to both equilibria we obtain:

$$\begin{aligned} \frac{d_1}{\Delta} &< n_1 \le \frac{d_1}{\Delta} + 1, \\ \frac{d_1}{\tilde{\Delta}} &< \tilde{n}_1 \le \frac{d_1}{\tilde{\Delta}} + 1, \end{aligned} \tag{15}$$

which imply

$$\tilde{n}_1 - n_1 < \frac{d_1}{\tilde{\Delta}} + 1 - \frac{d_1}{\Delta} = 1 + \frac{d_1}{\Delta}(\eta - 1) < 1 + n_1(\eta - 1).$$

Rearranging yields $\tilde{n}_1 < \eta n_1 + 1$ and since η, n_1 and $\tilde{n}_1$ are integers we conclude that $\tilde{n}_1 \le \eta n_1$. Multiplying both sides of this inequality by $\tilde{\Delta}$ yields $\tilde{n}_1\tilde{\Delta} \le n_1\Delta$ as expected. Now, let h be an integer satisfying: $1 < h \le q - 1$. From Proposition 5 we conclude:

$$\begin{aligned} 2r^{h-1}\frac{d_1}{\Delta} + n_{h-1} &\le n_h < 2r^{h-1}\frac{d_1}{\Delta} + n_{h-1} + 1, \\ 2r^{h-1}\frac{d_1}{\tilde{\Delta}} + \tilde{n}_{h-1} &\le \tilde{n}_h < 2r^{h-1}\frac{d_1}{\tilde{\Delta}} + \tilde{n}_{h-1} + 1. \end{aligned} \tag{16}$$

Rearranging:

$$\begin{aligned} \tilde{n}_h - n_h &< 2r^{h-1}d_1\left(\frac{1}{\tilde{\Delta}} - \frac{1}{\Delta}\right) + \tilde{n}_{h-1} - n_{h-1} + 1, \\ &= 2r^{h-1}\frac{d_1}{\Delta}(\eta - 1) + \tilde{n}_{h-1} - n_{h-1} + 1. \end{aligned} \tag{17}$$

The induction hypothesis yields $\tilde{n}_{h-1} \le \eta n_{h-1}$, or $\tilde{n}_{h-1} - n_{h-1} \le n_{h-1}(\eta - 1)$. From Eq. (16) we have $2r^{h-1}\frac{d_1}{\Delta} \le n_h - n_{h-1}$. Substituting these two inequalities into Eq. (17) yields:

$$\tilde{n}_h - n_h < (n_h - n_{h-1})(\eta - 1) + n_{h-1}(\eta - 1) + 1,$$

or $\tilde{n}_h < n_h\eta + 1$. Since $\tilde{n}_h, n$ and η are integers we obtain $\tilde{n}_h \leq n_h\eta$. Multiplying both sides by $\tilde{\Delta}$ yields $\tilde{n}_h\tilde{\Delta} \leq n_h\Delta$ as expected. Finally consider $h = q$. There are two possibilities. If

$$\tilde{n}_1 + \sum_{k=2}^{k=q} \tilde{\Psi}_k \geq \tilde{K},$$

then $\tilde{q} = q$. In this case $\tilde{n}_q = \tilde{K}$ which implies that $\tilde{n}_q\tilde{\Delta} = n_q\Delta$. If

$$\tilde{n}_1 + \sum_{k=2}^{k=q} \tilde{\Psi}_k < \tilde{K},$$

then $\tilde{q} > q$. In this case $\tilde{n}_q < \tilde{K}$ which implies that $\tilde{n}_q\tilde{\Delta} < n_q\Delta$. Hence we have proved that $\tilde{n}_h\tilde{\Delta} \leq n_h\Delta$, $\forall h \leq q$ and $\tilde{q} \geq q$. ■

Proof of Proposition 7

Suppose that $d_1\tau(q,r) \geq K\Delta$. First we prove that $\tilde{q}(\eta) = q$, $\forall \eta \geq 1$. Suppose on the contrary that $\tilde{q}(\eta) > q$. In this case $\tilde{n}_q < \tilde{K}$. Furthermore, in equilibrium, when he faces a spread of $\tilde{n}_{q+1}$,the investor is better off submitting a $\tilde{n}_q -$limit order rather than a market order. The two remarks imply

$$0 < \tilde{n}_q\tilde{\Delta} - d_1 T(\tilde{n}_q) < \tilde{K}\tilde{\Delta} - d_1 T(\tilde{n}_q) = K\Delta - d_1 T(\tilde{n}_q)$$

Using Proposition 4 we observe that $T(\tilde{n}_q) = \tau(q,r)$. Hence the previous inequality implies that

$$0 < K\Delta - d_1\tau(q,r),$$

in contradiction to our assumption. This shows that $d_1\tau(q,r) \geq K\Delta$ is a sufficient condition for $\tilde{q}(\eta) = q$, $\forall \eta \geq 1$. In order to prove that this condition is also necessary we need the following lemma.

