

Securities market structure, trading fees and investors' welfare.*

Jean-Edouard Colliard
Paris School of Economics
48 boulevard Jourdan
75014 Paris, France
colliard@ens.fr

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

November 2010

Abstract

We consider a riskless asset (a “zero coupon bond”) that trade on a dealer (OTC) market or a limit order market. In the limit order market, investors can choose to be "makers" (post limit orders) or "takers" (hit limit orders) whereas in the dealer market they must trade at dealers' quotes. Moreover, in the limit order market, investors pay a trading fee to the operator of this market (“the matchmaker”). We show that, for some parameter values, an increase in the matchmaker's trading fee can raise investors' ex-ante expected welfare. Actually, it forces makers to post more aggressive offers and thereby it raises the likelihood of a direct trade between investors. Thus, a reduction in the matchmaker's trading fee (due for instance to increased competition between trading platforms) can counter-intuitively raise the OTC market share and impair investors' welfare. However, investors are always better off with a zero trading fee rather than the fee set by a for-profit monopolist matchmaker. Finally, the model has testable implications for the effects of a change in trading fees and their breakdown between makers and takers on various measures of market liquidity.

Keyword: Limit order markets, trading fees, make/take fees, inter-market competition, liquidity, OTC markets, securities market structure.

*We thank Bruno Biais, Hans Degryse, Stefano Lovo, Albert Menkveld, Sophie Moinas, Mark Van Achter and participants at the conference on the Industrial Organization of Securities Markets in Frankfurt and seminars at the Toulouse School of Economics and the Paris School of Economics for their comments. All errors are ours.

1 Introduction

The industrial organization of securities markets is changing fast, both in Europe and in North America. For instance, in recent years, new trading platforms (BATS, Chi-X, EdgeX, Turquoise etc...) have challenged incumbent stock exchanges (e.g., NYSE-Euronext, the Nasdaq, the London Stock Exchange etc...). As a consequence, incumbents' market share has declined and trading is more fragmented. For instance, the market share of the NYSE (as a fraction of total trading volume) fell from 66% in January 2007 to about 40% in March 2009. A similar evolution is observed in Europe where, for instance, the market share of the London Stock Exchange dropped from about 70% to 60% in 2009.

The new platforms are often organized as limit order markets where traders can submit limit orders (post prices at which they are willing to trade) or submit market orders (hit limit orders). Platforms refer to investors submitting limit orders as *makers* and to investors submitting market orders as *takers*. Indeed, makers “build up” the liquidity of the market by posting offers while takers “consume” this liquidity by hitting offers. Platforms earn a fee each time a maker is matched with a taker and makers and takers are often charged different fees.¹ As competition among platforms is heating up, these fees have steadily declined.

The effects of this evolution are very much debated but not well understood yet. For instance, to date, the SEC has received more than two hundred comments on its concept release regarding the organization of U.S. equities markets and many of these comments address the question of fees.² Similarly, a recent consultation paper by the Committee of European Securities Regulators (CESR) raises the following question: “*What are the impacts of current fee structures on trading platforms, participants, their trading strategies and the wider market and its efficiency?*”³

Addressing these questions requires to understand how trading fees affect the “make or take decision,” (the choice between a market or a limit order) and the impact of this decision on investors' welfare. However, this analysis does not exist in the literature. Our

¹For instance, in 2009, in each transaction, NYSEArca (a trading platform owned by the NYSE) was charging \$0.0030 (per share) to the taker and rebating \$0.0024 (per share) to the maker. The net revenue for NYSEArca was therefore \$0.0006 (per share traded).

²See <http://www.sec.gov/comments/s7-02-10/s70210.shtml>

³See “Call for Evidence: Micro-Structural Issues of the European equity markets.” Available at <http://www.cesr.eu/>.

goal here is to fill this gap. Our key and new finding is that an increase in trading fees can, surprisingly, *increase* investors' welfare because it forces makers to post offers with a higher fill rate. This effect reduces the likelihood that investors resort to an intermediary (a dealer), which improves welfare (as intermediation involves a dead-weight cost). Thus, surprisingly, unbridled competition among trading platforms may lead to *too low* trading fees, even from investors' point of view. Moreover, we use the model to compare investors' welfare in different market structures. Interestingly, we find that, for some parameter values, investors' welfare can be reduced by the presence of dealers who stand ready to buy or sell the security continuously.

Our model features the market for a riskless security populated by buyers (investors with a high private value for the security) and sellers (investors with a low private value).⁴ Buyers and sellers arrive sequentially and have a deadline to carry out their trade. In our baseline model, upon arrival, an investor can choose to trade either in a dealer market or in a limit order market. This case is relevant since, in equities markets, limit order markets often face competition from over the counter (OTC) dealer markets.⁵ In the limit order market, the investor can choose to submit a market order (act as taker) or to post a limit order (act as maker). With a limit order, he obtains a better execution price but he runs the risk that his order will remain unfilled by the time his deadline is reached. If this happens, the investor can, in last resort, trade in the dealer market before leaving the market.

Dealers in the OTC bear an order processing cost each time they execute a trade. As they make zero profits, the cost of trading in the OTC market is entirely determined by the size of the order processing cost. In contrast, in the limit order market, investors bear no order processing cost since, in this market structure, no intermediaries use resources to step in between final investors. However, investors prefer fast to slow execution, *other things equal*. Thus, liquidity provision in the limit order market is not free: makers bear a "waiting cost" for which they must receive a compensation. Moreover, we assume that each time a maker and a taker are matched in the limit order market, they pay a fee to

⁴Our model of limit order trading builds upon Foucault (1999). However, this model considers a single limit order market and does not study how trading fees affect the make-take decision for investors.

⁵The fraction of trades taking place in these OTC markets can be significant. For instance, for stocks constituents of the FTSE 100, OTC trades account for about 46% of all trades as of November 2009 (source: Thomson-Reuters).

the operator of this market (the “matchmaker”). As observed in practice, we allow this fee to be different for a maker and a taker and we consider both the effects of changing the make/take fee breakdown and the total fee.

As a benchmark, we first analyze the allocation of matches (between investors or between investors and dealers) that maximizes investors’ welfare in this setting (the “unconstrained first best” for investors). In any trading mechanism, incentives compatibility constraints will in general prevent investors from achieving this allocation.⁶ Thus, this benchmark yields an upper bound for the gains from trade that investors can obtain in the trading mechanisms considered in the paper. Not surprisingly, in this case, the trading fee must be set to zero since a positive fee simply reduces the gains from trade available to investors.

This benchmark is useful to delineate the sources of inefficiencies that arise in the market structures we consider. We identify two sources of inefficiencies. First, the dealer market can “crowd out” the limit order market “too frequently” relative to the first best. Intuitively, it can be efficient to concentrate all trades in the dealer market since this is a way to save on waiting costs. But this is more efficient only if the order processing cost is low enough, i.e., below a threshold. Instead, in equilibrium, the liquidity of the limit order market dries up for values of the order processing cost well above this threshold. The reason is that makers compete with dealers. Hence, the dealer market reduces the rents makers can extract from takers and can therefore lead to an outcome where it is never individually optimal to submit a limit order. In this case, the limit order market breaks down.

The second source of inefficiency is more mundane: makers can strategically choose limit orders with low execution probabilities to obtain a higher surplus in case of execution. As a consequence, the rate at which trades happen on the limit order market can be too low relative to the rate required to achieve the first best for investors. In this case, it is possible to improve investors’ welfare in equilibrium by raising the trading fee. To see why, consider an increase in the matchmaker’s trading fee. This increase reduces the surplus to be split between makers and takers. However, takers can claim a higher fraction of this

⁶It is worth stressing that the “first best” here is the allocation of trades that maximizes investors’ welfare assuming that the central planner choosing this allocation has full knowledge of the investors’ type. Clearly, no mechanism can do better than this allocation.

dwindling “cake.” Indeed, their outside option (an immediate trade in the dealer market) becomes relatively more attractive and therefore their market power is higher. For some parameter values, this shift in market power from makers to takers forces the former to post more aggressively priced limit orders, i.e., limit orders that have a higher likelihood of execution. As a result, for intermediate values of dealers’ order processing cost, an increase in the matchmaker’s trading fee can, counter-intuitively, make the rate of trades on the limit order market closer to the first best and raise investors’ welfare.

In the last part of the paper, we contrast two types of market structures: (i) a single matchmaker and a dealer market and (ii) two competing matchmakers and a dealer market. We first derive the optimal pricing policy of the matchmakers in each case. Competition among matchmakers drives the trading fee to zero while, not surprisingly, a single matchmaker uses its monopoly power to charge a higher trading fee. However, a single matchmaker has one benefit for investors: it reduces the range of values for the order processing cost such that the matching rate on the limit order market is too low. The reason is that the matchmaker has all incentives to set a high trading fee if this does not come at the cost of a lower trading rate. Despite this benefit, we show that investors’ welfare is higher with two competing matchmakers because a single matchmaker leaves too little surplus to investors. We also show that for some parameter values investors’ welfare is higher when the dealer market is shut down.

The model yields a rich crop of predictions that could be tested using the ongoing battle among stock markets in the U.S. and in Europe. The most surprising and novel predictions come from the effect of the trading fee on limit order execution probabilities and investors’ welfare. First, the model implies that, for some parameter values, a decrease in the trading fee can trigger a decrease in limit orders’ execution probabilities, other things equal. As a consequence, the market share of the dealer market increases, despite the fact that the cost of trading in the limit order market (even including the bid-ask spread) has declined. Testing for this effect provides a sharp test of our theory since it is clearly non standard. Interestingly, the recent entry of new trading platforms in European equities markets seems to coincide with both lower trading fees and an increase in the market share of the OTC market (the dealer market in our set-up). This evolution is puzzling for many analysts but it is a possible outcome in our model.

Second, and relatedly, a decrease in trading fee can lower investors' welfare for reasons explained previously. Thus, an intensification of competition among matchmakers does not necessarily improve investors' welfare. Hollifield et al.(2006) empirically estimate investors' welfare in a limit order market (the Vancouver Stock Exchange). Interestingly, they find empirically that the opportunity cost associated with unfilled limit orders is the main source of inefficiency in a limit order market. In our model, an intensification of inter-market competition (entry of a new limit order market or a decrease in order processing costs) can sometimes lower ex-ante gains from trade precisely because it results in smaller fill rates for limit orders. Hollifield et al.(2006)'s methodology could be used to test this prediction.

Last, the model makes predictions about the effect of a change in make/take fees on the bid-ask spread. In equilibrium, the "raw" traded bid-ask spreads (i.e., the difference between ask and bid prices at which trades take place) decreases in the take fee and increases in the make fee. Thus, an increase in the total fee can lead to wider or tighter raw bid-ask spreads depending on whether the take or the make fee increases. In contrast, the *cum fee* bid-ask spread (i.e., the difference between the ask price plus the take fee and the bid price minus the take fee) always increases in the total fee and is independent of the make/take fee breakdown. For instance, when the take fee increases, makers must post more attractive bids to prevent investors acting as takers from switching to being makers. But the reduction in bid-ask spread is less than the increase in the take fee so that the burden of an increase in the take fee is borne by makers and takers.

Our analysis is related to theories of "competition for order flow" in securities markets (e.g., Pagano (1989), Glosten (1994), Hendershott and Mendelson (2000), Parlour and Seppi (2003), Viswanathan and Wang (2002), Foucault and Menkveld (2008) or Degryse et al.(2009)).⁷ These theories usually do not consider the possibility for investors to act as a maker or a taker. Thus, they do not analyze how this choice is affected by a change in trading fee and more generally by a change in the degree of competition between market platforms as we do here. Instead, the literature has focused on liquidity externalities and network effects (e.g., Pagano (1989) or Hendershott and Mendelson (2000)), which are

⁷There is also a rich empirical literature on this topic (e.g., Barclay et al.(2003), Biais et al. (2004), Boehmer and Boehmer (2004), Defontnouvelle et al.(2003), O'Hara and Ye (2009), or Cantillon and Yin (2010)).

absent from our analysis.

More generally, our paper contributes to the literature on competition between markets (e.g., Yavas (1992), Gehrig (1993), Silber (1996) or Rust and Hall (2006)). This literature also takes trading fees as given (or ignores these fees). In contrast we explicitly model the choice of their trading fee by competing platforms. This is important to analyze the efficiency of various trading arrangements. Last our paper is also related to Degryse et al.(2010). They consider the effect of clearing and settlement fees on investors' order placement strategies in a single market environment. Their approach is complementary since clearing and settlement fees add to the trading fee paid by investors to trading platforms. However, they are in general not under the control of trading platforms.⁸

The paper is organized as follows. Section 2 describes the model. Section 3 derives the equilibrium of the model. Section 4 analyze the implications of the model for liquidity and investors' welfare. Then in Section 5, we derive the optimal pricing policy of a matchmaker in various market structures and we compare investors' welfare in these market structures. Section 6 concludes.

2 Model

2.1 Market participants

Buyers and Sellers. We consider the market for a riskless security that pays a single cash flow v_0 at a random date \tilde{T} . Specifically, at each date $t = 0, 1, 2, \dots$, there is a probability $(1 - \rho)$ that the asset pays its cash-flow. If date t is not the terminal date, then a new investor arrives in the market to buy or sell one share of the security. The investor has a deadline of one period to carry out his transaction, after which he leaves the market forever. An investor's valuation for the security is either high, $v_H = v_0 + L$ or low, $v_L = v_0 - L$ with equal probabilities.⁹ Investors with a high valuation want to buy the security whereas investors with a low valuation want to sell it. We denote by $\Delta v \stackrel{def}{=} v_H - v_L = 2L$, the size of the gains from trade between buyers and sellers.

⁸For instance, all trades in U.S. equities markets are cleared and settled by unique agencies (respectively, NSCC and DTCC).

⁹Heterogeneity in investors' private value generates gains from trade as in many other models of limit order trading (e.g., Goettler et al.(2009) or Hollifield et al.(2004)). See Duffie et al. (2005) for economic interpretations.

Investors also differ in terms of impatience: patient investors' discount factor is $\widehat{\delta}_H$ whereas impatient investors' discount factor is $\widehat{\delta}_L < \widehat{\delta}_H$ with $\widehat{\delta}_L > 0$. The fraction of patient investors is denoted by π . In practice, investors' preference for quick execution arises from the need to synchronize trades across different securities (e.g., for arbitrageurs) or replicate a security (e.g., for index fund managers). The discount factor captures this preference for quick execution rather than the time value of money (as in Foucault et al.(2005) or Goettler et al.(2009)).

Trading venues. Each investor can trade either in a *dealer market* or in a *limit order market*, as shown on Figure 1.

Insert Figure 1 here

The Dealer Market (DM). In this market, dealers continuously post ask and bid prices denoted A^m and B^m at which they stand ready to buy or sell one share of the security. These market-makers value the security at v_0 . To process an order, they bear a cost $\lambda < L$. Hence, competition among market-makers implies that:

$$A^m = v_0 + \lambda,$$

$$B^m = v_0 - \lambda.$$

When he contacts a dealer, an investor buys or sells the security at the dealer's quotes and exits the market forever. Investors do not pay a fee to trade in the dealer market.¹⁰ Thus, when a trade takes place on the dealer market, the surplus accruing to the investor is $G^d = L - \lambda$ and the dealer has a zero profit. The order processing cost, λ , is therefore a deadweight cost. If $\lambda \geq L$, investors cannot trade at a profit with dealers and the dealer market is inactive.

The Limit Order Market (LOM). Alternatively, the investor can choose to trade in the limit order market. He must then choose to submit either a limit order or a market order. If an investor submits a buy (resp. sell) market order then the investor immediately trades at the best available ask (resp. bid) price and exits the market. If instead the investor submits a limit order, he specifies a bid or an ask price at which he is willing to trade. This offer is stored in the limit order book, waiting for future execution. Limit

¹⁰In dealer markets, commissions are in general factored into quotes.

orders are valid for one period since this is the deadline of all investors. Thus, either a limit order is filled after one period or it is cancelled. If the limit order is cancelled, then the investor trades with a dealer and exits the market.¹¹ That is, investors with unfilled limit orders use the dealer market *in last resort*. Following the terminology used by trading platforms, we call "*makers*" the investors posting quotes and "*takers*" the investors hitting quotes.

The *limit order book* is the set of offers posted in the limit order market at any point in time. As limit orders are valid for only one period, at each date t , the limit order book has three possible states: (i) it contains a sell limit order, (ii) it contains a buy limit order, or (iii) it is empty. Let A_t and B_t be the ask and bid prices posted in the limit order market at the beginning of period t . When an investor has posted a sell (resp. buy) limit order at date $t - 1$, then A_t (resp. B_t) is endogenous and will be determined below. Otherwise if there is no sell limit order in the book, we set $A_t = \bar{A} = +\infty$. Similarly, if there is no buy limit order in the book we set $B_t = \bar{B} = -\infty$.

