DISCUSSION PAPER SERIES

No. 8395

TRADING FEES AND EFFICIENCY IN LIMIT ORDER MARKETS

Jean-Edouard Colliard and Thierry Foucault

FINANCIAL ECONOMICS and INDUSTRIAL ORGANIZATION

Centre for Economic Policy Research

www.cepr.org

Available online at: www.cepr.org/pubs/dps/DP8395.asp

ISSN 0265-8003

TRADING FEES AND EFFICIENCY IN LIMIT ORDER MARKETS

Jean-Edouard Colliard, Paris School of Economics
Thierry Foucault, HEC, Paris and CEPR

Discussion Paper No. 8395
May 2011

Centre for Economic Policy Research
77 Bastwick Street, London EC1V 3PZ, UK
Tel: (44 20) 7183 8801, Fax: (44 20) 7183 8820
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 8395

May 2011

ABSTRACT

Trading Fees and Efficiency in Limit Order Markets*

We study competition between a dealer (OTC) market and 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 an increase in the matchmaker's trading fee can raise investors' ex-ante expected welfare. Actually, it induces makers to post more aggressive offers and thereby it raises the likelihood of a direct trade between investors. For this reason as well, a reduction in the matchmaker's trading fee can counter-intuitively raise the OTC market share. However, entry of a new matchmaker results in an improvement in investors' welfare, despite its negative effect on trading fees. 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.

JEL Classification: G00, G18, G20 and L10
Keywords: inter-market competition, limit order markets, liquidity, make/take fees, OTC markets and trading fees

Jean-Edouard Colliard
Paris School of Economics
48 Boulevard Jourdan
75014 Paris
FRANCE

Email: colliard@pse.ens.fr

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=173576

Thierry Foucault
HEC, Paris
1 rue de la Libération
78351 Jouy-en-Josas
FRANCE

Email: foucault@hec.fr

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=121810

* We thank Bruno Biais, Amil Dasgupta, Hans Degryse, Estelle Cantillon, Thomas Gehrig, Patrick Legros, Stefano Lovo, Albert Menkveld, Sophie Moinas, Mark Van Achter and participants at the conference on the Industrial Organization of Securities Markets in Frankfurt, the 2011 MTS conference in London, and seminars at ECARES, the Toulouse School of Economics, the Paris School of Economics and ENSAE for their comments. The authors gratefully acknowledge the Institut Louis Bachelier for its financial support. All errors are ours.

Submitted 04 May 2011

1 Introduction

The industrial organization of securities markets is changing fast, both in Europe and in North America. 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, incumbent markets match a smaller fraction of trades in their listings and trading is more fragmented. For instance, the market shares of Nasdaq and the NYSE (as a fraction of total trading volume) were respectively only 22.8% and 27% in April 2009. A similar evolution is observed in Europe where the market share of incumbent markets has sharply declined since 2007 (for instance, the market share of the London Stock Exchange in the constituent stocks of the FTSE100 was about 60% in 2009, down from 80% in 2007).[1]

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 both sides are often charged different fees.[2] As competition among platforms is heating up, these fees steadily decline.

The effects of this evolution are very much debated but not well understood yet. For instance, the SEC has to date received more than two hundred comments on its concept release regarding the organization of U.S. equities markets, many of which stress the importance of trading fees to understand the new trading landscape in U.S. securities markets.[3] The reason is that, even

[1] The figures for U.S exchanges are from "*NYSE, Nasdaq lose market share of U.S. equity trading in April.*" Bloomberg Business Week, May 03, 2010. The figures for European markets are based on authors' calculations using data from Thomson-Reuters.

[2] For 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 \$0.0023 (per share) to the maker. The net revenue for NYSEArca was therefore \$0.0007 (per share traded).

[3] See http://www.sec.gov/comments/s7-02-10/s70210.shtml

though trading fees per share are small, they represent a significant source of revenue for trading platforms given the high level of trading activity in equities markets. Moreover, as pointed out by the joint CFTC-SEC Advisory Committee on Emerging Regulatory Issues, the differentiation of trading fees between makers and takers could alter the balance between liquidity supply and liquidity demand. Similarly, trading fees are at the heart of recent regulatory debates in Europe. For instance, a recent consultation paper by the Committee of European Securities Regulators (CESR) asks: "*What are the impacts of current fee structures on trading platforms, participants, their trading strategies and the wider market and its efficiency?".*[4]

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 goal here is to fill this gap. An intriguing finding is that an increase in trading fees can, surprisingly, *increase* investors' welfare. Indeed, it may induce makers to post offers with a higher execution probability. Hence, trades are intermediated by dealers less frequently and as a result welfare improves because investors save on dealers' dead-weight intermediation cost.

Our model features the market for a security populated by buyers (investors with a high private value for the security) and sellers (investors with a low private value).[5] Buyers and sellers arrive sequentially and have a deadline to carry out their trade. Upon arrival, an investor can choose to trade either in a limit order market or in a dealer market. In this way, we account for the fact that limit order markets often face competition from over the counter (OTC) dealer markets. For instance, U.S. treasuries simultaneously trade in limit order markets (E.Speed and BrokerTec) and a dealer market while in European and U.S. equities markets, brokerage firms

[4] See "Call for Evidence: Micro-Structural Issues of the European equity markets." Available at http://www.cesr.eu/.

[5] This model builds upon Foucault (1999), who does not study how trading fees affect the make-take decision for investors.

can “internalize” the execution of an order by passing it to their market-making branch. 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.

Dealers in the OTC market bear an order processing cost each time they execute a trade. This cost sets the bid-ask spread in the OTC market as competition among dealers drive their profits to zero. In the limit order market, there is no order processing cost since 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 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 the effects of changing both the make/take fee breakdown and the total fee.

As a benchmark, we analyze the allocation of matches (between investors or between investors and dealers) that maximizes investors’ welfare (the first best for investors). In any trading mechanism, incentive 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 with the trading mechanisms considered in the paper. Not surprisingly, the first best requires a zero trading fee since this fee simply reduces the gains from trade available to investors.

By comparing investors’ actual expected gains from trade in equilibrium to this benchmark,

[6] The first best is the allocation of trades that maximizes investors’ welfare assuming that the central planner choosing this allocation knows the investors’ type. Clearly, no mechanism can do better than this allocation.

we identify two sources of inefficiencies. First, the dealer market can "crowd out" the limit order market for too large a set of parameters relative to the first best. Intuitively, when the dealers' order processing cost is low enough, it can be efficient to concentrate all trades in the dealer market since this is a way to save on waiting costs. However, in equilibrium, the limit order market stops attracting orders even if the order processing cost is too large for the concentration of trading in the dealer market to be efficient. Indeed, a reduction in dealers' order processing cost reduces the rents that makers can extract from takers as the bid-ask spread in the dealer market is tighter. As a result, it may not be individually optimal for any trader to act as a maker (as an immediate trade in the dealer market yields a higher profit), even though it would be collectively optimal that some investors play this role. In this case, the limit order market is, inefficiently, inactive.

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 that in the first best for investors. An increase in the trading fee can lessen this inefficiency. 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 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 induces the former to post more aggressively priced limit orders, i.e., limit orders that have a higher likelihood of execution. As a result, an increase in the matchmaker's trading fee can, counter-intuitively, raise investors' welfare because it makes the likelihood of direct trades between buyers and sellers higher.

This effect raises the possibility that unbridled competition among trading platforms may lead to *too low* trading fees, even from the investors' point of view. To study whether this happens, we endogenize trading fees and compare investors' welfare in two types of market structures: (i) a single matchmaker and a dealer market and (ii) two competing matchmakers and a dealer market. 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. For this reason, there is a zone of parameters for which the trading rate in the (consolidated) limit order market is higher with a single matchmaker. In this case, competition among matchmakers has a welfare cost for investors: it lowers the likelihood of direct trades among them. However, the benefit of a lower transfer to the matchmaker always outweigh this cost. As a result investors are always better off when two matchmakers compete for their orders. We also find that there exist parameter values for which investors' welfare is higher when OTC trading is banned because the possibility of using the dealer market in last resort can induce makers to choose limit orders with low execution probabilities.

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 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 induces makers to place limit orders with lower execution probabilities, other things equal. As a consequence, the market share of the dealer market increases, even though trading on the limit order market becomes cheaper (even after accounting for the bid-ask spread). In other words, the correlation between trading fees and trading volume on a limit order market can be *positive*. Testing for this effect provides a sharp test of our theory since it is clearly non standard. Interestingly, Malinova and Park (2011) find evidence in line with this prediction of our model. Moreover, the recent entry of new trading platforms in U.S. and European equities markets coincide with both lower trading fees and an increase in the market share of the OTC market (the dealer market in our set-up).[7] 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 reduce investors' welfare for reasons explained previously. Thus, an intensification of competition among matchmakers does not necessarily improve investors' welfare. Hollifield et al. (2006) estimate investors' welfare in a limit order market (the Vancouver Stock Exchange). They find that the opportunity cost associated with unfilled limit orders is the main source of inefficiency in this 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 execution probabilities for limit orders. Hollifield et al. (2006)'s methodology could be used to test this prediction.