Lemma 6 *Suppose that $d_1\tau(q,r) < \Delta\left[K - \frac{q}{\eta}\right]$. Then a decrease in the tick size from Δ to $\tilde{\Delta} = \Delta/\eta$ increases the length of the book (i.e. $\tilde{q}(\eta) > q$).*

Proof. Suppose on the contrary that $\tilde{q}(\eta) = q$. Proposition 5 implies that

$$\tilde{\Psi}_h \leq int(2r^{h-1}\frac{d_1}{\tilde{\Delta}}) + 1 \leq \frac{2r^{h-1}d_1}{\tilde{\Delta}} + 1, \;\; \forall 2 \leq h \leq q.$$

Furthermore

$$\tilde{n}_1 \leq \frac{d_1}{\tilde{\Delta}} + 1$$

It follows that:

$$\tilde{K} = \tilde{n}_1 + \sum_{h=2}^{q} \tilde{\Psi}_h \leq \frac{d_1}{\tilde{\Delta}}(1 + 2\sum_{h=1}^{q-1} r^{h-1}) + q.$$

Multiplying by $\tilde{\Delta}$ both sides of the inequality and using the fact that $K\Delta = \tilde{K}\tilde{\Delta}$ yield

$$\Delta \left[K - \frac{q}{\eta} \right] \leq d_1\tau(q, r)$$

- a contradiction. ■

Observe that if $d_1\tau(q, r) < K\Delta$,there exists η such that $d_1\tau(q, r) < \Delta \left[K - \frac{q}{\eta} \right]$. It then follows from the previous lemma that the condition $d_1\tau(q, r) \geq K\Delta$ is necessary for the length of the book to be unchanged when the tick size is reduced.

Now we observe that the stationary probability distribution of spreads is not affected by a reduction in the tick size if this reduction leaves unchanged the length of the book. This implies that the ex-ante expected waiting time does not change when the reduction in tick size has no impact on the length of the book. Furthermore Lemma 3 implies that each one of the q spreads (weakly) decreases when the tick size decreases. This implies that the expected spread weakly decreases with the tick size under the condition of the proposition. ■

Proof of Proposition 8

The first part of the proposition follows from Lemma 6 that we have established in the proof of the previous proposition. For the second part of the proposition, we use Lemma 7 below. Let $u_1, ..., u_q$ and $\tilde{u}_1, ..., \tilde{u}_{\tilde{q}}$ denote the stationary probabilities of the inside spreads in equilibrium when the tick sizes are Δ and $\tilde{\Delta}$, respectively. Similarly we denote by EC and $E\tilde{C}$ the ex-ante expected cost of waiting in the two equilibria.

Lemma 7 *If $q < \tilde{q}$ then $u_h > \tilde{u}_h$ for $h = 1, ..., q$.*

Proof. Let h be a spread with $h = 2, ..., q$. From Eq.(6) we have:

$$\begin{aligned}
\frac{u_h}{\tilde{u}_h} &= \frac{\theta^{q-h}(1-\theta)^{h-2}}{\theta^{q-1} + \sum_{i=2}^{q} \theta^{q-i}(1-\theta)^{i-2}} \cdot \frac{\theta^{\tilde{q}-1} + \sum_{i=2}^{\tilde{q}} \theta^{\tilde{q}-i}(1-\theta)^{i-2}}{\theta^{\tilde{q}-h}(1-\theta)^{h-2}} = \\
&= \frac{1}{\theta^{\tilde{q}-q}} \cdot \frac{\theta^{\tilde{q}-1} + \sum_{i=2}^{\tilde{q}} \theta^{\tilde{q}-i}(1-\theta)^{i-2}}{\theta^{q-1} + \sum_{i=2}^{q} \theta^{q-i}(1-\theta)^{i-2}} = \frac{\theta^{\tilde{q}-1} + \sum_{i=2}^{\tilde{q}} \theta^{\tilde{q}-i}(1-\theta)^{i-2}}{\theta^{\tilde{q}-1} + \sum_{i=2}^{q} \theta^{\tilde{q}-i}(1-\theta)^{i-2}} > 1
\end{aligned}$$