The owner of the limit order market (the "matchmaker") collects a fee, $f^T \geq 0$, each time a transaction occurs. This fee is split between each side (maker/taker) of a transaction as follows: the taker pays f_m and the maker pays f_l so that $f^T = f_l + f_m$. Following practice, we refer to f_l as being the "*make fee*" and to f_m as being the "*take fee*". For simplicity, we set the cost of processing trades for the matchmaker to zero. We denote by

$$G^l \stackrel{def}{=} \Delta v - f^T = 2L - f^T,$$

the size of the gains from trade net of the fees charged by the matchmaker. When a trade takes place on the limit order market, the total surplus is $G^l + f^T > G^d$, i.e., conditional on a trade, the limit order market is a more efficient technology to match buy and sell orders.¹²

¹¹For simplicity, we assume that if a limit order is unfilled at, say, date t , there is a small delay (less than one period) between the moment at which the investor with the unfilled limit order exits the market (after trading in the dealer market) and the moment at which a new limit order at date t (if any) is submitted in the limit order book. In this way, we avoid the problem that the investor with the unfilled limit order may want to retrade against the new limit order in the book.

¹²Studies of bid-ask spreads on Nasdaq and the NYSE when these markets were, respectively, similar to a dealer market and a limit order market have shown that the average bid-ask spread on Nasdaq was higher than on the NYSE, in part because real costs of intermediation were higher on Nasdaq (see Stoll (2000)). The real cost of intermediation in a dealer market includes labor costs but also the cost of capital associated with inventory risk. This cost is absent from our model but will add up to the cost of intermediation in the dealer market. Fink et al.(2006) also provides evidence consistent with the view that limit order markets are less costly trading technologies.

To sum up, when they arrive in the market, investors can trade immediately, either at dealers' quotes or at the best quotes posted in the limit order market. Alternatively, they can post an offer to buy or sell the security in the limit order market. In this case, they take the risk that this offer will not be hit. When an investor is indifferent between the two trading venues, we assume that he trades on the limit order market.

Payoffs. Let $\delta_i \equiv \rho \widehat{\delta}_i$. For brevity, we will refer to δ_i as investor i 's discount factor. Consider a buyer with a discount factor δ_i who arrives at date t . If he contacts a dealer upon arrival, he obtains a payoff

$$G^d = v_H - A_m = L - \lambda \geq 0.$$

If instead, the buyer submits a buy market order, his payoff is

$$U_m^{bu}(A_t, f_m) \stackrel{def}{=} v_H - A_t - f_m, \quad (1)$$

and if he posts a buy limit order at price B , his expected payoff is

$$U_l^{bu}(B, f_l, \delta_i) \stackrel{def}{=} \delta_i [\phi_{ex,t}^{bu}(B)(v_H - B - f_l) + (1 - \phi_{ex,t}^{bu}(B))G^d], \quad (2)$$

where $\phi_{ex,t}^{bu}(B)$ is the execution probability of a buy limit order posted at price B at date t , conditional on continuation of the trading game at date $t + 1$.¹³

Similarly, the payoff of a seller who contacts a dealer is G^d . If instead the seller submits a market order at date t , her payoff is:

$$U_m^{se}(B_t, f_m) \stackrel{def}{=} B_t - v_L - f_m, \quad (3)$$

whereas her payoff with a sell limit order at price A is:

$$U_l^{se}(A, f_l, \delta_i) \stackrel{def}{=} \delta_i [\phi_{ex,t}^{se}(A)(A - v_L - f_l) + (1 - \phi_{ex,t}^{se}(A))G^d], \quad (4)$$

where $\phi_{ex,t}^{se}(A)$ is the execution probability of a sell limit order posted at price A at date t , conditional on continuation of the trading game at date $t + 1$.

Timing. We solve the model backward. First, for fixed fees of the trading platform, we solve for investors' order placement strategies in equilibrium (see below). Second, we solve for the optimal fees of the trading platform (f_l , f_m and f^T). In choosing its fees, the trading platform correctly anticipates how these fees affect traders' order submission choices.

¹³If the asset pays off at date $t + 1$, the limit order posted at date t does not execute and the investor posting this order gets a zero payoff.

2.2 Possible regimes for the limit order market

Consider a buyer arriving at date t when the ask price in the market is A_t . Let B_t^* be the optimal bid price of this buyer if he submits a limit order. By definition

$$B_t^* \in \text{Argmax}_B U_l^{bu}(B, f_l, \delta_i).$$

When the buyer arrives, he chooses one of three options: (i) a buy market order on the limit order market, (ii) a buy limit order at price B_t^* , or (iii) a buy order in the dealer market. We denote these three options: b_m , B_t^* and b_d , respectively. The optimal bid price for a buyer does not depend on his discount factor since this factor simply scales the payoff of a buy limit order (see equation (2)).¹⁴ But the choice among the three possible decisions for the investor depends on the discount factor. The buyer's choice is denoted by $O_b(\delta_i, A_t) \in \{b_m, B_t^*, b_d\}$.

Now consider a seller who arrives at date t when the bid price in the market is B_t . We denote by A_t^* the optimal ask price posted by this investor if she submits a sell limit order. She has three options: (i) a sell market order on the limit order market, (ii) a sell limit order at price A_t^* or (iii) a sell order in the dealer market. We denote these three options: s_m , A_t^* and s_d , and we denote the seller's choice among these options by $O_s(\delta_i, B_t) \in \{s_m, A_t^*, s_d\}$. We refer to $O_s(\cdot)$ and $O_b(\cdot)$ as the sellers and the buyers' order placement strategies.

An equilibrium is a set of order placement strategies $O_s^*(\cdot)$ and $O_b^*(\cdot)$ such that (i) $O_s^*(\delta_i, B_t)$ maximizes the expected payoff of a seller with type δ_i when she arrives in the market given that she expects other participants to follow strategies $O_s^*(\cdot)$ and $O_b^*(\cdot)$ and (ii) $O_b^*(\delta_i, A_t)$ maximizes the expected payoff of a buyer with type δ_i when he arrives in the market given that he expects other participants to follow strategies $O_s^*(\cdot)$ and $O_b^*(\cdot)$. Investors' order placement strategies do not depend on time and the history of the market until their arrival date. Thus, we focus on Markov perfect equilibria. This focus is natural since investor' payoffs do not directly depend on the history of the game (see Maskin and Tirole (1997)).

At date t , the Bellman equation for a buyer's problem is

$$V_b(\delta_i, A_t) = \text{Max}\{v_H - A_t - f_m, U_l^{bu}(B_t^*, f_l, \delta_i), G^d\}$$

¹⁴In fact for a given bid B , we have $U_l^{bu}(B, f_l, \delta_L) = \left(\frac{\delta_L}{\delta_H}\right) U_l^{bu}(B, f_l, \delta_H)$.

Let $A^{r*}(\delta_i)$ be such that

$$v_H - A_t^{r*}(\delta_i) - f_m = \text{Max}\{U_l^{bu}(B_t^*, f_l, \delta_i), G^d\} \quad (5)$$

The buyer optimally picks a buy market order iff $A_t \leq A^{r*}(\delta_i)$. Thus, $A_t^{r*}(\delta_i)$ is the highest ask price at which the buyer is willing to submit a market order on the limit order market.

We refer to $A_t^{r*}(\delta_i)$ as the buyer's cut-off price at date t .

Similarly, a seller will submit a sell market order iff $B_t \geq B_t^{r*}(\delta_i)$ where

$$B_t^{r*}(\delta_i) - v_L - f_m = \text{Max}\{U_l^{se}(A_t^*(v_L), f_l, \delta_i), G^d\}. \quad (6)$$

Thus, $B_t^{r*}(\delta_i)$ is the smallest bid price at which the seller is willing to submit a market order on the limit order market. We refer to this bid as the seller's cut-off price.

Lemma 1 *Buyers' cut-off prices decrease in δ_i and sellers' cut-off prices increase in δ_i . That is, $A_t^{r*}(\delta_H) \leq A_t^{r*}(\delta_L)$ and $B_t^{r*}(\delta_H) \geq B_t^{r*}(\delta_L)$.*

Impatient investors are more willing to pay a concession to trade upon arrival since they receive a smaller expected payoff with a limit order. Thus, impatient buyers' (resp. sellers') cut-off price is higher (resp., lower) than patient buyers' cut-off price. Now consider a sell limit order posted at price A_t . This order executes at date t only if the investor arriving at this date is a buyer with a cut-off price higher than A_t . As $A_t^{r*}(\delta_H) < A_t^{r*}(\delta_L)$, we deduce that the execution probability of the sell limit order placed at A_t is

$$\phi_{ex}^{se}(A_t) = \begin{cases} \frac{\rho}{2} & \text{if } A_t \leq A^{r*}(\delta_H), \\ \frac{(1-\pi)\rho}{2} & \text{if } A^{r*}(\delta_H) < A_t \leq A^{r*}(\delta_L), \\ 0 & \text{if } A_t > A^{r*}(\delta_L). \end{cases} \quad (7)$$

Similarly, the execution probability of a buy limit order posted at price B_t is:

$$\phi_{ex}^{bu}(B_t) = \begin{cases} \frac{\rho}{2} & \text{if } B_t \geq B^{r*}(\delta_H), \\ \frac{(1-\pi)\rho}{2} & \text{if } B^{r*}(\delta_L) \leq B_t < B^{r*}(\delta_H), \\ 0 & \text{if } B_t < B^{r*}(\delta_L). \end{cases} \quad (8)$$

Thus, when a buyer chooses a limit order, he optimally posts either a high bid price equal to $B^{r*}(\delta_H)$ or a low bid price equal to $B^{r*}(\delta_L)$. We refer to the first bid as being a high fill rate limit order and the second bid as being a low fill rate limit order. The order with a low fill rate yields a greater surplus in case of execution but it executes less frequently. Similarly,

makers on the sell side choose either a limit order with a high fill rate ($A^* = A^{r^*}(\delta_H)$) or a low fill rate ($A^* = A^{r^*}(\delta_L)$).

An investor will not use a limit order if he can obtain a larger expected trading profit by trading immediately on the dealer market. This is the case for a buyer of type δ_i iff:

$$\text{Max}\{U_i^{bu}(B^{r^*}(\delta_H), f_l, \delta_i), U_i^{bu}(B^{r^*}(\delta_L), f_l, \delta_i)\} < G^d. \quad (9)$$

Similarly, a seller with type δ_i never submits a limit order iff:

$$\text{Max}\{U_i^{se}(A^{r^*}(\delta_H), f_l, \delta_i), U_i^{se}(A^{r^*}(\delta_L), f_l, \delta_i)\} < G^d. \quad (10)$$

If these conditions are satisfied for all investors then the dealer market “crowds out” the limit order market: investors never submit a limit order and therefore no trade happens on the limit order market. Otherwise, since impatient investors obtain a smaller expected payoff when they use limit orders, there are two possibilities: (i) conditions (9) and (10) are satisfied for patient investors only or (ii) conditions (9) and (10) are satisfied for patient and impatient investors. In the first case, only patient investors act as makers while in the second case both patient and impatient investors act as makers (for some states of the limit order book). We refer to equilibria of the first type as "specialized equilibria" and to equilibria of the second type as "unspecialized equilibria."

In summary, there are five possible types of equilibria (“regimes”) for the market:¹⁵

1. **Unspecialized/High Fill Rate (type #1):** The equilibrium is unspecialized and when they submit a limit order, investors choose a limit order with high execution probability. Hence on the *equilibrium path*, patient and impatient investors submit limit orders when the limit order book lacks liquidity on their side and market orders otherwise. They only use the dealer market in last resort.
2. **Unspecialized/Low Fill Rate (type #2):** The equilibrium is unspecialized and when they submit a limit order, investors choose a limit order with a low execution probability. Hence on the *equilibrium path*, patient investors always submit a limit

¹⁵For brevity, for each type of equilibrium, we just describe investors’ actions on the "equilibrium path," i.e., given the states of the limit order book that arise in equilibrium. Of course, a full description of investors’ order placement strategies in equilibrium require to specify their action for all possible states of the limit order book, even those that are unobserved on the equilibrium path. This specification is readily deduced from the type of equilibrium (specialized/unspecialized) and the type of limit orders (high fill rate/low fill rate) that investors use.

order when they arrive in the market. Impatient investors submit a limit order if the limit order book lacks liquidity on their side and a market order otherwise. All investors only use the dealer market in last resort.

3. **Specialized/Low Fill Rate (type #3):** The equilibrium is specialized. Patient investors behave as in the unspecialized equilibrium with a low fill rate and are the sole investors submitting limit orders. Impatient investors never submit a limit order. They contact the dealer if the limit order book lacks liquidity on their side when they arrive in the market and they submit a market order otherwise.
4. **Specialized/High Fill Rate (type #4):** The equilibrium is specialized. Impatient investors behave as in the specialized equilibrium with a low fill rate and patient investors behave as in an unspecialized equilibrium with a high fill rate.
5. **Dealer Market Only (type #5):** The dealer market crowds out the limit order market. When they arrive in the market, all investors immediately trade in the dealer market.

In the rest of the paper we focus on the case in which parameter values satisfy the following condition:

$$\mathbf{C.1:} \quad \frac{2\pi}{1-\pi}(1-\delta_L) < \delta_H - \delta_L < \frac{2\pi}{1-\pi}. \quad (11)$$

Note that this condition requires $\pi \leq \frac{1}{3}$ since $\delta_j \in (0, 1]$. As shown in the next section, under Condition **C.1**, each type of equilibrium can occur. Thus, by focusing on this range of parameters, our analysis covers all possible cases that can emerge in equilibrium. In contrast, for other parameter values, only a subset of all possible equilibria will emerge. Sometimes, for brevity, we shall refer to an equilibrium by its shorthand, e.g., a type #1 for the unspecialized/high fill rate equilibrium. We label a type #5 equilibrium, the equilibrium in which the limit order market is inactive.

3 Equilibria

We first describe the type of equilibrium that is obtained for each values of the parameters.

To this end, we define the following variables $\kappa_1 = \frac{2\pi - (1-\pi)(\delta_H - \delta_L)}{2\pi + \delta_H(1+\pi) - \delta_L(1-\pi)}$, $\kappa_2 = \frac{\delta_L(1-\pi)}{2(1-\delta_L\pi)}$,

$\kappa_3 = \frac{\delta_H(1-\pi)-2\pi}{2(1-2\pi-\delta_H\pi)}$, $\kappa_4 = \frac{\delta_H}{2}$. Under Condition C.1, $\kappa_1 \leq \kappa_2 \leq \kappa_3 \leq \kappa_4$. Let $\Gamma(\lambda, f^T) \equiv \frac{G^d}{G^l} = \frac{L-\lambda}{2L-f^T}$.