The model also generates implications about the effect of a change in make/take fees on the bid-ask spread. In equilibrium, the "raw" traded bid-ask spread (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 a wider or a tighter raw bid-ask spread 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 the 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),

[7] For the U.S. market, see "Nasdaq frets over internalization" by John d'Antona in Traders' Magazine (March 2010). For European markets, see the 2011 response of the Federation of European Securities Exchanges (FESE) to the consultation on the review of the markets in financial instruments directive (MiFID).

Foucault and Menkveld (2008), or Degryse et al. (2009)).[8] These theories usually do not consider the possibility for investors to act as a maker or a taker.[9] 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 absent from our analysis.[10]

More generally, our paper contributes to the theoretical literature on competition between markets for real or financial assets (e.g., Yavas (1992), Gehrig (1993), Spulber (1996) or Rust and Hall (2006); see Cantillon and Yin (2010b) for a review of the issues specific to this type of competition). This literature also takes trading fees as given (or ignores these fees). Hence it focuses on the demand for trading services taking the supply side as given. In contrast we model both the demand and the supply side by explicitly modelling the choice of their fees by trading platforms. This approach is important to analyze the efficiency of various trading arrangements in securities markets. Last our paper is also related to Degryse et al. (2010) who study the effect of clearing and settlement fees on investors' order placement strategies. Their approach is complementary since clearing and settlement fees add to the trading fee paid by investors to trading platforms.

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. In Section 5, we derive the optimal pricing policy of a matchmaker in various market

[8] There 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), Foucault and Menkveld (2008), O'Hara and Ye (2009), or Cantillon and Yin (2010a)).

[9] Foucault, Kadan and Kandel (2010) study the role of make and take fees but they do not allow investors to choose between limit and market orders as we do here. They emphasize the importance of the minimum price variation ("the tick size") for the determination of the optimal make and take fees breakdown. In contrast, the tick size is set to zero in our analysis, which explains why we find that the make/take fee breakdown is neutral (e.g., it does not affect the trading rate and investors' welfare) in contrast to Foucault, Kadan, and Kandel (2010).

[10] As pointed out by O'Hara and Ye (2011), the development of smart routing technologies have resulted in "*a single virtual market with many points of entry*," (page 14), considerably lessening the role of liquidity externalities.

structures and we compare investors' welfare in these market structures. Section 6 concludes. An appendix provides the proofs of the claims in the paper (those that do not appear in the paper for brevity are available on a companion Internet Appendix).

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 $\widetilde{T}$. Specifically, at each date $\tau = 0, 1, 2,....,$ there is a probability $(1 - \rho)$ that the asset pays its cash-flow. If date τ is not the terminal date, then a new investor arrives in the market to buy or to 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.[11] Investors with a high valuation want to buy the security whereas investors with a low valuation want to sell it. Hence, the size of the gains from trade between buyers and sellers is equal to $2L$.

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 (as, for instance, in 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.

[11] Heterogeneity 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.

Insert Figure 1 here

The Dealer Market (DM). In this market, dealers continuously post ask and bid prices denoted A^d and B^d at which they stand ready to buy or sell one share of the security. They value the security at v_0 and to process an order, they bear a cost λ. Competition among market-makers implies that they charge zero expected profits prices

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

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

When he contacts a dealer, an investor buys or sells the security at the dealer's quotes and exits the market forever.[12]Thus, when a trade takes place on the dealer market, the surplus accruing to the investor is $G^d \equiv L - \lambda$. 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 posts 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 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.[13] That is, investors with unfilled limit orders use the dealer market *in last resort*. Following the terminology used

[12] In dealer markets, commissions are in general factored into quotes. Thus, investors do not pay a fee to trade in the dealer market.

[13] If a limit order is unfilled at, say, date τ, 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 investor arrives. Hence, the only exit option for the first investor is to trade with a dealer.

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 τ, 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_τ and B_τ be the ask and bid prices posted in the limit order market at the beginning of period τ. When an investor posts a sell (resp. buy) limit order at date $\tau - 1$, then A_τ (resp. B_τ) is endogenous and will be determined below. Otherwise if there is no sell limit order in the book, we set $A_\tau = \overline{A} = +\infty$. Similarly, if there is no buy limit order in the book we set $B_\tau = \overline{B} = -\infty$.

The owner of the limit order market (the "matchmaker") collects a fee, $\overline{f} \geq 0$, each time a transaction occurs. This charge is split between the two sides (maker/taker) in the transaction as follows: the taker pays f_{ta} and the maker pays f_{ma} so that $\overline{f} = f_{ma} + f_{ta}$. Following practice, we refer to f_{ma} as the "*make fee*" and to f_{ta} as the "*take fee*". For simplicity, we set the cost of processing trades for the matchmaker to zero. We denote by

$$G^l \overset{def}{=} 2L - \overline{f},$$

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 $G^l + \overline{f} > G^d$, i.e., conditional on a trade, the limit order market is a more efficient technology to match buy and sell orders.[14]

To sum up, when they arrive in the market, investors can trade immediately, either at dealers'

[14] Studies 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.

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 taken. 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 $i \in \{H, L\}$. 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 τ. If he contacts a dealer upon arrival, he obtains a payoff

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

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

$$U^{bu}_{ta}(A_\tau, f_{ta}) \overset{def}{=} v_H - A_\tau - f_{ta}, \tag{1}$$

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

$$U^{bu}_{ma}(B, f_{ma}, \delta_i) \overset{def}{=} \delta_i \left[\phi^{bu}_\tau(B)(v_H - B - f_{ma}) + (1 - \phi^{bu}_\tau(B))G^d\right], \tag{2}$$

where $\phi^{bu}_\tau(B)$ is the execution probability of a buy limit order posted at price B at date τ, conditional on continuation of the trading game at date $\tau + 1$.[15]

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

$$U^{se}_{ta}(B_\tau, f_{ta}) \overset{def}{=} B_\tau - v_L - f_{ta}, \tag{3}$$

[15] If 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.

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

$$U^{se}_{ma}(A, f_{ma}, \delta_i) \overset{def}{=} \delta_i \left[\phi^{se}_{\tau}(A)(A - v_L - f_{ma}) + (1 - \phi^{se}_{\tau}(A))G^d\right], \tag{4}$$

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

2.2 Possible regimes for the limit order market

We focus on Markov perfect equilibria, i.e., equilibria in which investors' order placement strategies do not depend on time and the history of the market until their arrival date.[16] We now formally define the notion of Markov equilibrium in our model.

Consider a buyer arriving at date τ when the ask price in the market is A_{τ}. Let B^* be the optimal bid price of this buyer if he submits a limit order. By definition

$$B^* \in \text{Argmax}_B \; U^{bu}_{ma}(B, f_{ma}, \delta_i).$$

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^*, or (iii) a buy order in the dealer market. We denote these three options: b_m, B^* 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)). However the choice among the three possible decisions for the investor depends on the discount factor. The buyer's choice is denoted by $O_b(\delta_i, A_t) \in \{b_m, B^*, b_d\}$.

Now consider a seller who arrives at date τ when the bid price in the market is B_{τ}. We denote by A^* the optimal ask price posted by this investor if she submits a sell limit order. She has

[16] This focus is natural since investor' payoffs do not directly depend on the history of the game (see Maskin and Tirole (1997)).

three options: (i) a sell market order on the limit order market, (ii) a sell limit order at price A^* or (iii) a sell order in the dealer market. We denote these three options: s_m, A^* and s_d, and we denote the seller's choice among these options by $O_s(\delta_i, B_\tau) \in \{s_m,\ A^*,\ s_d\}$. We refer to $O_s(\cdot)$ and $O_b(\cdot)$ as the sellers and the buyers' order placement strategies.

Definition 1 *An equilibrium is a set of order placement strategies $O_s^*(\cdot)$ and $O_b^*(\cdot)$ such that (i) $O_s^*(\delta_i, B_\tau)$ 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) $O_b^*(\delta_i, A_\tau)$ 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 $O_b^*(\cdot)$.*

At date τ, the Bellman equation for a buyer's problem is

$$V_b(\delta_i, A_\tau) = Max\{v_H - A_\tau - f_{ta}, U_{ma}^{bu}(B^*, f_{ma}, \delta_i), G^d\}.$$

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

$$v_H - A^{r*}(\delta_i) - f_{ta} = Max\{U_{ma}^{bu}(B^*, f_{ma}, \delta_i), G^d\}. \tag{5}$$

The buyer optimally picks a buy market order iff $A_\tau \leq A^{r*}(\delta_i)$. That is, $A^{r*}(\delta_i)$ is the highest ask price at which the buyer is willing to submit a market order on the limit order market. We refer to $A^{r*}(\delta_i)$ as the buyer's cut-off price at date τ.