The proof for the case $h = 1$ is similar. ■

Define $w_h := \frac{u_h}{1-u_q}$ for $h = 1, ..., q-1$ and $\tilde{w}_h := \frac{\tilde{u}_h}{1-\tilde{u}_{\tilde{q}}}$ for $h = 1, ..., \tilde{q}-1$. Then $EC = d_1 \sum_{h=1}^{q-1} w_h T(n_h)$ and $E\tilde{C} = d_1 \sum_{h=1}^{\tilde{q}-1} \tilde{w}_h T(\tilde{n}_h)$. It follows from Lemma 7 that $\tilde{w}_h < w_h$ for $h = 1, ..., q-1$. Since $T(n_h) = T(\tilde{n}_h)$ for $h = 1, ..., q-1$ and since $T(n_h)$ is increasing in h we have:

$$\begin{aligned}
E\tilde{C} - EC &= d_1 \sum_{h=1}^{q-1} \tilde{w}_h T(\tilde{n}_h) + d_1 \sum_{h=q}^{\tilde{q}-1} \tilde{w}_h T(\tilde{n}_h) - d_1 \sum_{h=1}^{q-1} w_h T(n_h) \\
&= d_1 \left[\sum_{h=q}^{\tilde{q}-1} \tilde{w}_h T(\tilde{n}_h) - \sum_{h=1}^{q-1} (w_h - \tilde{w}_h) T(\tilde{n}_h) \right] \\
&> d_1 \left[\sum_{h=q}^{\tilde{q}-1} \tilde{w}_h T(\tilde{n}_q) - \sum_{h=1}^{q-1} (w_h - \tilde{w}_h) T(\tilde{n}_q) \right] \\
&= d_1 T(\tilde{n}_q) \left[\sum_{h=q}^{\tilde{q}-1} \tilde{w}_h - \sum_{h=1}^{q-1} (w_h - \tilde{w}_h) \right] = 0
\end{aligned}$$

Thus $E\tilde{C} > EC$ as required. ■

Proof of Proposition 9

Part 1: The length of the book is the smallest integer q such that $j_1^R + \sum_{k=2}^{k=q} \Psi_k \geq K$. Since j_1^R and Ψ_k increase with d_1, for all $k \geq 2$, the length of the book cannot increase when d_1 increases.

Part 2: The ex-ante expected waiting time is given by $ET = \sum_{k=1}^{k=q} (\frac{u_h}{1-u_q}) T(n_h)$. The ratio $(\frac{u_h}{1-u_q})$ does not depend on q (See Eq.(5) and (6)). Furthermore $T(n_h)$ does not depend on q. Hence each term in the sum which gives ET is unaffected by a change in the length of the book. However the number of terms increases with the length of the book. It immediately follows that ET decreases or is unchanged when d_1 increases.

Part 3: By induction on h. Let d_1^* be patient traders' waiting cost after the decrease in the order arrival rate (i.e. $d_1 < d_1^*$). We denote $\tilde{n}_h$ the h^{th} spread after this decrease. For $h = 1$ we know that:

$$n_1 = int(\frac{d_1}{\Delta}) + 1 \text{ and } \tilde{n}_1 = int(\frac{d_1^*}{\Delta}) + 1 \ ,$$

and since $d_1 < d_1^*$, then $n_1 \ \cdot \ \tilde{n}_1$. For $h = 2, ..., \tilde{q}$ ¡ 1, we have from Proposition 5:

$$\begin{aligned} n_h &= int(2r^{h-1}\frac{d_1}{\Delta}) + n_{h-1} + 1, \\ \tilde{n}_h &= int(2r^{h-1}\frac{d_1^*}{\Delta}) + \tilde{n}_{h-1} + 1. \end{aligned}$$

Therefore, from the induction hypothesis and since $d_1 < d_1^*$ we obtain that $n_h \ \cdot \ \tilde{n}_h$. Finally for $h = \tilde{q}$, we have $\tilde{n}_{\tilde{q}} = K$ and $n_{\tilde{q}} \ \cdot \ K$ (the inequality is strict if $\tilde{q} < q$) so that $n_{\tilde{q}} \ \cdot \ \tilde{n}_{\tilde{q}}$ and the result is proved. ■

Proof of Proposition 10

Assume that traders follow the same trading strategies as in the equilibrium in which they are not allowed to queue. We identify below a condition under which these strategies still form an equilibrium when traders are allowed to queue at the inside spread. Consider a patient trader who faces a spread equal to n_h. If he improves upon the inside spread, he optimally chooses a limit order which creates a spread equal to n_{h-1} given the strategies followed by the future traders. Hence, we only need to find a condition under which this trader is better off undercutting the inside spread rather than queuing at the best quotes.