Proposition 1 *The values of the parameters being fixed, there is a unique Markov Perfect Equilibrium. The type of the equilibrium is as follows:*

1. *When $\Gamma(\lambda, f^T) \leq \kappa_1$ then the unique equilibrium is an unspecialized/high fill rate equilibrium. In this equilibrium, bid and ask quotes posted by sellers and buyers are*

$$\begin{aligned} A^* &= v_H - f_m - \frac{\delta_H}{2 + \delta_H}(G^l + G^d), \\ B^* &= v_L + f_m + \frac{\delta_H}{2 + \delta_H}(G^l + G^d). \end{aligned}$$

2. *When $\kappa_1 < \Gamma(\lambda, f^T) \leq \kappa_2$ then the unique equilibrium is an unspecialized/low fill rate equilibrium. In this equilibrium, bid and ask quotes posted by buyers and sellers are*

$$\begin{aligned} A^* &= v_H - f_m - \frac{\delta_L}{2 + \delta_L(1 - \pi)} \left((1 - \pi)G^l + (1 + \pi)G^d \right), \\ B^* &= v_L + f_m + \frac{\delta_L}{2 + \delta_L(1 - \pi)} \left((1 - \pi)G^l + (1 + \pi)G^d \right). \end{aligned}$$

3. *When $\kappa_2 < \Gamma(\lambda, f^T) \leq \kappa_3$ then the unique equilibrium is a specialized/low fill rate equilibrium. In this equilibrium, bid and ask quotes posted by buyers and sellers are*

$$\begin{aligned} A^* &= v_0 - f_m + \lambda, \\ B^* &= v_0 + f_m - \lambda. \end{aligned}$$

4. *When $\kappa_3 < \Gamma(\lambda, f^T) \leq \kappa_4$ then the unique equilibrium is a specialized/high fill rate. In this equilibrium, bid and ask quotes posted by investors are as in the unspecialized/high fill rate equilibrium (Case 1).*

5. *When $\Gamma(\lambda, f^T) > \kappa_4$, the limit order market is inactive.*

Thus, the equilibrium type is determined by the position of $\Gamma(\lambda, f^T)$ relative to the thresholds, κ_j . Hence, ultimately, the type of equilibrium depends on dealers' order

Order Processing Cost: λ	Matchmaker's fee: f^T			
	0	0.5	0.9	1.25
0.7	(#1, 50%, 33%)	(#2, 40%, 29%)	(#3, 40%, 8%)	(#4, 50%, 9%)
0.6	(#2, 40%, 29%)	(#3, 40%, 8%)	(#4, 50%, 9%)	#5
0.5	(#3, 40%, 8%)	(#4, 50%, 9%)	#5	#5
0.4	(#4, 50%, 9%)	(#4, 50%, 9%)	#5	#5

Table 1: Equilibrium outcomes for various values of the order processing cost in the dealer market and the fee in the limit order market. For each value of (λ, f^T) shown in the table, we give (i) the equilibrium type, (ii) the execution probability of a limit order and (iii) the trading rate (i.e., the unconditional probability of a trade on the platform). Other parameter values are $L = 1$, $\delta_H = 0.8$, $\delta_L = 0.5$ and $\pi = 0.2$.

processing cost, λ , and the total fee charged by the matchmaker, f^T . To see this more clearly, let $\kappa_0 = 0$,

$$\lambda_k \equiv L(1 - 2\kappa_k), \quad (12)$$

and

$$f_k^T(\lambda) = \frac{\lambda - \lambda_k}{\kappa_k}. \quad (13)$$

Observe that $f_k^T(\lambda)$ increases in κ_k so that $f_1^T(\lambda) \leq f_2^T(\lambda) \leq f_3^T(\lambda) \leq a_4^T(\lambda)$. The condition $\kappa_{k-1} < \Gamma(\lambda, f^T) \leq \kappa_k$ is equivalent to $f_{k-1}^T(\lambda) < f^T \leq f_k^T(\lambda)$. Thus, for a fixed value of λ , an equilibrium of type k is obtained iff $f_{k-1}^T(\lambda) < f^T \leq f_k^T(\lambda)$ as shown on Figure 2. It is worth stressing that for $f^T = 0$, an equilibrium of type k is obtained iff $\lambda_k < \lambda \leq \lambda_{k-1}$ since $f_k^T(\lambda)$ increases with λ and is zero for $\lambda = \lambda_k$.

[Insert Fig.2 here]

To illustrate the effect of a change in λ or f^T on the equilibrium type, we first consider a numerical example. Parameters are set at: $L = 1$, $\pi = 0.2$, $\delta_H = 0.8$, $\delta_L = 0.5$. In this case, Table 1 gives for various pairs (λ, f^T) , (i) the equilibrium type, (ii) the execution probability for limit orders and (iii) the trading rate on the limit order market, i.e., the unconditional probability that a trade takes place on this market in each period (we explain in Section 4.2 how this trading rate is obtained).

Table 1 shows that limit orders' fill rate can increase when the matchmaker's fee becomes higher or the order processing cost in the dealer market, λ , becomes smaller. For

instance, when $\lambda = 0.5$, an increase in the trading fee charged by the limit order market from $f^T = 0$ to $f^T = 0.5$ leads to an increase in the fill rate on the limit order market. Or, when $f^T = 0.5$, a decrease in the order processing cost on the dealer market from $\lambda = 0.6$ to $\lambda = 0.5$ leads to an increase in the fill rate on the limit order market. As we shall see, these effects are important to understand the impact of a change in the trading fee or market structure on investors' welfare.

To gain intuition, suppose first that $f^T = 0$. For high values of λ , the dealer market is not very competitive. In this case, as shown in Table 1, the equilibrium is unspecialized and the fill rate is high (type #1). When λ declines, a trade on the dealer market yields a higher surplus. Thus, the cost of an unfilled limit order is reduced and submitting a limit order becomes more attractive for newcomers. A consequence, makers' market power declines and makers must leave a high surplus to patient takers to keep their execution probability high. For this reason, for intermediate values of λ ($\lambda \in [\lambda_3, \lambda_1]$), makers submit limit orders unattractive for patient investors but that secure a high surplus in case of execution. As these limit orders are hit only by impatient investors, they have a low execution probability. For low values of λ ($\lambda \in [\lambda_4, \lambda_3]$), dealers' quotes are very competitive and makers obtain a small surplus even if they choose quotes with a low fill rate. Thus, they optimally switch back to a high fill rate strategy since the loss in surplus is small compared to the increase in fill rate. For this reason, limit orders' fill rate increases when λ falls from 0.5 to 0.4 for $f^T = 0$ in 1.

Now consider the effect of an increase in the matchmaker's fee f^T , for a fixed value of λ . This increase reduces the surplus to be split between the maker and the taker when a transaction takes place on the limit order market ($G^l = 2L - f^T$). Thus, other things equal, an increase in f^T raises the attractiveness of the dealer market for a newcomer and reduces makers' market power, very much as an increase in λ does. As a consequence, the effect of an increase in f^T on makers' strategy is similar to that of a decrease in λ . In particular, it can induce makers to post more attractive offers, which therefore execute more frequently. This explains why, for instance, for $\lambda = 0.5$, an increase in f^T from 0 to 0.5 results in an increase in the trading rate on the limit order market.

Remark: We now briefly discuss what happens when condition C.1 is not satisfied. First consider the case in which $\delta_H < \frac{2\pi}{1-\pi}(1 - \delta_L) + \delta_L$. In this case either the proportion

of patient traders is high (greater than $1/3$) or/and $(\delta_H - \delta_L)$ is small. Thus, the heterogeneity among traders is small. For this reason, traders never submit limit orders with a low fill rate and only equilibria with a high fill rate arise (or the limit order market is inactive). Expressions for the quotes in the these equilibria are as in Proposition 1.¹⁶

When $\delta_H > \frac{2\pi}{1-\pi} + \delta_L$, the heterogeneity between patient and impatient traders is high and/or the proportion of patient traders is small. Thus, traders have no "natural" incentives to bid aggressively since there are very few patient traders and the difference between patient and impatient traders' reservation prices is high. For this reason, the unspecialized/high fill rate equilibrium does not exist. Otherwise, the possible equilibria (and the existence conditions) are as in Proposition 1.¹⁷

4 Implications

4.1 Bid-ask spreads and trading fees

Stoll (2000) reviews various measures of illiquidity in financial markets. He refers to the *traded spread* as the difference between the average price of trades at the ask side and the average price of trades at the bid side. In our model, this difference is $A^* - B^*$. Thus, we refer to $S_p = A^* - B^*$ as the *traded bid-ask spread on the limit order market*. The traded bid-ask spread underestimates the actual trading cost since it does not account for the taker fee. Thus, we also define the *cum fee bid-ask spread*, that is: $S_p^c \stackrel{def}{=} A^* - B^* + 2f_m$, which is the difference between the ask price cum fee and the bid price net of fee.

Figure 3 shows the cum fee bid-ask spread in equilibrium (plain line) and the bid-ask spread in the dealer market (dashed line) as a function of λ , the order processing cost in the dealer market.

[Insert Fig. 3 here]

¹⁶More specifically, an unspecialized/high fill rate equilibrium is obtained when $\frac{G^d}{G^l} \leq \kappa'_1$, a specialized/high fill rate equilibrium is obtained when $\kappa'_1 < \frac{G^d}{G^l} \leq \kappa'_2$ and the dealer market crowds out the limit order market when $\frac{G^d}{G^l} > \kappa'_2$ with $\kappa'_1 = \frac{\delta_L}{2+\delta_H-\delta_L}$ and $\kappa'_2 = \kappa_4$.

¹⁷Observe that for every $\delta_L > 0$, the set of parameters (π, δ_H) for which C.1 is satisfied is non empty. For instance for $\delta_L \approx 0$, it is sufficient to choose $\pi \in [\frac{\delta_H}{3}, \frac{\delta_H}{3(1-\delta_L)+\delta_H}]$, for C.1 to be satisfied. For $\delta_L = 0$, C.1 is never satisfied. This is the reason why we impose $\delta_L > 0$. When $\delta_L = 0$, impatient investors never submit limit orders since they obtain a zero payoff with a limit order. Thus, the equilibria are necessarily specialized and can be of types 3 or 4.

When the limit order market is active, the cum fee bid-ask spread is always lower in equilibrium than the bid-ask spread on the dealer market.¹⁸ Indeed, otherwise, it would never be optimal to submit a market order on the limit order market. In general, the cum fee bid-ask spread is strictly smaller than the bid-ask spread in the dealer market, except in the specialized/low fill rate equilibrium where the costs of a market order in either market are equal. It is also easily checked that the cum fee bid-ask spread is always positive. In contrast, the total fee being fixed, when f_m is positive and large (i.e., f_l is negative), the traded bid-ask spread can be negative. Yet, buying the security at the ask price and reselling it at the bid price would not be profitable because, cum fee, the bid-ask spread is positive.

As explained previously, makers have less market power when the bid-ask spread on the dealer market declines. Thus, *in a given equilibrium*, makers must therefore post more attractive bid-ask spreads when λ goes down, as shown on Figure 3. We therefore obtain the following result.

Corollary 1 :*For fixed fees of the trading platform, in a given equilibrium, the traded bid-ask spread and the bid-ask spread cum fees increase with the order processing cost on the dealer market.*

Corollary 1 is valid only for changes in λ that do not change the type of equilibrium, i.e., *small changes* in λ . For large changes, the effect of λ on the bid-ask spread in the limit order market is ambiguous. As an example, consider Table 1 again and suppose that $f^T = 0$. If $\lambda = 0.7$, a type #1 equilibrium is obtained and the cum fee bid-ask spread is $S_p^c = 0.68$. In this equilibrium, quotes have a high execution probability as they attract market orders from impatient and patient investors. But if $\lambda = 0.6$, a type #2 equilibrium is obtained: makers choose quotes that attract only impatient investors. As these investors have a high willingness to pay for immediacy, makers can afford to post much less aggressive quotes than in a type #1 equilibrium, at the cost of a lower execution probability. As a result, the cum fee bid-ask spread is $S_p^c = 1.13$ and is greater than when $\lambda = 0.6$, despite the fact that the dealer market appears more competitive in this case.

¹⁸The traded spread is $S_p = S_p^c - 2f_m$. Thus, it is also smaller than the bid-ask spread in the dealer market when $f_m > 0$. In contrast, when $f_m < 0$, the traded bid-ask spread can exceed the bid-ask spread in the dealer market.

Corollary 2 : *Suppose that the parameters are such that the equilibrium is unspecialized or specialized with a high fill rate:*

1. *The traded bid-ask spread in the limit order market decreases in the take fee and increases in the make fee.*
2. *The cum fee bid-ask spread increases in the total fee charged by the platform.*
3. *The total fee being fixed, the cum fee bid-ask spread does not depend on the allocation of the fee between makers and takers (i.e., it does not depend on f_l and f_m).*

To understand the first part of the corollary, consider first an increase in the take fee, f_m . Other things equal, this increase reduces one-for-one the concession that investors are willing to pay to trade upon arrival with a market order. That is, buyers' cut-off prices decline and sellers' cut-off prices increase, each by an amount equal to the take fee (see equations (5) and (6)). As a consequence, investors submitting limit orders must post more attractive quotes and the traded bid-ask spread narrows. This reduction in bid-ask spreads implies that the expected payoff with a limit order drops, which makes investors more willing to pay a concession for immediate execution. This feedback partially, but not fully, counterweights the initial change in investors' cut-off prices and the bid-ask spread. Thus, the net effect of an increase in the take fee is an increase in the bid price and a decrease in the ask price posted by investors submitting limit orders. But the decrease is less than one-for-one, that is:

$$-1 < \frac{1}{2} \frac{\partial S_p}{\partial f_m} < 0 \quad (14)$$

Hence, in equilibrium, the increase in the take fee is not entirely neutralized by a decrease in the traded bid-ask spread and therefore the cum fee half bid-ask spread increases in the take fee but at a rate less than one. Indeed, since $S^c = S + 2f_m$, we deduce from equation (14) that:

$$\frac{1}{2} \frac{\partial S_p^c}{\partial f_m} = \frac{1}{2} \frac{\partial S_p}{\partial f_m} + 1 > 0 \text{ and } \frac{1}{2} \frac{\partial S_p^c}{\partial f_m} < 1. \quad (15)$$

Hence, ultimately, the burden of a higher take fee is *shared* between makers and takers.

Now consider the effect of an increase in the make fee. Other things equal, an increase in the make fee reduces the expected payoff of investors submitting limit orders. As a

consequence, all investors are ready to pay larger concessions to get immediate execution. That is, other things equal, buyers' cut-off price increases and sellers' cut-off price decreases when the make fee increases (see equations (5) and (6)). This effect enables investors submitting limit orders to charge less competitive quotes, unless their quotes are constrained by those posted in the dealer market. But, as seen in Figure 2, this constraint does not bind for the equilibria considered in Corollary 2. Thus, in these equilibria, the traded bid-ask spread widens when the make fee increases because investors are willing to pay greater concessions to avoid the make fee. As a result, the expected payoff with a limit order is higher, which partially counterweight the impact of the increase in the make fee on investors' cut-off prices. Thus, the half traded bid-ask spread increases in the make fee but at a rate less than one:

$$0 < \frac{1}{2} \frac{\partial S_p}{\partial f_l} < 1. \quad (16)$$

This is also the case for the bid-ask spread cum fee since $\frac{1}{2} \frac{\partial S_p^c}{\partial f_l} = \frac{1}{2} \frac{\partial S_p}{\partial f_l}$. Hence, the increase in the make fee is not entirely "passed-through" by investors submitting limit orders to investors submitting market orders. Rather, as the increase in the take fee, the increase in the make fee is ultimately shared between both types of investors.

The last part of the corollary shows that changing the make/take fee breakdown, while keeping the total trading fee on the limit order market constant, does not affect the cum fee bid-ask spread. For instance, a decrease in the make fee by one cent triggers a drop of less than one cent in the half bid-ask spread (Part 1 of Corollary 2). If it is neutralized by an increase in one cent in the take fee, the half bid-ask spread drops further by less than one cent (Part 1 of Corollary 2 again) and the cumulative drop in the half bid-ask spread is just equal to one cent in equilibrium. That is the relative cost advantage granted to makers at the expense of takers is completely neutralized by the drop in the half traded bid-ask spread in equilibrium.

The specialized/low fill rate equilibrium requires a separate analysis. Indeed, in this equilibrium, the constraint that the quotes cum fee in the limit order market must be as attractive as dealers' quotes is binding (see Figure 3). Now, consider first an increase in the make fee. As explained previously, this increase reduces the expected payoff with a limit order and makes investors more willing to pay large concessions for immediate execution. But investors submitting limit orders cannot take advantage of this greater willingness to

pay for immediacy as their quotes cum fees would then become uncompetitive relative to an immediate trade in the dealer market. Hence, makers cannot pass through the increase in the make fee to takers, even partially, as in the other equilibria. Thus, the traded and cum fee bid-ask spread is inelastic to a change in the make fee in the specialized/low fill rate equilibrium.

Now consider an increase in the take fee. Following this increase, investors submitting buy (resp. sell) limit orders must increase (resp. reduce) their bid (ask) price by an amount just equal to the increase in the take fee as otherwise the payoff of a sell (buy) market order is less than the payoff obtained with a trade in the dealer market. As a consequence, the half traded bid-ask spread falls one-for-one with an increase in the take fee and the cum fee bid-ask spread is independent of the take fee. Thus, in a specialized/low fill rate equilibrium, an increase in the take or the make fee is entirely borne out by makers.

Overall, this analysis generates several interesting testable implications regarding the effect of a change in make/take fees on bid-ask spreads: (i) an increase in the make fee has a positive effect on the traded bid-ask spread, (ii) an increase in the take fee has a negative effect on the traded bid-ask spread, (iii) for a fixed fee structure, the cum fee bid-ask spread is independent of the make/take fee breakdown and (iv) an increase in the total fee increases the cum fee bid-ask spread or has no effect. Moreover, the model implies that the effect of an increase in the total fee on the traded bid-ask spread is ambiguous, as it depends on whether this increase is achieved by raising the take fee or the make fee. An increase in the take fee should decrease the traded bid-ask spread while an increase in the make fee should increase the bid-ask spread.