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

$$B^{r*}(\delta_i) - v_L - f_{ta} = Max\{U_{ma}^{se}(A^*, f_{ma}, \delta_i), G^d\}. \tag{6}$$

That is, $B^{r*}(\delta_i)$ 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 *Impatient investors are more likely to act as takers than makers. That is, buyers' cut-off prices decrease in δ_i and sellers' cut-off prices increase in δ_i ($A^{r*}(\delta_H) \leq A^{r*}(\delta_L)$ and $B^{r*}(\delta_H) \geq B^{r*}(\delta_L)$).*

Impatient investors are more willing to pay a concession to trade upon arrival since, other things equal, they receive a smaller expected payoff with a limit order. For this reason, the impatient buyers' (resp. sellers') cut-off price is higher (resp., lower) than the patient buyers' cut-off price. Now consider a sell limit order posted at price A. This order executes if and only if the next investor is a buyer with a cut-off price higher than A. As $A^{r*}(\delta_H) < A^{r*}(\delta_L)$, we deduce that the execution probability of the sell limit order placed at A is

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

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

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

Consequently, 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 a high "fill rate" (i.e., execution probability) limit order and the second bid as 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

$$Max\{U^{bu}_{ma}(B^{r*}(\delta_H), f_{ma}, \delta_i), U^{bu}_{ma}(B^{r*}(\delta_L), f_{ma}, \delta_i)\} < G^d. \tag{9}$$

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

$$Max\{U^{se}_{ma}(A^{r*}(\delta_H), f_{ma}, \delta_i), U^{se}_{ma}(A^{r*}(\delta_L), f_{ma}, \delta_i)\} < G^d. \tag{10}$$

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:[17]

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

[17] For 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.

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 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:

$$\textbf{C.1}\text{: } \frac{2\pi}{1-\pi}(1-\delta_L) < \delta_H - \delta_L < \frac{2\pi}{1-\pi}. \tag{11}$$

Note that this condition requires $\pi \leq \frac{1}{3}$ since $\delta_j \in (0,1]$. Under Condition C.1, each type of equilibrium can occur (see next section). In this way, our analysis covers all possible cases that can emerge in equilibrium. In contrast, for other parameter values, some equilibria do not exist. The results however still hold when Condition C.1 is not satisfied. Sometimes, for brevity, we shall refer to an equilibrium by its shorthand, e.g., a type #1 for the unspecialized/high fill rate equilibrium.

3 Equilibria

In this section, we describe the conditions under which a given type of equilibrium obtains taking the trading fees charged by the matchmaker as given. These fees are endogenized in Section 5. To describe the equilibria, let us define $\kappa_0 = 0$, $\kappa_1 = \frac{2\pi-(1-\pi)(\delta_H-\delta_L)}{2\pi+\delta_H(1+\pi)-\delta_L(1-\pi)}$, $\kappa_2 = \frac{\delta_L(1-\pi)}{2(1-\delta_L\pi)}$, $\kappa_3 = \frac{\delta_H(1-\pi)-2\pi}{2(1-2\pi-\delta_H\pi)}$, and $\kappa_4 = \frac{\delta_H}{2}$. Under Condition C.1, $\kappa_0 < \kappa_1 \leq \kappa_2 \leq \kappa_3 \leq \kappa_4$. Moreover, let us define

$$\lambda_k \equiv L(1-2\kappa_k), \tag{12}$$

and

$$\overline{f_k}(\lambda) = Max\left\{0, \frac{\lambda-\lambda_k}{\kappa_k}\right\}, \text{ for } k \in \{0,1,2,3,4\}. \tag{13}$$

Observe that $\overline{f_0}(\lambda) = 0$ and that $\overline{f}_k(\lambda)$ increases in κ_k so that $\overline{f_0}(\lambda) \leq \overline{f}_1(\lambda) \leq \overline{f}_2(\lambda) \leq \overline{f}_3(\lambda) \leq \overline{f}_4(\lambda)$.

Proposition 1 *The values of the parameters being fixed, there is a unique Markov Perfect Equilibrium. This equilibrium is of type $k \in \{1,2,3,4\}$ if and only if $\overline{f}_{k-1}(\lambda) \leq \overline{f} < \overline{f}_k(\lambda)$. Otherwise, if $\overline{f} > \overline{f}_4(\lambda)$, the dealer market crowds out the limit order market.*

1. *In a type #1 or #4 equilibrium, the bid and ask quotes posted by makers are:*

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

2. In a type #2 equilibrium, the bid and ask quotes posted by makers are:

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

3. In a type #3 equilibrium, the bid and ask quotes posted by makers are:

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

Figure 2 provides a visual representation of Proposition 1 by giving the type of equilibrium obtained for each value of the matchmaker's trading fee ($\overline{f}$) and the order processing cost in the dealer market (λ). As shown on Figure 2, when the matchmaker's trading fee is zero ($\overline{f} = 0$), an equilibrium of type k is obtained iff $\lambda_k < \lambda \leq \lambda_{k-1}$ since $\overline{f}_k(\lambda)$ increases with λ and is zero for $\lambda = \lambda_k$.

[Insert Fig.2 here]

To further gain insights on the role of the matchmaker's trading fee and the order processing cost, Table 1 gives, for specific values of the parameters, (i) the equilibrium type, (ii) the execution probability for limit orders (the fill rate), 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 computed).

[Insert Table 1 about here]

Table 1 shows that the limit orders' fill rate is non monotonic in the matchmaker's fee, $\overline{f}$, or the order processing cost in the dealer market, λ. For instance, when $\lambda = 0.7$, an increase in the trading fee charged by the limit order market from $\overline{f} = 0$ to $\overline{f} = 0.5$ leads to a decrease in limit orders' fill rate but an increase in the trading fee from $\overline{f} = 0.5$ to $\overline{f} = 1.25$ has the opposite effect. Similarly, when $\overline{f} = 0$, a decrease in the order processing cost from $\lambda = 0.7$ to $\lambda = 0.6$ leads to a decrease in limit orders' fill rate but a decrease in in the order processing cost from $\lambda = 0.6$ to $\lambda = 0.4$ has the opposite effect.

To understand this observation, notice that the maker and the taker who are part to a transaction on the limit order market share a surplus equal to $G^l = 2L - \overline{f}$. The division of this surplus between both sides (makers/takers) is determined by makers' market power, which ultimately depends on the trading fee and the order processing cost.

To see this, let us first assume that $\overline{f} = 0$ and consider the effect of a decrease in the order processing cost, λ, starting from $\lambda = L$. This decrease raises *both* the surplus that a taker can obtain by trading immediately in the dealer market and the payoff of a maker if her limit order does not execute. The first effect works to reduce makers' market power (they must post more attractive offers to attract takers) whereas the second effect works to increase makers' market power (they can post less attractive offers since their last resort option is more valuable). When λ falls below λ_1, the second effect dominates and makers choose limit orders with a lower execution probability to extract more surplus from takers. When λ falls below $\lambda_3 < \lambda_1$, the first effect dominates and makers must leave more surplus to takers. Thus, they choose offers with a higher execution probability.

Now consider the effect of an increase in the trading fee starting from $\overline{f} = 0$. This increase reduces the surplus to be shared between makers and takers. Thus, for a fixed value of λ, it raises the relative value of a trade in the dealer market both for takers and makers when their limit order does not execute. The first effect reduces makers' market power whereas the second effect increases makers' market power. The second effect dominates when $\overline{f}$ increases above $\overline{f}_1(\lambda)$ and, as a result, makers choose limit orders with lower execution probabilities. But when $\overline{f}$ increases above $\overline{f}_3(\lambda) > \overline{f}_1(\lambda)$, the first effect dominates and makers post more aggressive offers, with a higher likelihood of being accepted by takers.

Less surprisingly, impatient investors never act as makers when λ is low enough ($\lambda < \lambda_2$) or $\overline{f}$ is high enough ($\overline{f} > \overline{f}_2(\lambda)$). Indeed, in these two cases, the payoff of a limit order relative to an immediate trade in the dealer market becomes too small for impatient investors.