Let $T(n_h, 2)$ be the expected waiting time of the trader if he decides to queue by placing an order at the inside quote. The trader is better off undercutting iff

$$n_{h-1}\Delta - T(n_{h-1})d_1 \geq n_h\Delta - T(n_h,2)d_1, \quad \forall h \geq 1,$$

or

$$(n_h - n_{h-1})\Delta \leq [T(n_h,2) - T(n_{h-1})]\, d_1 \quad \forall h \geq 1. \tag{18}$$

We now identify a condition under which this no queuing condition holds. This requires computation of $T(n_h, 2)$.

Let $T^{sa}(n_h)$ be the expected waiting time for one trader (say i) posting a spread n_h, when the next person trades in the **same** direction as trader i (for instance if trader i is a buyer then the next trader is a buyer as well). This situation never occurs on the equilibrium path. It may occur however if a trader deviates from the equilibrium strategy by deciding to queue. Hence considering this situation is helpful to compute $T(n_h, 2)$. The next trader can either be a patient trader or an impatient trader. If the trader is patient and $h \geq 2$, he submits a limit order which creates a spread equal to n_{h-1}. After an expected time equal to $T(n_{h-1})$, this order will be cleared off and the order book will be back to the initial situation. If the trader is impatient, he will submit a market order. Following this order, the new spread posted in the book can be n_{h+1} or n_h. The second case occurs when the market order is executed at one of the two border prices, A or B, (because at these prices depth is infinite). From this point on, the expected waiting time for trader i will be $T(n_{h+1})$ or $T(n_h)$. Since $T(n_h) < T(n_{h+1})$, we deduce a lower bound for $T^{sa}(n_h)$, namely

$$T^{sa}(n_h) > 1 + \theta(T(n_{h-1}) + T^{sa}(n_h)) + (1 - \theta)T(n_h) \;\; \forall h \geq 2,$$

or

$$T^{sa}(n_h) > \frac{1}{1-\theta}\left\{1 + \theta T(n_{h-1}) + (1-\theta)T(n_h)\right\} \;\; \forall h \geq 2.$$

From Proposition 4, we know that $T(n_h) - T(n_{h-1}) = 2r^{h-1}$. Furthermore $r \equiv \frac{\theta}{1-\theta}$. Using these results we can rewrite the previous inequality as:

$$T^{sa}(n_h) > 2T(n_h) \ \forall h \geq 2. \tag{19}$$

For $h = 1$, we can follow the same reasoning. The only difference is that all the traders (patient or impatient) submit a market order when they face spread with size n_1. We obtain

$$T^{sa}(n_1) > 2T(n_1) = 2.$$

Now consider a trader who decides to queue when the inside spread is n_h. The trader who is in front of him in the queue has an expected waiting time equal to $T(n_h)$. Once this trader is executed, the deviant acquires price and time priority and his expected waiting time is $T^{sa}(n_h)$. It follows that:

$$T(n_h, 2) = T(n_h) + T^{sa}(n_h).$$

Using Inequality (19), we deduce a lower bound for $T(n_h, 2)$:

$$T(n_h, 2) > 3T(n_h) \ \forall h \geq 2.$$

For $h = 1$, we obtain

$$T(n_1, 2) > T(n_1) + 2T(n_1) = 3.$$

Substituting the lower bound for $T(n_h, 2)$ in Condition (18), we rewrite the no queuing condition as

$$(n_h - n_{h-1})\Delta \leq \{3T(n_h) - T(n_{h-1})\} d_1 \quad \forall h \geq 2. \tag{20}$$

Furthermore, using Proposition 5, we deduce that

$$(n_h - n_{h-1})\Delta < \Delta + 2r^{h-1} d_1$$

Hence a sufficient condition for Condition (20) is

$$\Delta + 2r^{h-1} d_1 \leq \{3T(n_h) - T(n_{h-1})\} d_1, \forall h \geq 2.$$

or

$$\Delta \leq 2T(n_h)d_1 \;\; \forall h \geq 2. \tag{21}$$

For $h = 1$, we follow the same reasoning and we obtain that a sufficient condition for no queuing when the inside spread is n_1 is

$$\Delta \leq 2T(n_1)d_1. \tag{22}$$

If Condition (22) holds true then Condition (21) is satisfied as well since $T(n_h)$ increases in h. Thus the sufficient condition for no queuing at any inside spread is:

$$\frac{d_1}{\Delta} \geq \frac{1}{2}.$$

■