4.2 Trading Rate, Make Rate and Fill Rate

In addition to the bid-ask spread, market participants often use the trading rate of a market and the fill rate for limit orders (the fraction of executed limit orders) as other measures of market liquidity. We now use Proposition 1 to study the determinants of the fill rate and the trading rate on the limit order market.¹⁹

When the limit order market is active, the investor who arrives at a given date can

¹⁹The trading rate on the dealer market is negatively related to the trading rate on the limit order market. Thus, a change in the parameter that increases the trading rate on the limit order market has the opposite effect on the trading rate in the dealer market.

be: (1) a patient investor who submits a limit order; (2) a patient investor who submits a market order; (3) an impatient investor who submits a limit order; (4) an impatient investor who submits a market order; (5) an impatient investor who trades upon arrival in the dealer market. Let $\varphi^k = (\varphi_1^k, \varphi_2^k, \varphi_3^k, \varphi_4^k, \varphi_5^k)$ be the vector giving the stationary probability of each of these events at date t in an equilibrium of type $\#k$ conditional on the market being still opened at date t . The likelihood of a trade on the limit order market in a given period is:

$$TR^k = \varphi_2^k + \varphi_4^k. \quad (17)$$

This probability also measures the average number of trades per period on the limit order market since it gives the fraction of periods in which a trade takes place on the limit order market. Thus, we call it the *trading rate* on the limit order market.

We denote by MR^k , the “make rate”, i.e., the likelihood that an investor arriving in a given period submits a limit order. By definition:

$$MR^k = \varphi_1^k + \varphi_3^k. \quad (18)$$

In equilibrium, in the limit order market, the number of filled limit orders must be equal to the number of executed market orders. Hence, the trading rate can also be written:

$$TR^k = FR^k \times MR^k. \quad (19)$$

where FR^k is the “fill-rate” (i.e., the fraction of executed limit orders) in an equilibrium of type k . Thus, the trading rate is high when the matchmaker attracts many limit orders (a high make rate) and these limit orders have a high execution probability (a high fill rate). Last, the likelihood of a trade on either market in a given period is $\varphi_2^k + \varphi_4^k + (\varphi_1^k + \varphi_3^k)(1 - FR^k) + \varphi_5^k = 1 - TR^k$.²⁰ Thus, the market share of the limit order market is $MS_l^k = \frac{TR^k}{1 - TR^k}$. As expected, it increases with the trading rate on this market.

Corollary 3 *The trading rate and the make rate in the limit order market are:*

$$\begin{aligned} TR^1 &= 33\%; \quad TR^2 = \frac{1 - \pi}{3 - \pi}; \quad TR^3 = \frac{\pi(1 - \pi)}{2}; \quad TR^4 = \frac{\pi}{2 + \pi}. \\ MR^1 &= 66\%; \quad MR^2 = \frac{2}{3 - \pi}; \quad MR^3 = \pi; \quad MR^4 = \frac{2\pi}{2 + \pi}. \end{aligned}$$

²⁰The likelihood of a trade in a given period is smaller than one. To see this suppose that there is a trade on the limit order market at, say, date $t - 1$. Then the limit order book is empty at date t and if a patient investor arrives he will submit a limit order. In this case, there is no trade in either market at date t .

Moreover, under C.1, for a given π , $TR^1 > TR^2 > TR^4 > TR^3$ and $MR^2 > MR^1 > MR^3 > MR^4$. The fill rate is 50% in equilibria of types #1 and #4 and $\frac{(1-\pi)}{2}$ in equilibria of types #2 and #3.

For a given value of π , the trading rate is entirely determined by the type of equilibrium which is obtained. Thus, the make/take fee breakdown has no effect on the trading rate since it does not affect the type of equilibrium in the limit order market.

In contrast, the trading fee has an impact on the trading rate since it is one determinant of the type of equilibrium. Observe that the trading rate is higher in a type #4 equilibrium than in a type #3 equilibrium. An intriguing implication is that an increase in the trading fee on the limit order market (or a decrease in the order processing cost on the dealer market) can, counter-intuitively, raise the trading rate on the limit order market (and therefore its market share). To see this, suppose that the trading fee and the order processing cost in the dealer market are such that an equilibrium of type #3 is obtained and consider Figure 2. In this case, as shown on Figure 2, there always exists a higher fee or a smaller order processing cost such that a type #4 equilibrium obtains.

For a numerical example, consider again Table 1 and suppose $\lambda = 0.6$. When $f^T = 0$, a type #2 equilibrium is obtained. As $\pi = 0.2$, the make rate is 71% and the fill rate is 40% (See Corollary 3). Thus, the trading rate is $71\% \times 40\% = 28\%$. When $f^T = 0.5$, a type #3 equilibrium is obtained. The make rate falls to 20% as impatient investors stop using limit orders and the fill rate remains at 40%. As a result the trading rate is only $20\% \times 40\% = 8\%$. But if $f^T = 0.9$, a type #4 equilibrium is obtained. The make rate is smaller (18.1%). But makers choose limit orders with a higher fill rate (50%). Hence, eventually the trading rate is higher and equal to $18.1\% * 50\% = 9\%$.

The reason for this counter-intuitive finding is as follows. Suppose that f^T and λ are such that a type #3 equilibrium is obtained. In this equilibrium, only patient investors submit limit orders. Thus, the make rate is low. Moreover, only impatient investors find optimal to submit market orders. Thus, the fill rate is low. As a consequence, the trading rate is very low. Now suppose that the matchmaker raises its fee or suppose that dealers' order processing cost declines. As explained after Proposition 1, either change reduces makers' market power and, if large enough, can force them to post quotes with a high fill rate. For this reason, the trading rate is strictly higher in a type #4 equilibrium than in

a type #3 equilibrium.

As our numerical example shows, the model does not imply that the trading rate on the limit order market always increases in the trading fee. It just features a range of values for the order processing cost in the dealer market (those resulting in a type #3 equilibrium) for which this can be the case.

4.3 Trading Fee and Investors' Welfare

We now use the findings of the previous sections to analyze the effect of trading fees on investors' welfare. We measure investors' welfare *ex-ante*, i.e., before investors learn their type (buyer/seller and patient/impatient) and their role (maker/taker). Our main result in this section is that a strictly positive trading fee can enhance investors' welfare in some cases.

Let $W(\lambda, f^T)$ be the expected ex-ante gains from trade for an investor when the order processing cost in the dealer market is λ and the trading fee charged by the matchmaker is f^T . Moreover, let $TR(f^T, \lambda)$ and $FR(f^T, \lambda)$ be respectively the trading rate on the platform and the fill rate for limit orders. Computations show that²¹

$$W(f^T, \lambda) = TR(f^T, \lambda)G^l + (1 - 2TR(f^T, \lambda))G^d - \underbrace{\bar{\delta}(f^T, \lambda) \left((G^l + S^c - f^T)FR(f^T, \lambda) + 2(1 - FR(f^T, \lambda))G^d \right)}_{\text{Waiting Costs}} \quad (20)$$

where $\bar{\delta}(f^T, \lambda) = ((1 - \delta_H)\varphi_1(\lambda, f^T) + (1 - \delta_L)\varphi_3(\lambda, f^T)) / 2$ and $\varphi_1(\lambda, f^T)$ (resp., $\varphi_3(\lambda, f^T)$) is the probability that a patient (resp. impatient) acts as a maker. This probability depends on the equilibrium type, as explained in Section 4.2. For instance, if λ and f^T are such that a type #3 equilibrium is obtained then $\varphi_1(\lambda, f^T) = \varphi_{1_j}^3$, where φ_1^3 is defined in Section 4.2.

The first component in investors' welfare is the weighted average of investors' surplus when a transaction takes place on the limit order market, G^l , and investors' surplus when a transaction takes place on the dealer market, G^d . The weights are the probabilities that, in each period, a transaction takes place on the limit order market ($TR(f^T, \lambda)$) or the dealer market ($1 - 2TR(f^T, \lambda)$). Investors' welfare is lower than this weighted average because makers incur a "waiting cost" since $\delta_j > 0$. The expected welfare loss due to this waiting

²¹For brevity, we derive the expression for investors' ex-ante expected gains from trade in the Internet Appendix for this paper.

cost is given by the last component in the expression for investors' welfare. Investors' welfare is independent of the make/take fee breakdown since this breakdown does not affect neither the division of gains from trade among makers and takers in equilibrium, nor investors' actions in equilibrium (hence the trading rate, the fill rate and the make rate).

Proposition 1 describes the equilibrium when investors optimally trade in the dealer market or in a limit order market. To measure the efficiency of this market structure, we need a benchmark. To this end, we first study the "role allocation" that maximizes investors' welfare given that investors arrive sequentially and have a deadline of one period. We refer to this allocation as the "unconstrained first best."²² A role allocation specifies the action taken by each type of investor when he arrives in the market (trade in the dealer market/act as a taker/act as a maker) and the division of surplus between the maker and the taker in each transaction. For instance, suppose that an impatient investor arrives after a trade just took place. A role allocation specifies whether the investor immediately trades in the dealer market or whether she acts as a maker ("waits"). In the second case, the role allocation also specifies whether she is matched with the next investor (or eventually trades in the dealer market) and, if she is matched, how the surplus is divided between the two investors.

Each equilibrium in Proposition 1 results in a specific role allocation but, as shown below, this role allocation does not in general maximizes investors' welfare. This is not surprising. In any feasible mechanism, incentives compatibility constraints reduce the gains from trade relative to the allocation of role maximizing gains from trade for investors.

A central planner in charge of choosing the role allocation that maximizes investors' welfare optimally sets the trading fee to zero. Indeed, this fee reduces the total surplus accruing to investors when a trade takes place and has no influence on investors' role since this role is chosen by the central planner, not investors. Thus, there is no loss in generality in setting the trading fee to zero to analyze the role allocation maximizing investors' welfare. Using this observation, we obtain the following result.²³

²²We say unconstrained since we derive this role allocation assuming that the planner in charge of implementing this allocation knows the types of investors. Another benchmark would be the "constrained first best", which is obtained by implementing the optimal direct mechanism that induce investors to reveal their type. Clearly, the trading mechanisms considered in the paper cannot achieve a higher welfare for investors than either benchmarks.

²³This result holds for all values of the parameters, even those such that Condition C.1 does not hold.

Proposition 2 (*first best*) Let $\lambda^* = \frac{L(1-\delta_H)}{3-\delta_H}$ and $\lambda^{**} = \frac{L(2(1-\delta_L)+\pi(\delta_H-\delta_L))}{3(2-\delta_L)-\pi(\delta_H-\delta_L)}$. For all values of the parameters, a central planner that maximizes investors' welfare sets the trading fee to zero, leaves no surplus to makers and chooses a role allocation that is

1. as in a type #5 equilibrium if $\lambda < \lambda^*$ (i.e., all trades take place on the dealer market).
2. as in a specialized/high fill rate equilibrium (type #4) if $\lambda^* \leq \lambda < \lambda^{**}$,
3. as in an unspecialized/low fill rate equilibrium if $\lambda^{**} \leq \lambda \leq L$.

To gain intuition on this result, suppose an investor arrives and there is no possible match with the previous investor. The trade-off for the central planner is whether the newcomer should act as maker or whether she should trade immediately on the dealer market. The first choice entails a waiting cost while the second choice entails an order processing cost. The solution to this trade-off depends on the size of the order processing cost, λ . If λ is small then the cost of an immediate trade in the dealer market is relatively small and it is never efficient to have an investor waiting (acting as makers). Thus, all trades take place in the dealer market. For intermediate values of λ , the cost of a trade in the dealer market is higher and it is efficient that the investor acts as a maker iff her waiting cost is relatively small (as in a specialized equilibrium). Last, if λ is high, it is efficient to have the investor acting as a maker whether she is patient or impatient (as in an unspecialized equilibrium).

A second feature of the first best role allocation is that if an investor acts as a maker then she is matched with the next investor whenever possible (e.g., the first investor is a buyer and the second a seller). Thus, as in high fill rate equilibria, the likelihood of a match is maximal ($\frac{1}{2}$). In this way, the central planner maximizes the return on the cost of waiting by minimizing the chance that an investor will eventually have to be matched with a dealer. Last, leaving surplus to makers is inefficient since the latter discount this surplus. Thus, the central planner leaves no surplus to makers.

Now, consider the role allocation that is obtained when investors optimally choose their order placement strategies. Proposition 1 implies that this role allocation never corresponds to the first best, even when the trading fee is zero. First, when the limit order market is active, makers obtain a strictly positive expected surplus. This is required as makers need to be incentivized to submit limit orders. Yet, this is inefficient since

investors discount delayed gains from trade. Second, and more interestingly, the dealer market crowds out the limit order market too “quickly.” Indeed, $\lambda^* < \lambda_4$. Thus, for $\lambda \in [\lambda^*, \lambda_4]$, the limit order market is inactive in equilibrium while the first best role allocation requires trades to happen directly between investors. Third, there is a range of values for λ ($\lambda \in (\lambda_3, \lambda_1)$) for which the likelihood of a match between two consecutive investors is less than $\frac{1}{2}$ in equilibrium (type #2 and type #3 equilibria). For these values of λ , makers choose quotes with low execution probabilities to extract more surplus from takers in case of execution. This behavior is individually optimal but socially inefficient since low likelihood of execution raises the chance that the cost of waiting for makers will be paid needlessly.

Thus, when $\lambda \in (\lambda_3, \lambda_1)$, unfilled limit orders are a source of inefficiency of the market structure considered in our paper. Interestingly, Hollifield et al.(2006) empirically show that unfilled limit orders constitute a major source of inefficiency for limit order markets. We now show that the trading fee can be used to alleviate this inefficiency and raise investors’ welfare. Intuitively, the reason is that a sufficiently large increase in trading fees can induce makers to switch from a strategy with a low fill rate to a strategy with a high fill rate, as explained in previous sections.

To see this, suppose that $\lambda \in (\lambda_3, \lambda_2]$. In this case, when $f^T \in [0, f_3^T(\lambda)]$, a type #3 equilibrium is obtained while for $f^T \in (f_3^T(\lambda), f_4^T(\lambda)]$, a type #4 equilibrium is obtained (see Figure 2). Thus, an increase in trading fee from $f^T = 0$ to, say, $f^T = f_3^T(\lambda) + \epsilon \leq f_4^T(\lambda)$ will shift the fill rate from low to high. The net effect on investors’ payoff is ambiguous. On the one hand, the higher fill rates for limit orders has a positive effect on welfare. On the other hand, the higher trading fee has a negative effect on investors’ welfare. However, for ϵ sufficiently small (i.e., f^T close to $f_3^T(\lambda)$), the first effect dominates as claimed in the next corollary.

Corollary 4 *Suppose that $\lambda \in (\lambda_3, \lambda_2]$. Then, there exists a value $\hat{\lambda} \in (\lambda_3, \lambda_2]$ such that for $\lambda_3 < \lambda \leq \hat{\lambda}$, investors’ welfare is higher when $f^T = f_3^T(\lambda) + \epsilon$ (where ϵ is very small but positive) than when $f^T = 0$.*

The same type of result can be obtained for $\lambda \in (\lambda_2, \lambda_1]$. Indeed, in this case, if $f^T = 0$, a type #2 equilibrium is obtained whereas if $f^T = f_3^T(\lambda) + \epsilon$ a type #4 equilibrium is

Order Processing Cost: λ	Investor's Aggregate Welfare		
	First Best	Equilibrium	
		Zero Trading Fee	Trading Fee: $f_3^T(\lambda) + \epsilon$
λ_3	0.687	49%	56%
0.82	0.685	47%	51%
0.85	0.682	44%	43%
0.88	0.679	41%	34%
0.90	0.677	39%	29%

Table 2: Trading fee and Investors' welfare

obtained. As the fill rate for limit orders is higher in a type #4 equilibrium, investors' welfare is higher when $f^T = f_3^T(\lambda) + \epsilon$ than when $f^T = 0$ if λ is sufficiently close to λ_2 .

We illustrate these findings with a numerical example in Table 2. In Table 2, parameter values are set at $L = 1$, $\pi = 0.297$, $\delta_H = 0.885$, $\delta_L = 0.067$. Given these values, we have $\lambda_3 = 0.802$ and $\lambda_2 = 0.95$. We therefore consider different values of λ in the interval $[0.802, 0.95]$. For each value of λ , we give investors' welfare in the first best in the second column. In the third and fourth column, we give investors' welfare as a percentage of their welfare in the first best when the trading fee is zero (third column) and when the trading fee is set at $f_3^T(\lambda) + \epsilon$ (we set $\epsilon = 10^{-9}$). In this example, it turns out that $\hat{\lambda} \approx 0.84$. Thus, for all values of $\lambda \in (0.802, 0.84)$, investors' welfare is strictly higher when the fee is set at $f_3^T(\lambda) + \epsilon$. It is worth stressing that the fee required to maximize investors' welfare can be large compared to gains from trade. For instance, in the example considered in Table 2, the fee must be set at about $f^T = 0.18$ (i.e., 26.2% of total gains from trade) when $\lambda = 0.82$.

Corollary 4 shows that a positive trading fee can be part of the optimal market structure even if the matchmaker incurs no cost to match trades. This is not the case for all values of λ , however, as shown by the next corollary.