4 Implications

4.1 Bid-ask spreads and trading fees

Bid-ask spreads are often used as measures of takers' trading costs in securities markets. Stoll (2000) refers to the *traded spread* as the difference between the average price of trades on the ask side and the average price of trades on the bid side. In our model, this difference is $A^* - B^*$. Thus, we refer to $S = A^* - B^*$ as the *traded bid-ask spread on the limit order market.* The traded bid-ask spread does not fully reflect takers' true trading costs since it does not account for the taker fee. Thus, we also define the *cum fee bid-ask spread*, that is: $S^c \overset{def}{=} A^* - B^* + 2f_{ta}$, 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]

When the limit order market is active, the cum fee bid-ask spread is always lower or equal to the bid-ask spread on the dealer market (the two spreads are just equal only in the specialized/low fill rate equilibrium).[18] Otherwise, it would never be optimal to submit a market order on the

[18] The traded spread is $S = S^c - 2f_{ma}$. Thus, it is also smaller than the bid-ask spread in the dealer market

limit order market. Moreover the cum fee bid-ask spread is always positive. In contrast, the total fee being fixed, when the maker fee, f_{ma}, is negative and sufficiently large in absolute value, 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.

Corollary 1 *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.*

Intuitively, makers compete with dealers: both supply liquidity to newcomers in the market. For this reason, in a given equilibrium (i.e., limit orders' fill rate being fixed), makers post more aggressive offers when dealers' order processing cost decline. However, this logic is valid only for changes in the order processing cost that do not change the type of equilibrium, i.e., *small changes* in λ. For large changes, the effect of dealers' order processing cost on the bid-ask spread in the limit order market is ambiguous because, as explained in the previous section, a decrease in λ has an ambiguous effect on makers' market power. As an example, consider Table 1 again and suppose that $\overline{f} = 0$. If $\lambda = 0.7$, a type #1 equilibrium is obtained and the cum fee bid-ask spread is $S^c = 0.68$. 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 $S^c = 1.13$ and is *greater* than when $\lambda = 0.7$.

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

when $f_{ma} > 0$. In contrast, when $f_{ma} < 0$, the traded bid-ask spread can exceed the bid-ask spread in the dealer market.

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 matchmaker.*

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_{ma} and f_{ta}).*

To understand the first part of the corollary, consider first an increase in the take fee, f_{ta}. 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 indirect effect partially, but not fully, countervails 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 to reduce the bid-ask spread but this reduction is less than one-for-one, that is

$$-1 < \frac{1}{2}\frac{\partial S_p}{\partial f_{ta}} < 0. \tag{14}$$

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 $S^c = S + 2f_{ta}$, we deduce from equation (14) that

$$\frac{1}{2}\frac{\partial S^c}{\partial f_{ta}} = \frac{1}{2}\frac{\partial S}{\partial f_{ta}} + 1 > 0 \text{ and } \frac{1}{2}\frac{\partial S^c}{\partial f_{ta}} < 1. \tag{15}$$

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, the buyers' cut-off price increases and the 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 countervails 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

$$0 < \frac{1}{2}\frac{\partial S}{\partial f_{ma}} < 1. \tag{16}$$

This is also the case for the bid-ask spread cum fee since $\frac{1}{2}\frac{\partial S^c}{\partial f_{ma}} = \frac{1}{2}\frac{\partial S}{\partial f_{ma}}$. Hence, the increase in the make fee is not entirely "passed-through" by investors submitting limit orders to investors submitting market orders. Rather, 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 spreads are 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 by makers.

Overall, this analysis generates several testable implications regarding the effect of a change in make/take fees on bid-ask spreads:

Implication 1: An increase in the make fee has a positive effect on the traded bid-ask spread.

Implication 2: An increase in the take fee has a negative effect on the traded bid-ask spread.

Implication 3: For a fixed total fee, the cum fee bid-ask spread is independent of the make/take fee breakdown.

Implication 4: An increase in the total fee increases the cum fee bid-ask spread or has no effect.

Implication 5: The effect of an increase in the total fee on the traded bid-ask spread depends on whether this increase is achieved by raising the take fee or the make fee. An increase in the take fee decreases the traded bid-ask spread while an increase in the make fee increases the bid-ask spread.

Malinova and Park (2011) consider a change in the level of trading fees and the fee structure of the Toronto Stock Exchange in 2005. An interesting feature of their paper is that this change unambiguously reduced the maker fee while increasing the take fee for some stocks and decreasing it for other stocks. Consistent with our Implication 4, Malinova and Park (2011), Table 7, find that the cum fee bid-ask spread increases for stocks that experience an increase in their total fee. In contrast, the cum fee bid-ask spread decreases (insignificantly) for stocks that experience a decrease in the total trading fee. They also show (see their Table 4) that the bid-ask spread decreases significantly for stocks that experience a decrease in the make fee and an increase in the take fee, in line with Implications 1 and 2. In contrast, they do not find a significant change in the bid-ask spread in stocks where both make and take fees decrease, in line with Implication 5.

4.2 Trading Rate, Make Rate and Fill Rate

In addition to the bid-ask spread, market participants often use the trading rate and limit orders' fill rates as other measures of market liquidity. We now use Proposition 1 to study the determinants of these variables.[19]

When the limit order market is active, the investor who arrives at a given date can 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 τ in an equilibrium of type $\#k$ conditional on the asset being still traded at date τ. The likelihood of a trade on the limit order market in a given period is

$$TR^k = \varphi_2^k + \varphi_4^k. \tag{17}$$

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

$$MR^k = \varphi_1^k + \varphi_3^k. \tag{18}$$

Clearing of the limit order market requires that the number of filled limit orders be equal to the number of executed market orders. Hence, the trading rate can also be written

$$TR^k = FR^k \times MR^k, \tag{19}$$

[19] The 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.

where FR^k is the "fill-rate" 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 $\varphi_2^k + \varphi_4^k + \left(\varphi_1^k + \varphi_3^k\right)(1 - FR^k) + \varphi_5^k = 1 - TR^k$.[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*

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

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. Hence, we obtain the following testable implication.

Implication 6: The total trading fee being fixed, a change in the make-take fee breakdown does not affect the trading rate on the limit order market.

In contrast, the trading fee has an impact on the trading rate since it is one determinant of the equilibrium type. 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-intuively,

[20] The 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 $\tau - 1$. Then the limit order book is empty at date τ and if a patient investor arrives he will submit a limit order. In this case, there is no trade in either market at date τ.

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 $\overline{f} = 0$, 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 $\overline{f} = 0.5$, 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 $\overline{f} = 0.9$, 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 $18.1\% * 50\% = 9\%$.

Implication 7: For some parameter values, an increase in the trading fee on the limit order market can raise the trading rate on this market.[21]

In line with this prediction, Malinova and Park (2011) empirically find an increase in the number of transactions for stocks that experience an increase in their trading fee in the field experiment considered in their paper.

Our model suggests the following explanation for this counter-intuitive finding. Suppose that $\overline{f}$ 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 it 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

[21] As our numerical example shows, the model does <u>not</u> 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.

processing cost declines. As explained after Proposition 1, both changes reduce the makers' market power and, if large enough, can induce 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.

One concern might be that the testable implications of the model only apply when the matchmaker is a monopolist. This is not the case. In Section 5.2, we extend the model to consider the case with two competing matchmakers and we show that our implications regarding bid-ask spreads and the trading rate still hold.

4.3 Trading Fee and Investors' Welfare

The previous findings raise the intriguing possibility that an increase in the trading fee may enhance investors' welfare for some parameter values. Indeed, as we just explained, a higher trading fee can increase the likelihood that investors will directly trade together on the limit order market and direct trades between investors result in higher gains from trade as they economize on dealers' intermediation cost. A rigorous analysis of this issue however requires to account for the waiting costs borne by investors posting limit orders.

Hence, we now study how investors' expected welfare varies with the trading fee in equilibrium. We measure investors' welfare *ex-ante*, i.e., before investors learn their type (buyer/seller and patient/impatient) and choose their role (maker/taker). We focus on investors' welfare since the matchmaker obviously benefits from an increase in the trading fee when this increase also raises the trading rate on the limit order market. Less obvious is whether an increase in the trading fee can also be beneficial to investors.

Let $W(\lambda, \overline{f})$ 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 $\overline{f}$. Moreover, let $TR(\overline{f}, \lambda)$ and $FR(\overline{f}, \lambda)$ be respectively the trading rate on the platform and the fill rate for limit orders. In absence of "waiting costs" for makers (i.e., $\delta_L = \delta_H = 1$), investors' expected gains from trade would simply be the 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, weighted by the probabilities that, in each period, a transaction takes place on the limit order market ($TR(\overline{f}, \lambda)$) or the dealer market ($1 - 2TR(\overline{f}, \lambda)$). Let $W_{base}(\lambda, \overline{f}) = TR(\overline{f}, \lambda)G^l + (1 - 2TR(\overline{f}, \lambda))G^d$ be this weighted average. Computations show that[22]

$$W(\overline{f}, \lambda) = W_{base}(\lambda, \overline{f}) - \underbrace{\overline{\delta}(\overline{f}, \lambda)\Big((G^l + S^c - \overline{f})FR(\overline{f}, \lambda) + 2(1 - FR(\overline{f}, \lambda))G^d\Big)}_{\text{Waiting Costs}} \qquad (20)$$

where $\overline{\delta}(\overline{f}, \lambda) = \left((1 - \delta_H)\varphi_1(\lambda, \overline{f}) + (1 - \delta_L)\varphi_3(\lambda, \overline{f})\right)/2$ and $\varphi_1(\lambda, \overline{f})$ (resp., $\varphi_3(\lambda, \overline{f})$) 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 $\overline{f}$ are such that a type #3 equilibrium is obtained then $\varphi_1(\lambda, \overline{f}) = \varphi_1^3$, where φ_1^3 is defined in Section 4.2. Investors' welfare is lower than $W_{base}(\lambda, \overline{f})$ because makers incur a "waiting cost" since $\delta_j < 1$. Investors' welfare is independent of the make/take fee breakdown since this breakdown affects 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).