Corollary 5 *Suppose that $\lambda \in (\lambda_4, \lambda_3]$ or $\lambda \in (\lambda_1, L]$. Then, the trading fee that maximizes investors' welfare is $f^T = 0$.*

If $\lambda \in (\lambda_4, \lambda_3]$ or $\lambda \in (\lambda_1, L]$ and the trading fee is zero, an equilibrium of type #4 or #1 is obtained. In this equilibrium, the fill rate is as high as in the first best allocation and matches are as in this allocation for a given sequence of arrivals. In this case, raising

the trading fee can only reduce investors' welfare since (i) it reduces the surplus to be split among makers and takers and (ii) it leaves unchanged or even decreases the fill rate. Thus, for extreme values of λ , the trading fee that maximizes investors' welfare is zero.

To sum up, for high or low values of the order processing cost in the dealer market, the trading fee that maximizes investors' welfare is zero. But, surprisingly, this fee can be strictly positive for intermediate values of the order processing cost as a higher trading fee is a way to induce makers to make offers with higher execution probabilities.

5 Market Structure and Investors' Welfare

In this section, we study which type of "market structure" maximizes investors' welfare. We consider several types of market structures: a single dealer market, a dealer market coexisting with one or two matchmakers or a market structure with only one or two matchmakers. We first derive the optimal pricing policy for the matchmaker(s) in each market structure and we then study which market structure results in the highest possible welfare for investors.(see Section 5.3).

5.1 Pricing policy of a single matchmaker

The per period expected profit of the matchmaker is equal to the trading rate on the limit order market times the total fee per trade on this market. As the trading rate does not depend on the breakdown of the total fee, the objective function of the matchmaker is:

$$\text{Max}_{f^T} \Pi(f^T, \lambda) \equiv TR_l(f^T, \lambda) \times f^T,$$

where $TR(f^T, \lambda)$ is the trading rate on the platform if its fee is f^T and the order processing cost on the dealer market is λ . As explained previously, if the platform sets a fee $f^T \in (f_{k-1}^T(\lambda), f_k^T(\lambda)]$ then a type k equilibrium is obtained and $TR_l(f^T, \lambda) = TR_l^k$ (see Proposition 1 and Corollary 3). Thus, in a type k equilibrium, the platform can increase its fee up to $f_k^T(\lambda)$ without changing its revenue per period. Setting a fee strictly larger than $f_4^T(\lambda)$ is never optimal for the matchmaker since it results in no trading on the limit order market. Also, setting its fee at $f_3^T(\lambda)$ cannot be optimal since by raising its fee at $f_4^T(\lambda)$, it attracts more trades and thereby it generates more revenue, as shown in Corollary 3. Thus, eventually, the matchmaker optimally chooses one of three fees: $f_1^T(\lambda)$,

$f_2^T(\lambda)$, or $f_4^T(\lambda)$. In making this choice, the platform faces the traditional price-quantity trade-off for a monopolist: the larger is the fee charged by the matchmaker, the smaller is the trading rate on the limit order market. The solution to this trade-off ultimately depends on the order processing cost in the dealer market, as shown in the next proposition. For this proposition, we use the following notations: $\lambda'_1 \equiv \left(\frac{(3-\pi)\kappa_1^{-1}-3(1-\pi)\kappa_2^{-1}-4\pi}{(3-\pi)\kappa_1^{-1}-3(1-\pi)\kappa_2^{-1}} \right) L$, $\lambda'_2 \equiv \left(\frac{(2+\pi)\kappa_1^{-1}-3\pi\kappa_4^{-1}-4(1-\pi)}{(2+\pi)\kappa_1^{-1}-3\pi\kappa_4^{-1}} \right) L$, $\lambda'_3 \equiv \left(\frac{(1-\pi)(2+\pi)\kappa_2^{-1}-\pi(3-\pi)\kappa_4^{-1}-4(1-2\pi)}{(1-\pi)(2+\pi)\kappa_2^{-1}-\pi(3-\pi)\kappa_4^{-1}} \right) L$. Under C.1, we have either $\lambda'_1 > \lambda'_2 > \lambda'_3 > \lambda_4$ or $\lambda'_3 > \lambda'_2 > \lambda'_1 > \lambda_4$.

Proposition 3 1. *If $\lambda < \lambda_4$, the limit order market is inactive (there is no positive fee for which the matchmaker can attract limit orders).*

2. *If $\lambda \geq \lambda_4$, the matchmaker's optimal fee is:*

$$f^{T*}(\lambda) = \begin{cases} \frac{\lambda-\lambda_1}{\kappa_1} & \text{if } \max(\lambda'_1, \lambda'_2) \leq \lambda \leq L, \\ \frac{\lambda-\lambda_2}{\kappa_2} & \text{if } \lambda'_3 \leq \lambda < \lambda'_1, \\ \frac{\lambda-\lambda_4}{\kappa_4} & \text{if } \lambda_4 \leq \lambda < \min(\lambda'_2, \lambda'_3). \end{cases}$$

Thus, the type of equilibrium obtained in the limit order market given the optimal fee set by the matchmaker is: a type #1 equilibrium if $\max(\lambda'_1, \lambda'_2) \leq \lambda \leq L$, and a type #4 equilibrium if $\lambda_4 \leq \lambda < \min(\lambda'_2, \lambda'_3)$, or a type #2 equilibrium if $\lambda'_3 \leq \lambda < \lambda'_1$ and $\lambda'_3 < \lambda'_1$.

Figure 4 shows the evolution of the fee set by the matchmaker as a function of order processing cost in the dealer market, λ when $\lambda'_1 > \lambda'_2 > \lambda'_3 > \lambda_4$.

Insert Fig.4 about here

Corollary 6 :*The type of equilibrium being fixed, the fee charged by the matchmaker increases in the order processing cost in the dealer market.*

Intuitively, as the order processing cost in the dealer market declines, the matchmaker faces increasing competition from the dealer market since the surplus that investors can obtain by immediately trading in the dealer market gets larger. For this reason, the matchmaker tends to choose smaller fees when the bid-ask spread in the dealer market becomes small.

Corollary 6 is valid only when the change in the bid-ask spread in the dealer market does not result in a change in the type of equilibrium obtained in the limit order market.

For instance, consider λ_H and λ_L such that $\lambda_H = \lambda'_2 + \epsilon$ and $\lambda_L = \lambda'_2 - \epsilon$. As shown on Figure 5, for ϵ small enough, the fee charged by the matchmaker will be higher when $\lambda = \lambda_L$ than when $\lambda = \lambda_H$ and yet $\lambda_H > \lambda_L$. Thus, an increase in the competitiveness of the dealer market does not always induce the matchmaker to reduce its fee. The reason is the following. In a given equilibrium, the matchmaker decreases its fee when λ decreases to maintain unchanged the trading rate. At some point, this is too costly and the matchmaker is better off raising discontinuously its fee at the cost of a loss in market share (a drop in the trading rate).

Entry of the dealer market curbs the matchmaker's market power and always results in a decrease in the trading fee. To see this, consider the polar case in which $\lambda = L$. As explained previously, this situation is akin to the case in which the dealer market does not exist and the matchmaker has therefore full monopoly power. As one would expect, in this case, the matchmaker optimally charges the largest possible fee: $f^{T^*}(L) = 2L$ (see Proposition 3). Thus, the matchmaker extracts all gains from trade ($G^l = 0$) and investors' payoff is zero, whether they use a market order or a limit order. In contrast, when $\lambda < L$, the fee charged by the matchmaker is strictly less than $2L$ and investors' expected payoffs are strictly positive.

Let $S^c(f^T, \lambda)$ be the cum fee bid-ask spread when the bid-ask spread in the dealer market is λ and the total fee charged by the matchmaker is f^T . Thus, when the matchmaker optimally sets its fee, the cum fee bid-ask spread is $S^c(f^{T^*}(\lambda), \lambda)$.

Corollary 7 *Suppose the matchmaker optimally sets its trading fee at $f^{T^*}(\lambda)$ and $\lambda > \lambda_4$ (the matchmaker is active).*

1. *In this case, the cum fee bid-ask spread in the dealer market is equal to the bid-ask spread in the dealer market ($S^c(f^{T^*}(\lambda), \lambda) = 2\lambda$) when $\lambda \leq \max(\lambda'_1, \lambda'_2)$ and smaller than the bid-ask spread in the dealer market when $\max(\lambda'_1, \lambda'_2) < \lambda$.*
2. *Moreover the cum fee bid-ask spread increases with the order processing cost in the dealer market: $\frac{dS^c_p(f^{T^*}(\lambda), \lambda)}{d\lambda} > 0$.*

The first part of the corollary helps to better understand the optimal pricing policy for the matchmaker. To see this, recall that the cum fee bid-ask spread increases with the

trading fee. Thus, the highest fee that the matchmaker can charge is the fee that makes all investors submitting market orders indifferent between trading in the dealer market (and paying λ) or the limit order market (and paying $S^c(f^T, \lambda)$). This fee is such that investors' surplus is $L - \lambda$ whether they trade in the dealer market or in the limit order market. We call this fee the “matching fee.” The matching fee is the highest possible fee for the matchmaker. But it has a cost: it precludes either the submission of market orders by patient investors or the submission of limit orders by impatient investors. To see this, suppose that the matchmaker picks the matching fee and patient investors submit market orders. In this case quotes must be such that patient investors just obtain $(L - \lambda)$ with a limit order. Indeed if they obtain more they would not submit market orders whereas if they obtain less they would never submit a limit order and the limit order market would be inactive. But in this case, impatient investors must receive a surplus strictly less than $(L - \lambda)$ with limit orders and therefore never act as makers. Alternatively, suppose that impatient investors submit limit orders. Then patient investors must obtain a surplus strictly higher than $(L - \lambda)$ and never submit market orders.

Thus, the matching fee precludes a type #1 equilibrium (in which both types of investors can be makers or takers) and hence a high trading rate. This equilibrium can happen only if the matchmaker's fee leaves a surplus strictly higher than $(L - \lambda)$ to investors submitting market orders. When $\lambda < \max(\lambda'_1, \lambda'_2)$, this surplus is too high and the matchmaker is better off with the matching fee whereas if $\lambda \geq \max(\lambda'_1, \lambda'_2)$, the matchmaker optimally chooses a strictly lower fee (so that $(S^c(f^{T*}(\lambda), \lambda) < 2\lambda)$).

The second part of the corollary shows that the cum fee bid-ask spread in the limit order market increases with the order processing cost in the dealer market even when the matchmaker's fee is endogenous (and hence react to the change in order processing cost). In this case, a reduction in dealers' order processing cost reduces the cum fee bid-ask spread on the limit order market because it reduces the market power of both the matchmaker (hence its fee) and the makers (who therefore post more aggressive quotes).

5.2 Competing matchmakers

We now extend the model to analyze the effect of competition between limit order markets. To this end, we assume that there are two matchmakers denoted 1 and 2. The fees and

quotes on the platform ran by matchmaker j are indexed by $j \in \{1, 2\}$. For instance, f_{mj} denotes the take fee on the platform ran by matchmaker j and A_j^* denotes the ask price posted by sellers on this platform in equilibrium. We refer to the set of offers/trades in both platforms as the "*consolidated market*."

Upon arrival, an investor observes the quote posted in each limit order market and decides whether to submit a market order, a limit order or to trade on the dealer market. Moreover, if the investor chooses a market order or a limit order, the investor also decides whether the order gets routed to matchmaker 1 or matchmaker 2. We provide a formal definition of the equilibrium in this case in Appendix B. For brevity, the proofs of the results in this section are given in the Internet Appendix.

Proposition 4 • *If the matchmakers charge different total fees ($f_1^T \neq f_2^T$), the limit order market with the largest total fee is inactive and the equilibrium is as described in Proposition 1 with a single limit order market.*

- *If the matchmakers charge the same total fees ($f_1^T = f_2^T$), the two limit order markets are active, the equilibrium is as in Proposition 1, but when an investor submits a limit order, he chooses to route his order to platform 1 with probability $\frac{1}{2}$ and to platform 2 with probability $\frac{1}{2}$. Both platforms are active if and only if $\lambda \geq \lambda_4$.²⁴*

Thus, our predictions regarding the bid-ask spread on the limit order market, the fill rates, the trading rates etc... are still valid for the consolidated market when two matchmakers, instead of one, compete for investors' order flow. In particular, for a given sequence of investors' arrivals, the dynamics of order flow will be identical whether there is a single matchmaker or two matchmakers (holding the total fee constant). The only difference with two coexisting matchmakers is that each only gets half of all trades (thus trading rates on individual platforms are divided by two). But, *for given fees*, from the point of view of investors, everything is as if trading was consolidated in a single market.

The market share of each matchmaker is determined by its total fee relative to its competitor's fee. However the breakdown of this fee between makers and takers is neutral. For instance, if matchmaker 2 charges a total fee that is strictly higher than matchmaker 1

²⁴In equilibrium, limit order traders use limit orders with the same execution probabilities on both platforms. For instance, if they submit a limit order with high fill rate on platform 1, they also do so when they submit limit orders on platform 2.

($f_{T1} < f_{T2}$) then it attracts no trading at all, even if it subsidizes one side ($f_{l2} < 0$ or $f_{m2} < 0$). Interestingly, it is often argued that by charging a low make fee, a trading platform can attract more limit orders and be therefore more attractive for investors submitting market orders. This logic is not as obvious as it seems since it does not hold here. The reason is as follows.

Suppose that $f_1^T = f_2^T$ and that initially both markets have the same make/take fee breakdown. Moreover, suppose that parameters are such that an equilibrium of types #1, #2 or #4 obtains. Now suppose that matchmaker 2 cuts its make fee and recovers the loss in revenues by increasing its take fee, so that its total fee is unchanged. Other things equal, the cut in the make fee increases the expected payoff for investors submitting limit orders on platform 2. But, for this reason and the fact that the take fee is higher on platform 2, investors are less willing to pay concessions for an immediate trade on this platform. Thus, as explained in Section 4.1 (see the discussion after Corollary 2), investors have to price their limit orders more aggressively on platform 2 and the traded bid-ask spread on platform 2 must fall until the point where the cum fee bid-ask spread is identical on both platforms. At this point, the division of gains from trade between makers and takers is identical in both markets (as cum fees quotes are identical) and investors submitting limit orders are therefore indifferent between routing their limit orders to platform 1 or platform 2.

As platforms can coexist with different make/take fees, the traded bid-ask spreads on both platforms can be very different. Interestingly, Corollary 2 implies that the platform with the smallest make fee (largest take fee) must feature a smaller traded bid-ask spread. But, in all cases, the cum fee bid-ask spread is identical on both platforms since their total fees are identical. These are two additional testable implications of the model.

Let $\Pi_j(f_j^T, f_{-j}^T; \lambda)$ be the expected profit of matchmaker j for a given choice of its fee (f_j^T), the fee chosen by its competitor (f_{-j}^T) and the order processing cost in the dealer market. Using Propositions 1 and 4, we deduce that:

$$\Pi_j(f_j^T, f_{-j}^T; \lambda) = \begin{cases} TR(f^T, \lambda) \times f^T & \text{if } f_j^T < f_{-j}^T, \\ 0.5 \times TR(f^T, \lambda) \times f^T & \text{if } f_j^T = f_{-j}^T, \\ 0 & \text{if } f_j^T > f_{-j}^T. \end{cases}$$

where $TR(f^T, \lambda) = TR^k$ if f^T and λ are such that $\kappa_{k-1} < \frac{L-\lambda}{2L-f^T} \leq \kappa_k$. The next proposition provides the Nash equilibrium of the stage in which the two matchmakers simulta-

neously choose their trading fees. We focus on the case $\lambda > \lambda_4$ as otherwise the dealer market crowds out the matchmakers.

Proposition 5 : *If $\lambda > \lambda_4$, both matchmakers optimally choose a zero total fee for any value of the bid-ask spread in the dealer market. The breakdown of this fee for each matchmaker is indeterminate (i.e., any menu (f_{mj}, f_{lj}) such that $f_{mj} + f_{lj} = 0$ can be sustained in equilibrium). The type of equilibrium in the consolidated limit order market is as given in Proposition 1 in the particular case in which $f^T = 0$.*

Thus, competition among matchmakers drives their total fee to zero. Hence, one expects the cum fee bid-ask spread in the limit order market to decline and the market share of the dealer market to fall after entry of a new matchmaker. Interestingly, the next corollary shows that this is not always the case.