Each equilibrium type in Proposition 1 is associated with a specific role allocation for investors, i.e., a specification of (a) 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 (b) the division of surplus between the maker and the taker in each transaction

To identify the sources of inefficiencies in equilibrium, it is useful to first analyze the first best "role allocation", i.e., the actions for each type of investors and the division of surplus when two consecutive investors are matched together that maximizes investors' ex-ante expected gains

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

from trade. We assume that the central planner in charge of implementing the first best role allocation cannot force investors to trade (hence, an investor's surplus when he trades cannot be negative) and is constrained by the fact that investors have a deadline of one period (i.e., the central planner cannot batch all trades at a single point in time, say date 1). Yet, he knows investors' types so that investors do not need to be incentivized to truthfully reveal their type. Each equilibrium in Proposition 1 results in a specific role allocation but, as shown below, this role allocation never corresponds to the first best. This is not surprising. In any feasible mechanism, incentive compatibility constraints reduce the gains from trade relative to the first best.

In the first best, the trading fee is always set to zero. Indeed, this fee reduces the total surplus accruing to investors when a trade takes place and it does not influence investors' actions since these actions are assigned to them by the central planner. Using this observation, we obtain the following result.

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 when they are matched with another investor 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/high fill rate equilibrium (type #1) 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 a 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 a maker). 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 she is patient (i.e., 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 (i.e., when one wants to buy and the other wants to sell). Thus, as in high fill rate equilibria, the likelihood of a match is maximal (i.e., equal to one-half). 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 for a too large set of parameters. Indeed, for $\lambda \in [\lambda^{*}, \lambda_4]$, the limit order market is inactive in equilibrium while the first best requires this market to be opened. 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 collectively inefficient for investors since a 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 $\overline{f} \in [0, \overline{f}_3(\lambda)]$, a type #3 equilibrium is obtained while for $\overline{f} \in (\overline{f}_3(\lambda), \overline{f}_4(\lambda)]$, a type #4 equilibrium is obtained (see Figure 2). Thus, an increase in trading fee from $\overline{f} = 0$ to, say, $\overline{f} = \overline{f}_3(\lambda) + \epsilon \leq \overline{f}_4(\lambda)$ will shift the fill rate from low to high. The net effect on investors' payoff is ambiguous. On the one hand, the higher fill rate 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., $\overline{f}$ close to $\overline{f}_3(\lambda)$), 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* $\widehat{\lambda} \in (\lambda_3, \lambda_2]$ *such that for* $\lambda_3 < \lambda \leq \widehat{\lambda}$, *investors' welfare is maximal when the trading fee is* $\overline{f} = \overline{f}_3(\lambda) + \epsilon$ *(where* ϵ *is very small but positive).*

The same type of result can be obtained for $\lambda \in (\lambda_2, \lambda_1]$. Indeed, in this case, a type #2 equilibrium is obtained when the trading fee is zero and there always exists a fee high enough (e.g., $\overline{f} = \overline{f}_3(\lambda) + \epsilon$) such that a type #4 equilibrium is obtained. As the fill rate for limit orders is higher in a type #4 equilibrium, investors' welfare is higher when $\overline{f} = \overline{f}_3(\lambda) + \epsilon$ than when $\overline{f} = 0$ if λ is sufficiently close to λ_2.

[Insert Table 2 about here]

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 show 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 $\overline{f}_3(\lambda) + \epsilon$ (we set $\epsilon = 10^{-9}$). In this example, it turns out that $\widehat{\lambda} \approx 0.84$. Thus, for all values of $\lambda \in (0.802, 0.84)$, investors' welfare is strictly higher when the fee is set at $\overline{f}_3(\lambda) + \epsilon$. 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 $\overline{f} = 0.18$ (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 $\overline{f} = 0$.*

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 post offers with higher execution probabilities.

5 Is competition among trading platforms good for investors?

The results of the previous section raise the puzzling possibility that competition among trading platforms might be excessive, even from the point of view of investors since their welfare can, for some parameter values, be higher when the trading fee is strictly positive. However, which market structure maximizes investors' welfare critically depends on the trading fees that competing matchmakers would choose. Hence, in this section, we first study the trading fee that arises in equilibrium when there is a single matchmaker (Section 5.1) and when there are two matchmakers (Section 5.2). In each case, the matchmaker(s) chooses its fee before the market opens, taking into account the impact of its fee on investors' order placement strategies in equilibrium. Then, we compare investors' welfare in each market structure given the trading fees associated with each market structure (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 matchmaker's problem is

$$\text{Max}_{\overline{f}}\ \Pi(\overline{f}, \lambda) \equiv TR(\overline{f}, \lambda) \times \overline{f},$$

where $TR(\overline{f}, \lambda)$ is the trading rate on the platform if its fee is $\overline{f}$ and the order processing cost on the dealer market is λ. As explained previously, if the platform sets a fee $\overline{f} \in (\overline{f}_{k-1}(\lambda), \overline{f}_k(\lambda)]$ then a type k equilibrium is obtained and $TR(\overline{f}, \lambda) = TR^k$ (see Proposition 1 and Corollary 3). Thus, in a type k equilibrium, the platform can increase its fee up to $\overline{f}_k(\lambda)$ without changing its revenue per period. Setting a fee strictly larger than $\overline{f}_4(\lambda)$ is never optimal for the matchmaker since it results in no trading on the limit order market. Moreover, the trading rate on the limit order market is higher when the matchmaker's fee is $\overline{f}_4(\lambda)$ than when it is $\overline{f}_3(\lambda)$ (Corollary 3). Hence, the fee $\overline{f}_3(\lambda)$ cannot be optimal for the matchmaker since it generates fewer trades and a lower revenue per trade. Thus, eventually, the matchmaker optimally chooses one of three fees: $\overline{f}_1(\lambda)$, $\overline{f}_2(\lambda)$, or $\overline{f}_4(\lambda)$. In making this choice, the platform faces the traditional price-quantity trade-off for a monopolist: the larger the fee charged by the matchmaker, the smaller 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

$$\overline{f}^*(\lambda) = \begin{cases} \frac{\lambda-\lambda_1}{\kappa_1} & \text{if} \quad \max(\lambda_1^{'}, \lambda_2^{'}) \leq \lambda \leq L, \\ \frac{\lambda-\lambda_2}{\kappa_2} & \text{if} \quad \lambda_3^{'} \leq \lambda < \lambda_1^{'}, \\ \frac{\lambda-\lambda_4}{\kappa_4} & \text{if} \quad \lambda_4 \leq \lambda < \min(\lambda_2^{'}, \lambda_3^{'}). \end{cases}$$

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* $\lambda_3^{'} \leq \lambda < \lambda_1^{\prime}$*.*

Figure 4 shows the optimal fee for 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.*

As explained previously, the situation in which $\lambda = L$ is akin to the case in which the dealer market does not exist and the matchmaker has therefore full monopoly power. As expected, in this case, the matchmaker optimally charges the largest possible fee: $\overline{f}^*(L) = 2L$ (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. As the order processing cost 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 λ decreases.

Let $S^c(\overline{f}, \lambda)$ be the cum-fee bid-ask spread for given values of the trading fee and the order processing cost. Thus, when the matchmaker optimally sets its fee, the cum fee bid-ask spread is $S^c(\overline{f}^*(\lambda), \lambda)$.

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

1. *In this case, the cum fee bid-ask spread is equal to the bid-ask spread in the dealer market*

($S^c(\overline{f}^*(\lambda), \lambda) = 2\lambda$) when $\lambda \leq \max(\lambda_1^{'}, \lambda_2^{'})$ *and is 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:* $\frac{dS^c(\overline{f}^*(\lambda),\lambda)}{d\lambda} > 0$.

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 or the limit order market. This fee, that we denote $\overline{f}^{\max}$, solves:

$$L - S^c(\overline{f}^{\max}, \lambda)/2 = L - \lambda,$$

that is $S^c(\overline{f}^{\max}, \lambda) = 2\lambda$. The previous corollary shows that the monopolist matchmaker optimally chooses a fee equal to $\overline{f}^{\max}$ when $\lambda \leq \max(\lambda_1^{'}, \lambda_2^{'})$. The cost of this strategy is that it precludes a type #1 equilibrium which has the highest possible trading rate.[23] This is the reason for which there exist values of λ ($\lambda > \max(\lambda_1^{'}, \lambda_2^{'})$) for which the monopolist matchmaker optimally chooses a fee smaller than $\overline{f}^{\max}$. When this happens a type #1 equilibrium is obtained (the trading rate is maximal) and the bid-ask spread in the limit order market is strictly smaller than in the dealer market.

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 reacts to the change in order processing cost). In this case, a

[23] Indeed, in such an equilibrium, impatient investors must sometimes act as makers (submit limit orders). However, for this to be optimal, they must obtain an expected surplus with a limit order at least equal to $L - \lambda$ (as otherwise they would trade in the dealer market immediately). As a consequence patient investors obtain an expected surplus strictly larger than $L - \lambda$ in a type #1 equilibrium. As they also sometimes submit market orders in a type #1 equilibrium, the matchmaker's fee must be smaller than $\overline{f}^{\max}$ for this equilibrium to obtain.

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, $f_{ta,j}$ 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 to matchmaker 2. When indifferent, we assume that the investor chooses either market with equal probabilities. The formal definition of the equilibrium in this case and the proofs of the results in this section are given in the Internet Appendix, for brevity.

Proposition 4

- *If the matchmakers charge different total fees ($\overline{f}_1 \neq \overline{f}_2$), the limit order market with the highest total fee is inactive and the equilibrium is as described in Proposition 1 with a single limit order market charging $\overline{f} = Min\{\overline{f}_1, \overline{f}_2\}$.*
- *If the matchmakers charge the same total fees ($\overline{f}_1 = \overline{f}_2$), the two limit order markets are active if and only if $\lambda \geq \lambda_4$ and investors behave as described in Proposition 1 in equilibrium. Moreover 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* $\frac{1}{2}$.[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 ($\overline{f}_1 < \overline{f}_2$) then it attracts no trading at all, even if it subsidizes one side ($f_{ma,2} < 0$ or $f_{ta,2} < 0$). This yields the following implication.

Implication 8: The market share of a matchmaker vis à vis another matchmaker is inversely related to its total fee but it is indedepent from its make/take fee breakdown.

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. The model does not vindicate this argument. The reason is as follows. Suppose that $\overline{f}_1 = \overline{f}_2$ and that initially both markets have the same make/take fee breakdown. 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 makers' expected payoff on platform 2. But, for this reason and the fact that the take fee is higher on platform 2, takers

[24] In equilibrium, makers 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.

are less willing to trade on platform 2. Thus, as explained in Section 4.1 (see the discussion after Corollary 2), 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 makers 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.[25]

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

$$\Pi_j(\overline{f}_j, \overline{f}_{-j}; \lambda) = \begin{cases} TR(\overline{f}, \lambda) \times \overline{f} & if\ \overline{f}_j < \overline{f}_{-j}, \\ 0.5 \times TR(\overline{f}, \lambda) \times \overline{f} & if\ \overline{f}_j = \overline{f}_{-j}, \\ 0 & if\ \overline{f}_j > \overline{f}_{-j}. \end{cases} \tag{21}$$

where $TR(\overline{f}, \lambda) = TR^k$ if $\overline{f}$ and λ are such that $\overline{f}_{k-1}(\lambda) < \overline{f} \leq \overline{f}_k(\lambda)$. The next proposition provides the Nash equilibrium of the game in which the two matchmakers simultaneously choose their trading fees and obtain payoffs given by (21). 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*

[25] In addition, the ask price of, say, platform 1 may be equal or smaller than the bid price on platform 2 if the make fee on platform 1 is negative. This "locked" or "crossed" markets quotes do not constitute an arbitrage since the true cost of trading cum fees are equal in the two markets. Crossed and locked quotes do arise in reality and several commentators have linked this apparent inefficiency to the practice of subsidizing makers (see Schmerklen (2003), "Nasdaq's battle over locked crossed markets," in Wall Street Technology).

of the bid-ask spread in the dealer market. The breakdown of this fee for each matchmaker is indeterminate (i.e., any menu $(f_{ta,j}, f_{ma,j})$ such that $f_{ta,j} + f_{ma,j} = 0$ 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 $\overline{f} = 0$.

Thus, competition among matchmakers drives their total fee to zero. Hence, one would expect 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. The next corollary shows that this intuition is not always correct, however.

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

1. *When $\lambda_3 \leq \lambda < \lambda_2$, the cum fee bid-ask spread does not depend on the number of 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 and the market share of the dealer market are smaller when there are two matchmakers.*
3. *When $\lambda < \lambda_3$ or $\lambda > \lambda_1$, the cum fee bid-ask spread is smaller when there are two matchmakers and the market share of the dealer market does not depend on the number of matchmakers.*

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 the 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 is just equal to the bid-ask spread on the dealer market (Corollary 7). The 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), the 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.

Implication 9: The entry of a new matchmaker can raise the market share of the OTC market, even though it results in smaller trading fees.

The implementation of new rules (known as MiFID) in 2007 has triggered the entry of new trading platforms (Chi-X, BATS, ...) operating limit order markets for stocks listed on the main European stock exchanges (London Stock Exchange, Deutsche Börse etc...). As a result, trading fees have considerably decreased. Simultaneously, however, the market share of OTC equities market for E.U stocks has been steadily increasing since 2007, in line with our last implication.[26]

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 a 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

[26] For instance, in Europe, the market share of OTC trading in equities markets is estimated at 36% (see FESE (2011)).

structure has costs 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 Proposition 6 below, this never happens because a monopolist matchmaker always sets its fee too high.

A second question is whether an OTC 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 are 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 two matchmakers only 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 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* $\lambda \in (\lambda_4, \bar{\lambda}) \bigcup [\lambda_1, L]$*).*

Table 3 illustrates Proposition 6 for the same parameter values as in Table 2. 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 $\lambda \in]\lambda_3, \lambda_1[$.

[Insert Table 3 about here]

Consider the case in which two competing matchmakers operate in parallel with a dealer market ("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 $\overline{f}_3(0.82) = 0.18$. In this case, investors' welfare is 51% of investors' welfare in the first best, which is even higher than what is obtained with two competing matchmakers only.

This observation suggests that it might be desirable in some cases to tax investors trading in limit order markets (whether the maker or the taker pays the tax is irrelevant since the make/take fee breakdown is neutral). Indeed, competition between matchmakers is sometimes excessive in the sense that it leads to too small a fee, which destroys makers' incentives to post offers with high execution probabilities. The justification for this tax is therefore very different from that usually given by the advocates of a Tobin tax: the goal here is not to curb excessive speculation but to indirectly tilt traders' order placement choices toward a more efficient outcome.

As in Section 4.3, we have focused our attention on investors' welfare because it is obvious that the matchmaker prefers to be in a monopolistic situation. The interesting question is whether investors may also be harmed by competition among matchmakers. We have checked that if the expected profit of the matchmaker can be redistributed to investors (e.g., by taxing the matchmaker), there exist parameter values where the market structure with a single matchmaker and a dealer market is strictly dominant for all participants. Again the reason is that a high trading fee is not necessarily just a transfer from investors to the matchmaker: it also leads to a more efficient allocation of trading between the limit order market and the dealer market. The welfare gain can then be redistributed to make all parties better off.

6 Conclusion

In this paper, we show that trading fees in a limit order market are more than just transfers from investors to owners of the market. Indeed, they indirectly affect makers' market power relative to takers and as a consequence the execution probabilities chosen by investors submitting limit orders. In particular, an increase in the trading fee on a limit order market has a non monotonic effect on limit order fill rates. Actually, 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, induces them to make offers with a higher execution probability. For this reason, a decrease in trading fees (due for instance to competition among limit order markets) 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 also use the model to analyze the effect of differentiating trading fees between makers and takers. This is important since the maker-taker pricing model is very controversial in the securities industry and the economic rationale for this business model is not well understood.

Moreover, recently, the joint CFTC-SEC task force on the flash crash of May 2010 has advocated differentiating make and take fees according to market conditions: "*A peak load pricing solution to encouraging liquidity could have both access fees and rebates rise in turbulent markets. If one Exchange has a higher access fee than another, then it will get fewer aggressive liquidity demanding trades. If an exchange has a higher rebate, it will get a disproportionate share of liquidity supplying limit orders to fill out its book.*" (see Summary report of the joint CFTC-SEC Advisory Committee on Emerging Regulatory issues, p.9).[27]

In our model, for a fixed trading fee, a change in the make/take fee breakdown 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: it has no effect on traders' order placement strategies, trading volume and welfare with and without competition among matchmakers. Only the total fee matters.