Corollary 8 *Suppose $\lambda > \lambda_4$.*

1. *When $\lambda_3 \leq \lambda < \lambda_2$, the cum fee bid-ask spread is identical when there is one or two matchmakers and the market share of the dealer market is higher when there are two matchmakers.*
2. *When $\lambda_2 \leq \lambda < \lambda_1$, the cum fee bid-ask spread is smaller and the market share of the dealer market is smaller when there are two matchmakers.*
3. *When $\lambda < \lambda_3$ or $\lambda > \lambda_1$, the cum fee bid-ask spread is smaller and the market share of the dealer market is identical.*

Thus, when $\lambda \in [\lambda_3, \lambda_2)$, entry of a new matchmaker leaves the bid-ask spread unchanged and *raises* the market share of the dealer market. The reason for this counter-intuitive result is that a drop in trading fee can induce makers to choose limit orders with lower fill rates. More formally, when $\lambda \in [\lambda_3, \lambda_2)$ and a single matchmaker operates, the trading fee chosen by the matchmaker is such that a type #4 equilibrium obtains (as $\lambda_2 < \min(\lambda'_2, \lambda'_3)$, see Corollary 3). Moreover, the fee is such that the cum fee bid-ask spread on the dealer market is just equal to the bid-ask spread on the dealer market (Corollary 7). Entry of an additional matchmaker drives the trading fee to zero. Thus, it increases makers' market power (as explained in previous section) and induces them to

choose bidding strategies with a low fill rate so that a type #3 equilibrium obtains (see Figure 2 for $\lambda \in [\lambda_3, \lambda_2)$). As the trading rate on the consolidated limit order market is smaller in a type #3 equilibrium than in a type #4 (see Corollary 3), entry of a new matchmaker is eventually associated with a higher market share for the dealer market. Moreover it does not change the bid-ask spread since the cum fee bid-ask spread is equal to the bid-ask spread in the dealer market in a type #3 equilibrium.

The previous corollary offers a potential explanation for the evolution of the market share of the OTC market for E.U equities markets. The implementation of new rules (so called MiFID regulation) in 2007 has triggered entry of many new trading platforms trading stocks listed on E.U incumbent exchanges, forcing these exchanges to cut their fees. Yet, the market share of the OTC equities market for E.U stocks has been steadily increasing since 2007, which is one possible outcome predicted by Corollary 8.

5.3 Market structure and investors' welfare

We now use the the findings of the previous sections to analyze which market structure yields the highest welfare for investors. Specifically, we run an horserace between four different market structures: (i) competing matchmakers without a dealer market (“CM” for short), (ii) a monopolist matchmaker with a dealer market (“MMD”), (iii) competing matchmakers with a dealer market (“CMD”) and (iv) a dealer market only (D).

The first question is whether it is optimal to have one or two matchmakers. Each market structure has cost and benefits to investors. With two matchmakers, the trading fee is zero. However, for values of $\lambda \in [\lambda_3, \lambda_1]$, the equilibrium with two matchmakers is such that the fill rate for limit orders is low while a single matchmaker sets its fee such that for these values of λ , the fill rate is high. Thus, a priori, investors' welfare could be higher with a single matchmaker. However, as shown in the next proposition, this never happens because a single matchmaker fee is too high.

A second question is whether a dealer market should coexist with the matchmakers. Suppose that investors can only trade in a limit order market with two matchmakers. This situation is as if $L = \lambda$ and a type #1 equilibrium is obtained. Now suppose that a dealer market is introduced. The benefits for investors is that, upon arrival, they can contact a dealer if their waiting cost is high and they have access to “last resort liquidity suppliers”

when their limit orders are unfilled. The cost however is that a dealer market may induce makers to choose bidding strategies with a lower fill rate or to specialize. This happens if $\lambda \in [\lambda_3, \lambda_1]$. For this reason, for this range of values for λ , the optimal market organization can feature only two matchmakers as shown in Proposition 6 below.

Proposition 6 :

1. When $\lambda \leq \lambda_4$, investors' welfare is maximal when investors can only trade in a dealer market.
2. When $\lambda > \lambda_4$, depending on the parameters π, δ_H, δ_L ,
 - either investors' welfare is always maximal when investors can trade in a dealer market and two competing matchmakers,
 - or there exists $\bar{\lambda} \in]\lambda_4, \lambda_1[$ such that for $\lambda \in [\bar{\lambda}, \lambda_1[$ investors' welfare is maximal with two competing matchmakers but no access to a dealer market, and maximal with two competing matchmakers and access to a dealer market otherwise (i.e., for $\lambda \in]\lambda_4, \bar{\lambda}[\cup]\lambda_1, +\infty[$).

Table 3 illustrates Proposition 6 for the same parameter values as in Table 2 ($L = 1, \pi = 0.297, \delta_H = 0.885, \delta_L = 0.067$). For these parameter values, $\lambda_4 = 0.11, \lambda_3 = 0.80, \lambda_2 = 0.95, \lambda_1 = 0.98$ and $\bar{\lambda} = \lambda_3$. Thus, the optimal organization for investors features two competing matchmakers operating in parallel with a dealer market for $\lambda \in [\lambda_4, \lambda_3]$ or $\lambda > \lambda_1$, a single dealer market when $\lambda < \lambda_4$, and two competing matchmakers without a dealer market for $\lambda \in]\lambda_3, \lambda_1[$.

Consider the case in which two competing matchmakers operate in parallel with a dealer market now (CMD). As explained in Section 4.3, in this market structure, investors' welfare can be improved by charging a higher fee when $\lambda \in [\lambda_3, \lambda_2]$. For instance when $\lambda = 0.82$, investors' welfare with two competing matchmakers and a dealer market is 47.5% of investors' welfare in the first best. However, investors' welfare can be improved by 3.5% (see Table 2) if the trading fee is set at $f_3^T(0.82) = 0.18$. In this case, investors' welfare is 51% of investors' welfare in the first best.

Competition between matchmakers however drives the trading fee to zero. One possible solution in this case is to set a floor on the fee that should be charged by matchmakers. But

		Investor's Aggregate Welfare				
		First Best	Market Structure			
			MMD	CMD	CM	D
Order Processing Cost: λ						
	0.2	0.84	95%	98%	41%	95%
	0.5	0.72	70%	84%	48%	70%
	0.7	0.7	43%	65.5%	50%	43%
	0.82	0.69	26%	47.5%	50.7%	26%
	0.99	0.67	7%	52.3%	52.1%	1.5%

Table 3: Market Structure and Investors' welfare

the determination of the optimal floor is delicate. For instance a too high floor may enable the matchmakers to sustain the monopoly solution for which investors' welfare is 21.5% less than when the matchmakers set a zero fee (since λ is high, a monopoly matchmaker doesn't face much competition from dealers).

6 Conclusion

In this paper, we have analyzed the effect of inter-market competition on trading platforms' optimal pricing policy and investors' order placement strategies. Our main finding is that an increase in the trading fee on a limit order market has a non monotonic effect on limit order fill rates. The reason is that this increase reduces the surplus to be split between makers and takers in each transaction. Thus, for a fixed division of this surplus, it makes the outside option of takers (an immediate trade in a dealer market) more attractive. As a consequence, makers' market power is reduced, which, for some parameter values, forces them to make offers with a higher execution probability. For this reason, a decrease in trading fees (due for instance to competition) does not always result in a higher market share for the limit order market or higher expected gains from trade (as unfilled limit orders result in a welfare loss).

We have also analyzed the effects of make/take fees. In our model, a change in make/take fees that leaves the total fee unchanged affects the raw bid-ask spread but it leaves the cum fee bid-ask spread unchanged. For this reason, it leaves the division of gains from trade between makers and takers unaffected. Thus, the make/take fee breakdown is neutral (i.e., it has no effect on trading volume and welfare). Only the total fee

matters. However, it is worth stressing that, in our setting, makers face no constraints on the prices that they can post. In reality, these prices must be posted on a grid with a fixed minimum price variation (e.g., 1 cent in the U.S). With such a friction, makers would not be able to fully neutralize the effect of a change in the make/take fee breakdown and this breakdown would therefore start playing a role. In fact Foucault, Kadan and Kandel (2009) develops a theory of optimal make/take fees in this case. In this theory investors cannot choose between limit and market orders. Thus, an interesting extension of our analysis would be to analyze the effect of a minimum price variation on the make/take fee breakdown.

7 Appendix

7.1 Appendix A: Proofs of the results with a single matchmaker

7.1.1 Proofs

Proof of Lemma 1. Using equations (3), we obtain

$$U_t^{bu}(B^*, f_l, \delta_L) = \left(\frac{\delta_L}{\delta_H} \right) U_t^{bu}(B^*, f_l, \delta_H). \quad (21)$$

Using this equation and equation (5), we deduce that

$$v_H - A_t^{r*}(\delta_L) - f_m \leq v_H - A_t^{r*}(\delta_H) - f_m, \quad (22)$$

which yields $A_t^{r*}(\delta_H) \leq A_t^{r*}(\delta_L)$. The same type of argument shows that $B_t^{r*}(\delta_i)$ increases in δ_i . ■

Proof of Proposition 1

First, we note that under Condition **C.1**, the set of parameters values such that $\kappa_{k-1} < \frac{G^d}{G^l} \leq \kappa_k$ is never empty. Second, by definition, a buyer (resp. seller) optimally submits a buy market order on the limit order market when the ask price posted in the limit order market is less (higher) than his (her) cut-off price. Hence, when we analyze investors' best responses in a given type of equilibrium, we just need to consider their best response when the limit order book is such that they optimally choose to submit a limit order (e.g., a buyer arrives and the posted ask price exceeds his cut-off price).

The steps to find the conditions under which a given type of equilibrium is obtained are identical for each step. Thus, for brevity, we just detail these steps for types #1 and

#2 equilibria. We provide the derivations for the other types of equilibria in the Internet Appendix for this paper.

Type 1 equilibrium: Assume that $\frac{G^d}{G^l} \leq \kappa_1$. In a type #1 equilibrium, patient and impatient buyers (resp. sellers) post a limit order when the limit order book features an ask (resp. bid) price higher (smaller) than their cut-off price. In this case, their expected payoff with their optimal limit order must be greater than the payoff they can obtain by trading immediately on the dealer market, G^d . Moreover, investors choose buy and sell limit orders with high fill rates. That is, $A^* = A^{r*}(\delta_H)$ and $B^* = B^{r*}(\delta_H)$ and limit orders at these prices execute with probability $\frac{1}{2}$. Using these remarks and equations (5) and (6), we deduce that

$$\begin{aligned} v_H - A^{r*}(\delta_H) - f_m &= \frac{\delta_H}{2}(v_H - B^{r*}(\delta_H) - f_l) + \frac{\delta_H}{2}G^d \\ B^{r*}(\delta_H) - v_L - f_m &= \frac{\delta_H}{2}(A^{r*}(\delta_H) - v_L - f_l) + \frac{\delta_H}{2}G^d \end{aligned}$$

Solving this system of equations yield closed-form solutions for $A^{r*}(\delta_H)$ and $B^{r*}(\delta_H)$ and therefore the ask and bid prices in a type #1 equilibrium.

We now check that the order placement strategy of each investor is a best response to other investors' order placement strategies. We first check that all investors are better off submitting a limit order rather than trading immediately in the dealer market. For instance consider a buyer who arrives when there is no sell limit order in the limit order book (i.e., $A_t = \bar{A}$). His expected utility with a limit order at $B^* = B^{r*}(\delta_H)$ is

$$U_l^{bu}(B^{r*}(\delta_H), f_l, \delta_L) = \frac{\delta_i}{2 + \delta_H}(G^l + G^d). \quad (23)$$

This expected is greater than his immediate surplus, G^d , if he buys in the dealer market iff

$$\frac{G^d}{G^l} \leq \left(\frac{2 + \delta_H - \delta_i}{\delta_i} \right)^{-1}. \quad (24)$$

Now, it can be checked that $\kappa_1 \leq \left(\frac{2 + \delta_H - \delta_i}{\delta_i} \right)^{-1}$. Thus, as $\frac{G^d}{G^l} \leq \kappa_1$, Condition (24) is satisfied. The same reasoning shows that as $\frac{G^d}{G^l} \leq \kappa_1$, sellers are better off submitting a limit order at A^* rather than trading immediately in the dealer market when they expect other investors to behave as in a type #1 equilibrium.

Now we check that when he submits a buy limit order, a buyer is better off choosing a limit order with high fill rate at price $B^* = B^{r*}(\delta_H)$ rather than a buy limit order with

low fill rate at price $B^{r*}(\delta_L)$. To see this, observe that in a type #1 equilibrium, impatient sellers' cut-off price solves

$$B^{r*}(\delta_L) - v_L - f_m = \frac{\delta_L}{2}(A^{r*}(\delta_H) - v_L - f_l) + \frac{\delta_L}{2}G^d,$$

that is

$$B^{r*}(\delta_L) = v_L + f_m + \frac{\delta_L}{2 + \delta_H}(G^l + G^d) \quad (25)$$

Thus, a buyer submitting a buy limit order with low fill rate expects a payoff

$$U_l^{bu}(B^{r*}(\delta_L), f_l, \delta_i) = \frac{\delta_i(1 - \pi)}{2} \left[G^l - \frac{\delta_L}{2 + \delta_H}(G^l + G^d) \right] + \frac{\delta_i(1 + \pi)}{2} G^d.$$

Therefore, using equation (23), the limit order with low fill rate is dominated by a limit order with high fill rate iff

$$\frac{(1 - \pi)}{2} \left[G^l - \frac{\delta_L}{2 + \delta_H}(G^l + G^d) \right] + \frac{(1 + \pi)}{2} G^d \leq \frac{1}{2 + \delta_H}(G^l + G^d).$$

After some algebra, this condition can be written,

$$\frac{[2\pi - (1 - \pi)(\delta_H - \delta_L)]}{[2\pi + (1 + \pi)\delta_H - (1 - \pi)\delta_L]} \geq \frac{G^d}{G^l}, \quad (26)$$

that is $\frac{G^d}{G^l} \leq \kappa_1$. In the same way, we can also show that a seller is better off submitting a sell limit order with high fill rate at $A^* = A^{r*}(\delta_H)$ rather than a buy limit order with low fill rate at $A^{r*}(\delta_L)$ iff $\frac{G^d}{G^l} \leq \kappa_1$.

Type 2 equilibrium: Assume that $\kappa_1 < \frac{G^d}{G^l} \leq \kappa_2$. In a type #2 equilibrium, patient and impatient buyers (resp. sellers) choose to post a limit order when the limit order book does not feature an ask price less (higher than) than their cut-off price. Hence, their expected payoff with their optimal limit order must be greater than the payoff they can obtain by trading immediately on the dealer market, G^d . Moreover, in a type 2 equilibrium, buy and sell limit orders have a low fill rate. That is, $A^* = A^{r*}(\delta_L)$ and $B^* = B^{r*}(\delta_L)$ and limit orders at these prices execute with probability $\frac{1-\pi}{2}$. Using these remarks and equations (5) and (6), we deduce that in a type #2 equilibrium

$$\begin{aligned} v_H - A^{r*}(\delta_L) - f_m &= \frac{\delta_L}{2}(1 - \pi)(v_H - B^{r*}(\delta_L) - f_l) + \frac{\delta_L}{2}(1 + \pi)G^d, \\ B^{r*}(\delta_L) - v_L - f_m &= \frac{\delta_L}{2}(1 - \pi)(A^{r*}(\delta_L) - v_L - f_l) + \frac{\delta_L}{2}(1 + \pi)G^d. \end{aligned}$$

Solving this system of equations yield closed-form solutions for $A^{r*}(\delta_H)$ and $B^{r*}(v_H, \delta_H)$ and therefore the ask and bid prices in a type #2 equilibrium.

We now check that the order placement strategy of each investor is a best response to other investors' order placement strategies. We first check that all investors are better off submitting a limit order rather than trading immediately in the dealer market. For instance consider a buyer who arrives when there is no sell limit order in the limit order book (i.e., $A_t = \bar{A}$). His expected utility with a limit order at $B^* = B^{r*}(\delta_L)$ is

$$U_l^{bu}(B^*, f_l, \delta_L) = \frac{\delta_i}{2 + \delta_i(1 - \pi)} \left((1 - \pi)G^l + (1 + \pi)G^d \right). \quad (27)$$

It is easily checked that $U_l^{bu}(B^*, f_l, \delta_L) \geq G^d$ iff $\frac{G^d}{G^l} \leq \kappa_2$ as assumed in this case. Thus, impatient buyers are better off submitting limit orders with low fill rates rather than trading in the dealer market when they expect other investors to behave as in a type #2 equilibrium. This is a fortiori true for patient buyers, and the proof is symmetric for sellers.