We see this irrelevance result as a useful benchmark to identify conditions under which make and take fees would matter. For instance, 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 cannot fully neutralize the effect of a change in the make/take fee breakdown and this breakdown should therefore start playing a role. Foucault, Kadan and Kandel (2009) develop a theory of optimal make/take fees in this case. However, in their theory, investors cannot choose between limit and market orders. An interesting extension of our analysis would be to analyze the optimal make/take fee breakdown in presence of a minimum price variation and check whether and how this breakdown could vary with market conditions, as suggested by the joint CFTC-SEC report.

Appendix

[27] Access fees is another name for take fees and rebates here refer to negative make fees.

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

$$U^{bu}_{ma}(B^*, f_{ma}, \delta_L) = \left(\frac{\delta_L}{\delta_H}\right) U^{bu}_{ma}(B^*, f_{ma}, \delta_H). \tag{22}$$

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

$$v_H - A^{r*}(\delta_L) - f_{ta} \leq v_H - A^{r*}(\delta_H) - f_{ta}, \tag{23}$$

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

Proof of Proposition 1. First, we observe that the condition $\overline{f}_{k-1}(\lambda) < \overline{f} \leq \overline{f}_k(\lambda)$ is equivalent to $\kappa_{k-1} < \Gamma(\lambda, f^T) \leq \kappa_k$ where $\Gamma(\lambda, \overline{f}) \equiv \frac{G^d}{G^l} = \frac{L-\lambda}{2L-\overline{f}}$. Under Condition **C.1**, the set of parameters values such that $\kappa_{k-1} < \Gamma(\lambda, \overline{f}) \leq \kappa_k$ 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 type of equilibrium. 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. *Type #1 equilibrium*: Assume that $\Gamma(\lambda, \overline{f}) \leq \kappa_1$. In a type #1 equilibrium, 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 $\frac{1}{2}$. Moreover they always use the dealer market in last resort. Hence their payoff with optimal limit orders exceeds their payoff if they immediately trade in the dealer market, G^d. Using these remarks and equations (5) and (6), we deduce that

$$v_H - A^{r*}(\delta_H) - f_{ta} = \frac{\delta_H}{2}(v_H - B^{r*}(\delta_H) - f_{ma}) + \frac{\delta_H}{2}G^d,$$
$$B^{r*}(\delta_H) - v_L - f_{ta} = \frac{\delta_H}{2}(A^{r*}(\delta_H) - v_L - f_{ma}) + \frac{\delta_H}{2}G^d.$$

Solving this system of equations yields 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., $A_t = \overline{A}$). His expected utility with a limit order at $B^* = B^{r*}(\delta_H)$ is

$$U^{bu}_{ma}(B^{r*}(\delta_H), f_{ma}, \delta_L) = \frac{\delta_i}{2+\delta_H}(G^l + G^d). \tag{24}$$

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

$$\frac{G^d}{G^l} \leq \left(\frac{\delta_i}{2+\delta_H-\delta_i}\right). \tag{25}$$

Now, it can be checked that $\kappa_1 \leq \left(\frac{\delta_i}{2+\delta_H-\delta_i}\right)$. Thus, as $\Gamma(\lambda, \overline{f}) \leq \kappa_1$, Condition (25) is satisfied. The same reasoning shows that as $\Gamma(\lambda, \overline{f}) \leq \kappa_1$, 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, the impatient sellers' cut-off price solves

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

that is

$$B^{r*}(\delta_L) = v_L + f_{ta} + \frac{\delta_L}{2+\delta_H}(G^l + G^d). \tag{26}$$

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

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

Therefore, using equation (24), 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,

$$\frac{[2\pi - (1-\pi)(\delta_H - \delta_L)]}{[2\pi + (1+\pi)\delta_H - (1-\pi)\delta_L]} \geq \frac{G^d}{G^l}, \tag{27}$$

that is $\Gamma(\lambda, \overline{f}) \leq \kappa_1$. 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 $\Gamma(\lambda, \overline{f}) \leq \kappa_1$.

Type #2 equilibrium: Assume that $\kappa_1 < \Gamma(\lambda, \overline{f}) \leq \kappa_2$. 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 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

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

Solving this system of equations yields closed-form solutions for $A^{r*}(\delta_H)$ and $B^{r*}(\delta_H)$ 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., $A_t = \overline{A}$). His expected utility with a limit order at $B^* = B^{r*}(\delta_L)$ is

$$U^{bu}_{ma}(B^*, f_{ma}, \delta_i) = \frac{\delta_i}{2 + \delta_L(1-\pi)}\left((1-\pi)G^l + (1+\pi)G^d\right). \tag{28}$$

It is easily checked that $U^{bu}_{ma}(B^*, f_{ma}, \delta_i) \geq G^d$ iff $\Gamma(\lambda, \overline{f}) \leq \kappa_2$ 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

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

that is:

$$B^{r*}(\delta_H) = v_L + f_{ta} + \frac{\delta_H}{2+\delta_L(1-\pi)}\left((1-\pi)G^l + (1+\pi)G^d\right), \tag{29}$$

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

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

Therefore, using equation (28), 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,

$$\frac{[2\pi - (1-\pi)(\delta_H - \delta_L)]}{[2\pi + (1+\pi)\delta_H - (1-\pi)\delta_L]} < \frac{G^d}{G^l}, \tag{30}$$

that is $\Gamma(\lambda, \overline{f}) > \kappa_1$. 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 $\Gamma(\lambda, \overline{f}) > \kappa_1$.■

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 τ. At this date, this market can be in six 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 investors' decisions in an equilibrium of type #1, we have

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

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

$$\mathbf{d}_k(\boldsymbol{\tau}) \equiv \frac{\mathbf{q_k}(\boldsymbol{\tau})}{\mathbf{1} - \mathbf{q_{0k}}(\boldsymbol{\tau})}.$$

If $\mathbf{d}_k(\boldsymbol{\tau}+\mathbf{1}) = \mathbf{d}_k(\boldsymbol{\tau}) = \mathbf{d}_k$, 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.[28] 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^k_j is the stationary probability of state j at any date

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

conditional on the cash-flow of the security not being paid at date τ and we obtain

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

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

Proof of Proposition 2. The proof is given in the Internet Appendix for brevity.■

Proof of Corollary 4. Suppose that $\lambda \in [\lambda_3, \lambda_2)$. If $\overline{f} = 0$ then 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 $\overline{f} = \overline{f}_3(\lambda) + \epsilon$ where ϵ is very small. Then a type #4 equilibrium is obtained and investors' welfare is

$$\lim_{\epsilon \to 0^+} W(\lambda, \overline{f}_3(\lambda) + \epsilon) = \frac{(L - \lambda)}{(2 + \pi)(2\pi - \delta_H(1 - \pi))} \left(4\pi(1 - \pi) + \delta_H(7\pi^2 + \pi - 2) \right).$$

Let us define $\Delta W(\lambda) = \lim_{\epsilon \to 0^+} W(\lambda, \overline{f}_3(\lambda) + \epsilon) - W(\lambda, 0)$. Under **C.1**, $\Delta W(\lambda)$ 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 Corollary 5. Directly follows from the argument after the corollary.■

Proof of Proposition 3. As explained in the text, the optimal fee for a monopolist matchmaker belongs to $\{\overline{f}_1(\lambda), \ \overline{f}_2(\lambda), \ \overline{f}_4(\lambda)\}$. If the platform chooses a fee equal to $\overline{f}_k(\lambda)$ then a type k

equilibrium is obtained. The expected profit of the platform is then $\Pi(\overline{f}_k, \lambda) = TR_l^k \times \overline{f}_k(\lambda)$. Using the expression for $\overline{f}_k(\lambda)$ (equation (13)) and TR_l^k (Corollary 3), we obtain that

$$\Pi(\overline{f}_1(\lambda), \lambda) \geq \Pi(\overline{f}_2(\lambda), \lambda) \Leftrightarrow \frac{L-\lambda}{2L} \leq \frac{2\pi}{(3-\pi)\kappa_1^{-1} - 3(1-\pi)\kappa_2^{-1}} \Leftrightarrow \lambda \geq \lambda_1^{'},$$

$$\Pi(\overline{f}_1(\lambda), \lambda) \geq \Pi(\overline{f}_4(\lambda), \lambda) \Leftrightarrow \frac{L-\lambda}{2L} \leq \frac{2(1-\pi)}{(2+\pi)\kappa_1^{-1} - 3\pi\kappa_4^{-1}} \Leftrightarrow \lambda \geq \lambda_2^{'},$$

$$\Pi(\overline{f}_2(\lambda), \lambda) \geq \Pi(\overline{f}_4(\lambda), \lambda) \Leftrightarrow \frac{L-\lambda}{2L} \leq \frac{2(1-2\pi)}{(1-\pi)(2+\pi)\kappa_2^{-1} - \pi(3-\pi)\kappa_4^{-1}} \Leftrightarrow \lambda \geq \lambda_3^{'}.$$

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

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

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

Proof of Proposition 6. 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.