Now we check that when he submits a buy limit order, a buyer is better off choosing a limit order with a low fill rate at $B^* = B^{r*}(\delta_L)$ rather than a buy limit order with high fill rate at $B^{r*}(\delta_H)$. To see this, observe that in a type 2 equilibrium, patient sellers' cut-off price solves

$$B^{r*}(\delta_H) - v_L - f_m = \frac{\delta_H}{2} (A^{r*}(\delta_L) - v_L - f_l) + \frac{\delta_H}{2} G^d,$$

that is:

$$B^{r*}(\delta_H) = v_L + f_m + \frac{\delta_H}{2 + \delta_L(1 - \pi)} \left((1 - \pi)G^l + (1 + \pi)G^d \right), \quad (28)$$

Thus, a buyer submitting a buy limit order with high fill rate expects a payoff

$$U_l^{bu}(B^{r*}(\delta_H), f_l, \delta_i) = \frac{\delta_i}{2} \left[G^l - \frac{\delta_H}{2 + \delta_L(1 - \pi)} \left((1 - \pi)G^l + (1 + \pi)G^d \right) \right] + \frac{\delta_i}{2} G^d.$$

Therefore, using equation (27), we deduce that the expected payoff with a buy limit order with low fill rate is higher iff

$$\frac{G^l}{G^d} (2\pi - (\delta_H - \delta_L)(1 - \pi)) < 2\pi + \delta_H(1 + \pi) - \delta_L(1 - \pi)$$

After some algebra, this condition can be written,

$$\frac{[2\pi - (1 - \pi)(\delta_H - \delta_L)]}{[2\pi + (1 + \pi)\delta_H - (1 - \pi)\delta_L]} < \frac{G^d}{G^l}, \quad (29)$$

that is $\frac{G^d}{G^t} > \kappa_1$. In the same way, we can also show that a seller is better off submitting a limit order with low fill rate at $A^* = A^{r*}(\delta_L)$ rather than a buy limit order with high fill rate at $A^{r*}(\delta_H)$ iff $\frac{G^d}{G^t} > \kappa_1$. ■

Proof of Corollary 1 Direct using the expressions for the quotes in Proposition 1. ■

Proof of Corollary 2. Direct using the expressions for the quotes in Proposition 1. ■

Proof of Corollary 3.

Consider the market for the security at date t . At this date, this market can be in 6 possible states: (0) closed because the asset has already paid its cash-flow; (1) active, a patient investor arrives and submits a limit order; (2) active, a patient investor arrives and submits a market order; (3) active, an impatient investor arrives and submits a limit order; (4) active, an impatient investor arrives and submits a market order; (5) active, an impatient investor arrives and trades upon arrival in the dealer market. Transitions from one state to another follows a Markov chain with the following transition matrix, \hat{P}_k

$$\hat{P}_k = \begin{pmatrix} 1 & \mathbf{0}' \\ (1 - \rho)\mathbf{1} & \rho\hat{M}_k \end{pmatrix}$$

where $\mathbf{0}$ and $\mathbf{1}$ are 5×1 vectors and \hat{M}_k is a 5×5 matrix that depends on the type of equilibrium, k . For instance, given equilibrium decisions in an equilibrium of type #1, we have

$$\hat{M}_1 = \begin{pmatrix} \frac{\pi}{2} & \frac{\pi}{2} & \frac{1-\pi}{2} & \frac{1-\pi}{2} & 0 \\ \pi & 0 & 1-\pi & 0 & 0 \\ \frac{\pi}{2} & \frac{\pi}{2} & \frac{1-\pi}{2} & \frac{1-\pi}{2} & 0 \\ \pi & 0 & 1-\pi & 0 & 0 \\ \pi & 0 & 1-\pi & 0 & 0 \end{pmatrix}$$

As state 0 is absorbing it is clear that after some time the process will be in state 0 and the only stationary distribution of this process gives a weight of 1 to this state (that is, the market closes with probability 1 when $\rho < 1$).

Let modify the matrix \hat{M}_k by deleting rows and columns corresponding to states that are never entered (for instance state 5 in a type #1 equilibrium) so that the matrix, now called M_k , is indecomposable, and let P_k be the transition matrix with this modified matrix.

For instance

$$M_1 = \begin{pmatrix} \frac{\pi}{2} & \frac{\pi}{2} & \frac{1-\pi}{2} & \frac{1-\pi}{2} \\ \pi & 0 & 1-\pi & 0 \\ \frac{\pi}{2} & \frac{\pi}{2} & \frac{1-\pi}{2} & \frac{1-\pi}{2} \\ \pi & 0 & 1-\pi & 0 \end{pmatrix}$$

and

$$P_1 = \begin{pmatrix} 1 & \mathbf{0}' \\ (1 - \rho)\mathbf{1} & \rho\hat{M}^k \end{pmatrix}$$

Now, we define $[q_{0k}(t), \mathbf{q}'_k(\mathbf{t})]$ as the probability distribution over all states at time t in an equilibrium of type k and we denote by $\mathbf{d}_k(\mathbf{t})$ the probability distribution overall all states conditional on the process not having been absorbed, that is,

$$\mathbf{d}_k(\mathbf{t}) \equiv \frac{\mathbf{q}_k(\mathbf{t})}{\mathbf{1} - \mathbf{q}_{0k}(\mathbf{t})}$$

If $\mathbf{d}_k(\mathbf{t} + \mathbf{1}) = \mathbf{d}_k(\mathbf{t}) = \mathbf{d}_k$, then \mathbf{d}_k is called a stationary conditional distribution. Darroch and Seneta (1965) show that \mathbf{d}_k is the left eigenvector of ρM_k corresponding to the maximum-modulus eigenvalue of ρM_k . In our setting it is easy to see that \mathbf{d}_k is just the stationary distribution associated with M_k .²⁵ We call φ^k this distribution, to which we add a 0 for each state we deleted when rewriting \hat{M}_k as M_k . Thus, φ_j^k is the stationary probability of state j at any date conditional on the cash-flow of the security not being paid at date t and we obtain

$$\begin{aligned} \varphi^1 &= \left(\frac{2\pi}{3}, \frac{\pi}{3}, \frac{2(1-\pi)}{3}, \frac{1-\pi}{3}, 0 \right) \\ \varphi^2 &= \left(\pi, 0, \frac{(1-\pi)(2-\pi)}{3-\pi}, \frac{1-\pi}{3-\pi}, 0 \right) \\ \varphi^3 &= \left(\pi, 0, 0, \frac{\pi(1-\pi)}{2}, \frac{(1-\pi)(2-\pi)}{2} \right) \\ \varphi^4 &= \left(\frac{2\pi}{2+\pi}, \frac{\pi^2}{2+\pi}, 0, \frac{\pi(1-\pi)}{2+\pi}, \frac{2(1-\pi)}{2+\pi} \right) \end{aligned}$$

The corollary then follows from equations (17), (18) and (19).

Remark: Probability φ_j^k can also be seen as the proportion of time spent in each state when time to absorption is long. To see this, assume the process starts in state $i \in \{1, \dots, 5\}$ with probability π_i , denote T_i the first time at which it reaches state 0 conditional on starting in state i , and N_{ij} be the number of visits of state j conditional on starting in state i . Consider

$$m_j = \lim_{t \rightarrow \infty} \sum_i \pi_i E \left(\frac{N_{ij}}{T_i} | T_i = t \right).$$

Variable m_j is the proportion of time spent in state j before absorption when the time to absorption is very long. Darroch and Seneta (1965) show that $m_j = w_j d_j$, where

²⁵Because this vector is by definition associated with the eigenvalue 1, M^k being stochastic this is the maximum-modulus eigenvalue.

$\mathbf{w} = (w_1, \dots, w_5)$ is the right eigenvector of ρM corresponding to the same eigenvalue as \mathbf{d} . As the probability of transition from any state to state 0 is always ρ , it is easy to show that we have $\mathbf{w} = \mathbf{1}$. Hence, $m_j = \varphi_j$ for $i \in \{1, \dots, 5\}$. ■

Proof of Proposition 2 (Sketch, see the Internet appendix for the full proof)

As explained in the text, the unconstrained first best is necessarily such that the trading fee is zero and makers' surplus is zero. These variables being fixed, the central planner can choose to allocate roles as in a type #1, #2, #3, #4 or #5 equilibrium. Let $W_j(\lambda)$ if the central planner chooses an role allocation identical to that obtained in a type #k equilibrium.

Suppose that roles are allocated as in a type #1 equilibrium. Then, $FR = \frac{1}{2}$, $TR = 33\%$, $\varphi_1 = \frac{2\pi}{3}$, $\varphi_3 = \frac{2(1-\pi)}{3}$. Thus, using the expression for investors' welfare (equation (20)) for these values of the parameters, we get that in this case:

$$W_1(\lambda) = \frac{(2L)}{3} + \frac{L - \lambda}{3} - \frac{2}{3}(\pi(1 - \delta_H) + (1 - \pi)\delta_L)(L - \lambda)$$

We can proceed in the same way to derive the expressions for investors' welfare in each case. Then direct comparisons of investors' welfare in each possible case yields the proposition. ■

Proof of Corollary 4 Suppose that $\lambda \in [\lambda_3, \lambda_2)$. If $f^T = 0$, a type #3 equilibrium obtains. In this case

$$W(\lambda, 0) = L(1 - \pi(1 - \delta_H)) - \lambda(1 - \pi(1 - \delta_H\pi)).$$

Now suppose that $f^T = f_3^T(\lambda) + \epsilon$ where ϵ is very small. Then a type #4 equilibrium is obtained. Investors' welfare is

$$W(\lambda, f_3^T(\lambda)) = \frac{(L - \lambda)}{(2 + \pi)(2\pi - \delta_H(1 - \pi))} (4\pi(1 - \pi) + \delta_H(7\pi^2 + \pi - 2))$$

Denoting $\Delta W(\lambda) = W(\lambda, f_3^T(\lambda)) - W(\lambda, 0)$, we can show that, under **C.1**, ΔW is linearly decreasing in λ and $\Delta W(\lambda_3) > 0$. Depending on the parameters, under **C.1**, $\Delta W(\lambda_2)$ can be either positive or negative. Thus there exists $\hat{\lambda} \in (\lambda_3, \lambda_2]$ such that $\Delta W(\lambda) \geq 0$ iff $\lambda_3 < \lambda \leq \hat{\lambda}$. ■

Proof of Proposition 3

If the platform chooses a fee equal to $f_k^T(\lambda)$ then a type k equilibrium is obtained. The expected profit of the platform is then $\Pi(f_k^T, \lambda) = TR_l^k \times f_k^T(\lambda)$. Using the expression for $f_k^T(\lambda)$ (equation (13)) and TR_l^k (Corollary 3), we obtain that

$$\begin{aligned} \Pi(f_1^T, \lambda) \geq \Pi(f_2^T, \lambda) &\Leftrightarrow \frac{L - \lambda}{2L} \leq \frac{2\pi}{(3 - \pi)\kappa_1^{-1} - 3(1 - \pi)\kappa_2^{-1}} \Leftrightarrow \lambda \geq \lambda'_1, \\ \Pi(f_1^T, \lambda) \geq \Pi(f_4^T, \lambda) &\Leftrightarrow \frac{L - \lambda}{2L} \leq \frac{2(1 - \pi)}{(2 + \pi)\kappa_1^{-1} - 3\pi\kappa_4^{-1}} \Leftrightarrow \lambda \geq \lambda'_2, \\ \Pi(f_2^T, \lambda) \geq \Pi(f_4^T, \lambda) &\Leftrightarrow \frac{L - \lambda}{2L} = \frac{2(1 - 2\pi)}{(1 - \pi)(2 + \pi)\kappa_2^{-1} - \pi(3 - \pi)\kappa_4^{-1}} \Leftrightarrow \lambda \geq \lambda'_3. \end{aligned}$$

The first and the second part of the proposition follows. ■

Proof of Corollary 6. Immediate from inspection of the expression for $f^{T*}(\lambda)$ in Proposition 3. ■

Proof of Corollary 7. Immediate from inspection of the expression for $f^{T*}(\lambda)$ in Proposition 3. ■

Proof of Proposition 6. (Sketch, see the Internet appendix for the full proof)

The proof relies on direct comparisons of investors' welfare under the different market structures. Although writing investors' welfare in each market structure is tedious, comparing the value of investors' welfare in each market structure is straightforward since investors' welfare is a linear function of λ . See the Internet appendix for the detailed proof. ■

7.2 Appendix B: Extension with Two Matchmakers

7.2.1 Equilibrium definition with two matchmakers

Let $U_{mj}^{bu}(A_j, f_m)$ and $U_{lj}^{bu}(B_j, f_{lj}, \delta_i)$ be respectively the payoffs of a market order at price A_j and a limit order at price B_j for a buyer with type δ_i . Similarly let $U_{mj}^{se}(B_j, f_{mj})$ and $U_{lj}^{se}(A_j, f_{lj}, \delta_i)$ be respectively the payoffs of a market order at price B_j and a limit order at price A_j for a seller with type δ_i . We denote by B_j^* the optimal bid price for a buyer if he submits a buy limit order on platform j and by A_j^* the optimal ask price for a seller if she submits a sell limit order on platform j . As limit orders stay only one period in the limit order market, there is at most one quote in the consolidated market at any given point in time. Thus, if an investor's limit order is unfilled after one period, the investor uses the

dealer market in last resort as in the baseline model. Consequently, investors' payoffs are defined as in equations (1), (2), (3) and (4).

The investor may be indifferent between submitting a limit order on platform 1 or on platform 2²⁶. This happens when $U_{l1}^{bu}(B_1^*(v_H), f_{l1}, \delta_i) = U_{l2}^{bu}(B_2^*(v_H), f_{l2}, \delta_i)$. In this case we assume that the investor routes his order to either platform with equal probabilities²⁷.

Let A_{tj} and B_{tj} be the best ask and bid prices posted on platform j at date t , with the convention that $A_{jt} = \bar{A}_j = \infty$ ($B_{jt} = \bar{B}_j = -\infty$) if there is no sell (buy) limit order on platform j at date t . Consider a buyer arriving at date t . He now has five options: (i) a buy market order on platform 1, (ii) a buy market order on platform 2, (iii) a buy limit order at price B_1^* on platform 1, (iv) a buy limit order at price B_2^* on platform 2 and (v) a buy order in the dealer market. We denote these five options: b_{m1} , b_{m2} , B_1^* , B_2^* and b_d , respectively and the buyer's choice among these options by $O_b(\delta_i, A_t) \in \{b_{m1}, b_{m2}, B_1^*(v_H), B_2^*(v_H), b_d\}$.

In the same way, we denote by, s_{m1} , s_{m2} , A_1^* , A_2^* and s_d , the five possible options for a seller arriving at, say, date t and we denote the seller's choice among these options by $O_s(\delta_i, B_t) \in \{s_{m1}, s_{m2}, A_1^*, A_2^*, s_d\}$.

The equilibrium is then defined as in the baseline model. We can also proceed as in the baseline model to solve for the equilibrium order placement strategies. In particular, let define $A_j^{r*}(\delta_i)$ and $B_j^{r*}(\delta_i)$ as the solutions to:

$$v_H - A_j^{r*}(v_H, \delta_i) - a_{jt} = \text{Max}\{U_l^{bu}(B_1^*(v_H), f_l, \delta_i), U_l^{bu}(B_2^*(v_H), f_l, \delta_i), G^d\}, \quad (30)$$

$$B_j^{r*}(v_L, \delta_i) - v_L - a_{jt} = \text{Max}\{U_l^{se}(A_1^*(v_L), f_l, \delta_i), U_l^{se}(A_2^*(v_L), f_l, \delta_i), G^d\}. \quad (31)$$

Thus, a buyer with type δ_i is better off submitting a buy (sell) market order on platform j rather than a limit order on any platform if and only $A_{jt} \leq A_j^r(v_H, \delta_i)$. Moreover if $A_{jt} \leq A_j^r(\delta_i)$ then the ask side on the other limit order book is empty since at any point in time there is at most one limit order in the consolidated market. Thus, a buyer with type δ_i submits a buy market order on platform j if and only if $A_{jt} \leq A_j^r(\delta_i)$, as in the

²⁶The investor may also be indifferent between submitting a market order on platform 1 or on platform 2. But this case never happens in our model. Actually, at any given point in time, only one platform will feature the best ask price or the best bid price. Thus, at a given point in time, an investor choosing a market order will never be indifferent between both platforms.

²⁷Without this assumption proposition 4 would still hold except for the routing probabilities. Proposition 5 would a fortiori hold since the platform attracting less orders in equilibrium would have larger incentives to undercut its rival.

baseline model. Similarly, a seller with type δ_i submits a sell market order to platform j iff $B_j^* \geq B_j^{r*}(\delta_i)$.

As in the baseline model, impatient investors are willing to pay larger concessions for an immediate trade on market j than patient investors. That is, $A_j^r(\delta_i)$ decreases in δ_i and $B_j^r(\delta_i)$ increases in δ_i . Thus, if a buyer decides to submit a limit order on platform j , he picks one of two prices, as in the baseline model. Either he places a limit order with low execution probability ($\frac{1-\pi}{2}$) at price $B_j^r(\delta_L)$ or he places a limit order with high execution probability ($\frac{1}{2}$) at price $B_j^r(\delta_H)$. The prices chosen by a seller are symmetric. Thus, as in the baseline case, there are four possible types of equilibria (specialized/unspecialized; low fill rate/high fill rate) in which at least one limit order market is active.

Moreover, as in the baseline model, an investor never submits a limit order if his/her expected payoff with such an order on either platform is too small relative to the payoff of an immediate trade on the dealer market. Last, investors use the services of both matchmakers if and only if the payoffs of an optimal sell limit order and an optimal buy limit order are identical on both platforms. Otherwise investors will only route their orders to the platform that yields the highest expected payoff and the other platform is inactive. For instance, if:

$$\begin{aligned} U_{l1}^{bu}(B_1^*, f_{l1}, \delta_i) &> U_{l2}^{bu}(B_2^*, f_{l2}, \delta_i), \\ U_{l1}^{se}(A_1^*, f_{l1}, \delta_i) &> U_{l2}^{se}(A_2^*, f_{l2}, \delta_i), \end{aligned}$$

then, buyers and sellers never find optimal to submit a limit order on platform 2 and therefore only platform 1 is active.

References

- [1] Barclay, M., T. Hendershott and T. McCormick, 2003, “Competition among Trading Venues: Information and Trading on Electronic Communication Networks”, *Journal of Finance*, 2637-2665.
- [2] Biais, B., C. Bisière and C. Spatt, 2004, “Imperfect Competition in Financial Markets”, working paper, Toulouse University.
- [3] Boehmer, B. and E. Boehmer, 2004, “Trading your Neighbor’s ETFs’: Competition and Fragmentation”, *Journal of Banking and Finance*, **27**, 1667-1703.
- [4] Cantillon, E. and Yin, P.L., 2010, “Competition between Exchanges: Lessons from the Battle of the Bund,” Working paper, MIT and Université Libre de Bruxelles.
- [5] Darroch J.N. and E. Seneta, 1965, “On Quasi-Stationary Distributions in Absorbing Discrete-Time Finite Markov Chains” *Journal of Applied Probability* 2 , 88-100.
- [6] DeFontnouvelle, P., R. Fishe and J. Harris, 2003, “The Behavior of Bid-Ask Spreads and Volume in Options Markets during the Competition for Listings in 1999”, *Journal of Finance*, 2437-2463.
- [7] Degryse, H., Van Achter, M., and G. Wuyts, 2009, “Dynamic Order Submission Strategies with Competition between a Dealer Market and a Crossing Network”, *Journal of Financial Economics* 91, 319-338.
- [8] Degryse, H., Van Achter, M., and G. Wuyts, 2010, “Internalization, Clearing and Settlement, and Stock Market Liquidity”, *mimeo*, Tilburg University.
- [9] Duffie, D., Garleanu N. and Pedersen, L. (2009) “Over-the-Counter Markets,” *Econometrica*, 73, 1815-1847.
- [10] Foucault T., Kadan O. and Kandel E. (2005), “Limit Order Book as a Market for Liquidity,” *Review of Financial Studies*, 18, 1171-1217.
- [11] Foucault, T., Kadan, O. and Kandel, E. (2009): “Liquidity Cycles, and Make/Take Fees in Electronic Markets” CEPR, Discussion Paper n°7551.

- [12] Foucault Thierry and Albert J. Menkveld, “Competition for Order Flow and Smart Order Routing Systems,” *Journal of Finance*, 63, 119-158, 2008
- [13] Gehrig, T., 1993 “Intermediation in Search Markets.” *Journal of Economics and Management Strategy* 2, 97–120.
- [14] Glosten, L., 1994, “Is the Electronic Order Book Inevitable”, *Journal of Finance*, **49**, 1127–1161.
- [15] Goettler, R. L., C. A. Parlour, and U. Rajan (2009). “Informed traders and limit order markets.” *Journal of Financial Economics* 93(1), 67–87.
- [16] Hendershott, T. and Mendelson, H., (2000), “Crossing Networks and Dealer Markets: Competition and Performance”, *Journal of Finance*, 55, 2071-2115.
- [17] Hollifield, B., Miller, R. A., and Sandas, P. (2004) “Empirical analysis of limit order markets.” *Review of Economic Studies* 71, 1027-1063.
- [18] Hollifield, B., Miller, R. A., Sandas, P., and Slive J. (2006) “Estimating the gains from trade in limit order markets.” *Journal of Finance* 61, 2753-2804.
- [19] Maskin, E. and Tirole, J. (1997) “Markov Perfect Equilibrium: I. Observable Actions,” *Journal of Economic Theory*, 191-219.
- [20] O’Hara, M. and Ye, M., 2009, “Is market fragmentation harming market quality,” working paper, Cornell university.
- [21] Pagano, M., 1989, “Trading Volume and Asset Liquidity”, *Quarterly Journal of Economics*, 104, 255-274.
- [22] Parlour, C., and D. Seppi, 2003, “Liquidity-Based Competition for Order Flow”, *Review of Financial Studies*, **16**, 301-343.
- [23] Rust, J. and G. Hall “Middlemen versus Market Makers: A Theory of Competitive Exchange,” *Journal of Political Economy* 111, 353-403.
- [24] Spulber, D. “Market making by price-setting firms.” *Review of Economics Studies* 63, 559–80.

- [25] Stoll, Hans R. (2000), "Friction", *Journal of Finance*, 55(4), 1479-1514
- [26] U.S. Securities and Exchange Commission, 2000, Release N°34-42450
- [27] Viswanathan, V. and J. Wang (2002), "Market Architecture: Limit Order Books vs. Dealership Markets", *Journal of Financial Markets*, **5**, 127-167.
- [28] Yavas, A., 1992 "Marketmakers versus matchmakers." *Journal of Financial Intermediation* 2, 33–58.

FIGURE 1

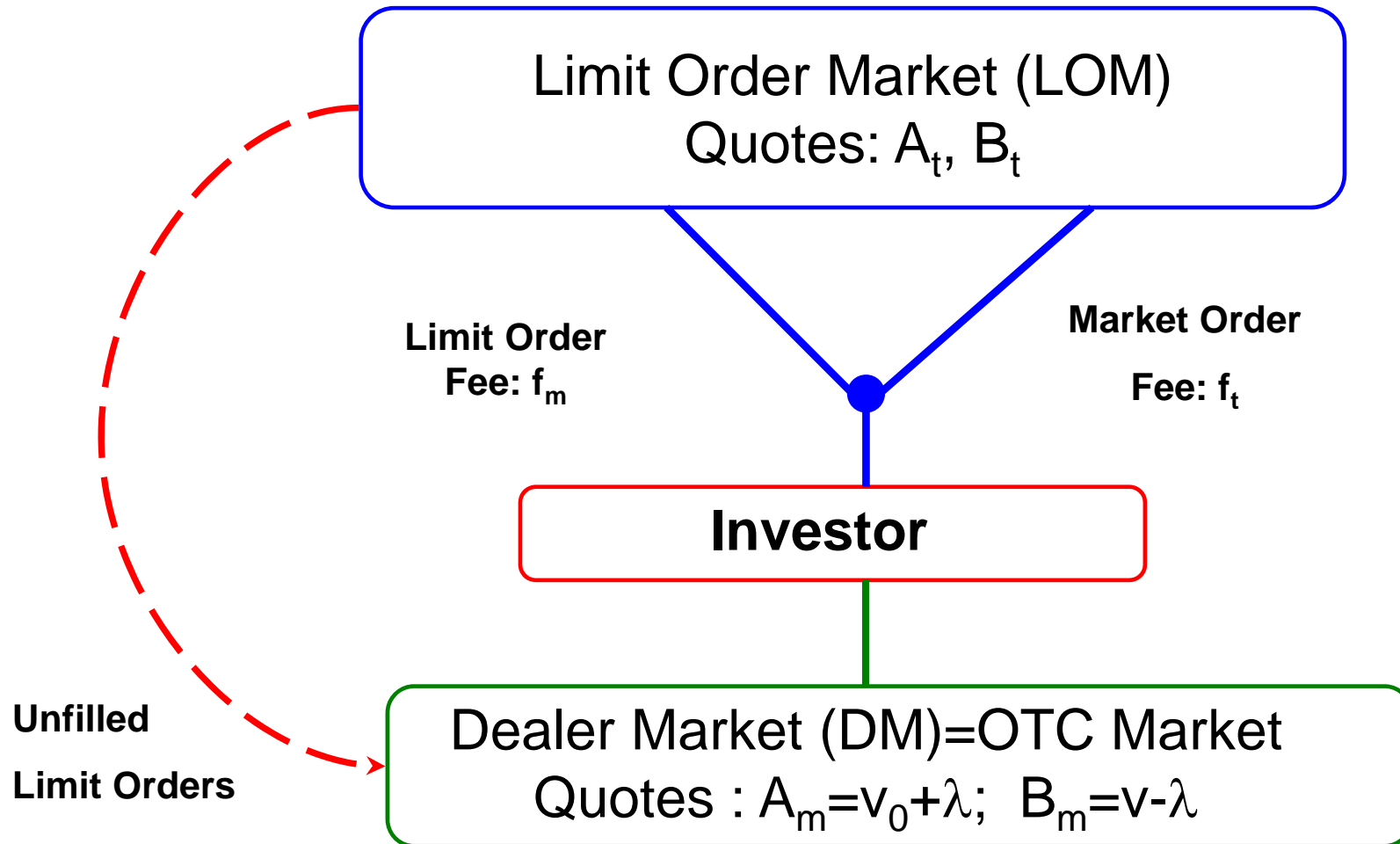


FIGURE 2

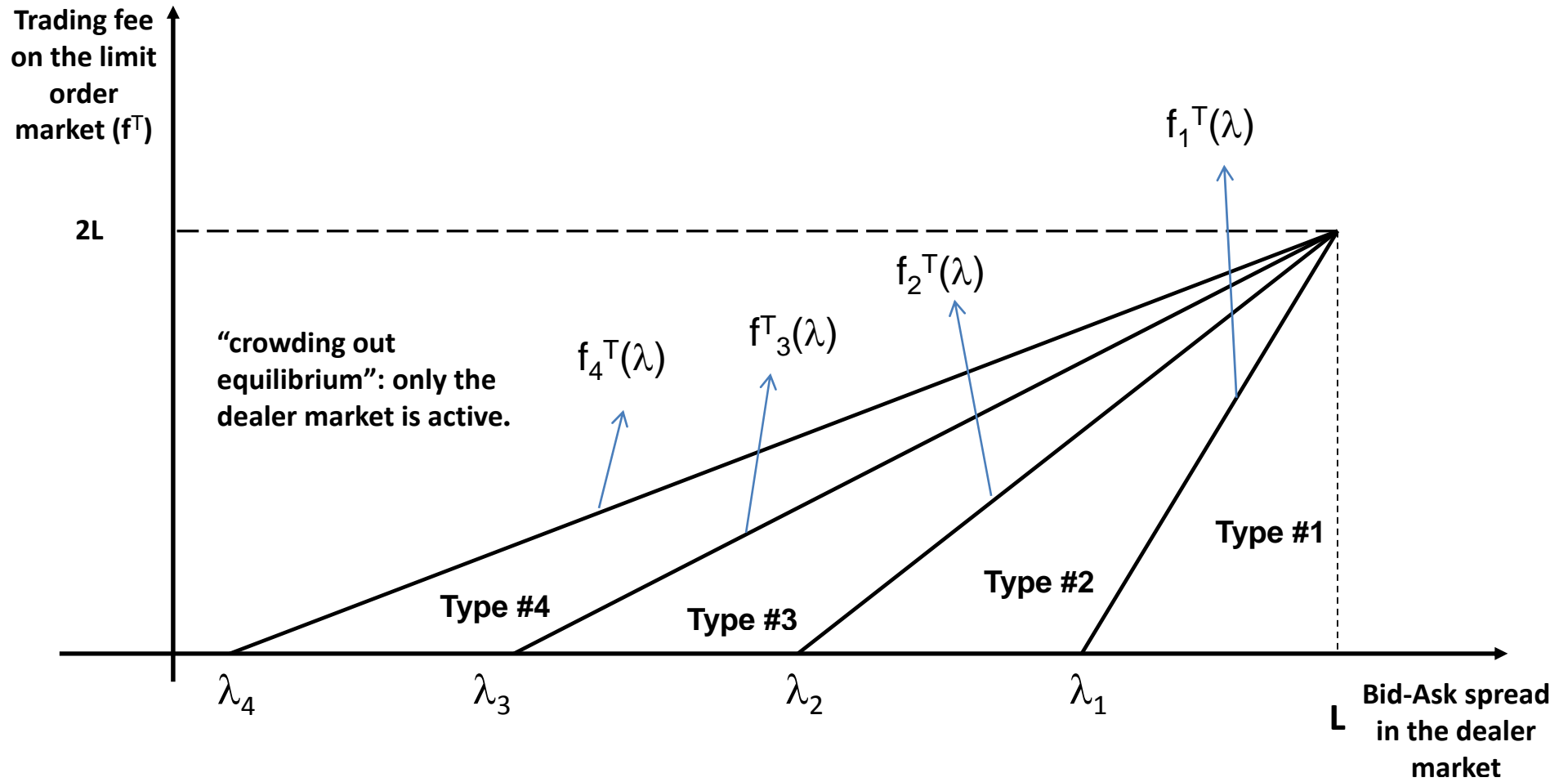


FIGURE 3

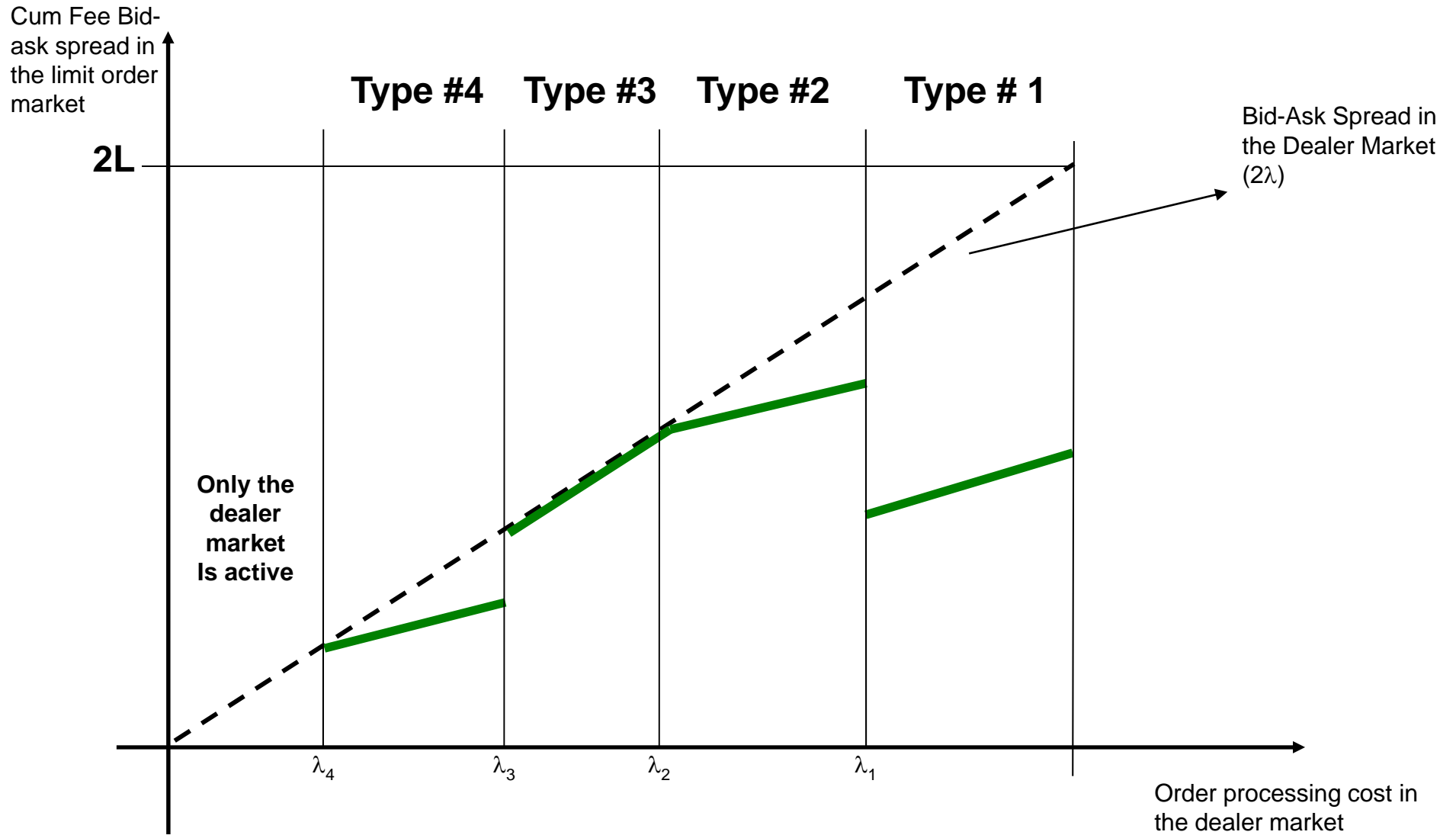


FIGURE 4