■

References

[1] Barclay, M., T. Hendershott, and T. McCormick, 2003, "Competition among Trading Venues: Information and Trading on Electronic Communication Networks," *Journal of Finance*, 58, 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., 2010a, “Competition between Exchanges: Lessons from the Battle of the Bund,” Working paper, MIT and Université Libre de Bruxelles.

[5] Cantillon, E. and Yin, P.L., 2010b “Competition between Exchanges: A Research Agenda,” forthcoming *International Journal of Industrial Organization.*

[6] CFTC-SEC, “Recommendations regarding Regulatory Responses to the Market Event of May 6, 2010.”

[7] 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.

[8] 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*, 58, 2437-2463.

[9] 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.

[10] Degryse, H., Van Achter, M., and G. Wuyts, 2010, “Internalization, Clearing and Settlement, and Stock Market Liquidity”, *mimeo*, Tilburg University.

[11] Duffie, D., Garleanu N., and Perdersen, L. 2009 “Over-the-Counter Markets,” *Econometrica*, 73, 1815-1847.

[12] Federation of European Securities Exchanges, 2011, "Response of the Federation of European Securities Exchanges to the European Commission Public Consultation on the Review of the Markets in Financial Instruments Directive" available at http://www.fese.be/.

[13] Foucault, T., Kadan, O. and Kandel, E., 2010, "Liquidity Cycles, and Make/Take Fees in Electronic Markets," Working paper, HEC, Paris.

[14] Foucault Thierry and Albert J. Menkveld, 2008, "Competition for Order Flow and Smart Order Routing Systems," *Journal of Finance*, 63, 119-158.

[15] Gehrig, T., 1993 "Intermediation in Search Markets." *Journal of Economics and Management Strategy*, 2, 97–120.

[16] Glosten, L., 1994, "Is the Electronic Order Book Inevitable", *Journal of Finance,* 49, 1127–1161.

[17] Goettler, R. L., C. A. Parlour, and U. Rajan, 2009, "Informed traders and limit order markets." *Journal of Financial Economics*, 93, 67–87.

[18] Hendershott, T. and Mendelson, H., 2000, "Crossing Networks and Dealer Markets: Competition and Performance", *Journal of Finance*, 55, 2071-2115.

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

[20] 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.

[21] Malinova, K. and A. Park, 2011, "Subsidizing liquidity: the impact of make and take fees on market quality," mimeo, University of Toronto.

[22] Maskin, E. and Tirole, J. (1997) "Markov Perfect Equilibrium: I. Observable Actions," *Journal of Economic Theory*, 191-219.

[23] O'Hara, M. and Ye, M., 2011, "Is market fragmentation harming market quality," forthcoming Journal of Financial Economics.

[24] Pagano, M., 1989, "Trading Volume and Asset Liquidity", *Quarterly Journal of Economics*, 104, 255-274.

[25] Parlour, C., and D. Seppi, 2003, "Liquidity-Based Competition for Order Flow", *Review of Financial Studies,* 16, 301-343**.**

[26] Rust, J. and G. Hall "Middlemen versus Market Makers: A Theory of Competitive Exchange," *Journal of Political Economy* 111, 353-403.

[27] Schmerken I., 2003, "Nasdaq's battle over locked crossed markets," Wall Street Technology.

[28] Spulber, D. "Market making by price-setting firms." *Review of Economics Studies* 63, 559–80.

[29] Stoll, Hans R., 2000, "Friction", *Journal of Finance*, 55, 1479-1514

[30] U.S. Securities and Exchange Commission, 2000, Release N°34-42450

[31] Yavas, A., 1992 "Marketmakers versus matchmakers," *Journal of Financial Intermediation*, 2, 33–58.

TABLES

	Matchmaker's fee: $\overline{f}$			
Order Processing Cost: λ	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 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 $\pi = 0.2$.

	Investor's Welfare		
	First Best	Equilibrium	
		Zero Trading Fee	Trading Fee: $\overline{f}_3(\lambda) + \epsilon$
Order Processing Cost: λ			
0.802	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. For each value of λ shown in the table, we give investors' ex-ante expected gains from trade in the first best (column 2), , when the trading fee is zero (column 3) and when the trading fee is sligthly above $\overline{f}_3(\lambda)$ ($\epsilon = 10^{-9}$). Other parameter values are $L = 1$, $\delta_H = 0.885$, $\delta_L = 0.067$ and $\pi = 0.297$.

	Investor's 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. For various values of λ, we give investors' ex-ante expected gains from trade in the first and in various market structures; "MMD": a monopolist matchmaker with a dealer market; "CMD": Two competing matchmakers with a dealer market; "CM": Two competing matchmakers; "D": a dealer market only (no matchmakers). A superscript "*" indicates which structure is optimal for each value of λ. Other parameter values are $L = 1$, $\delta_H = 0.885$, $\delta_L = 0.067$ and $\pi = 0.297$.

FIGURES

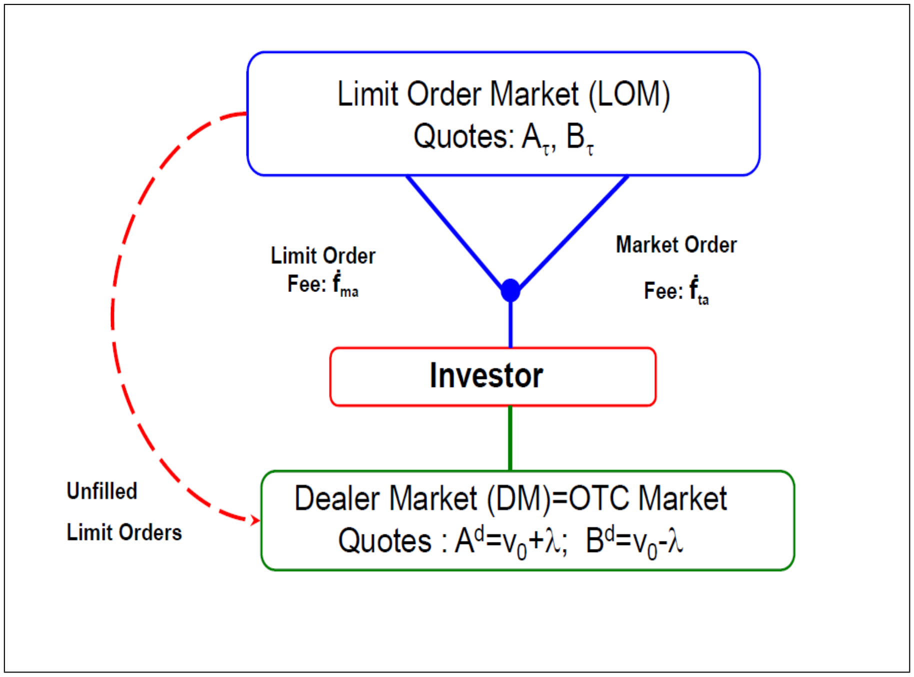

Figure 1: Trading Venues

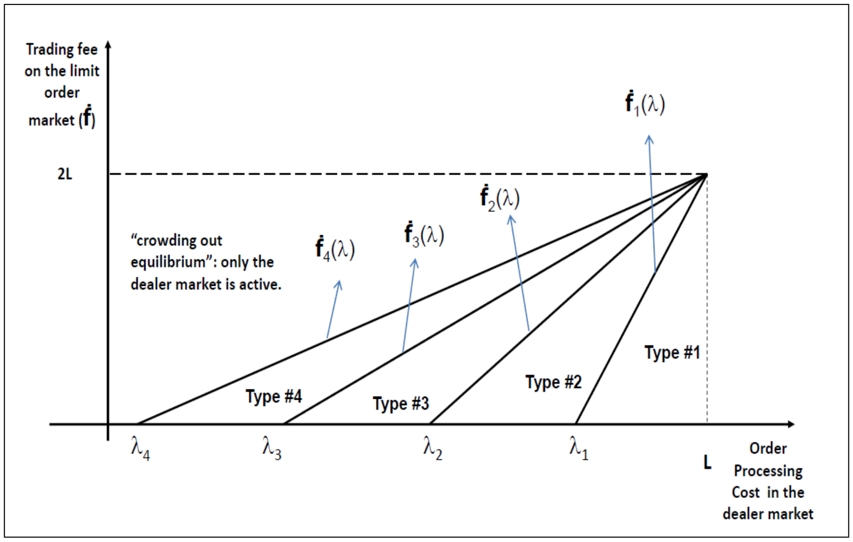

Figure 2: Equilibrium Types as a function of λ and $\overline{f}$.

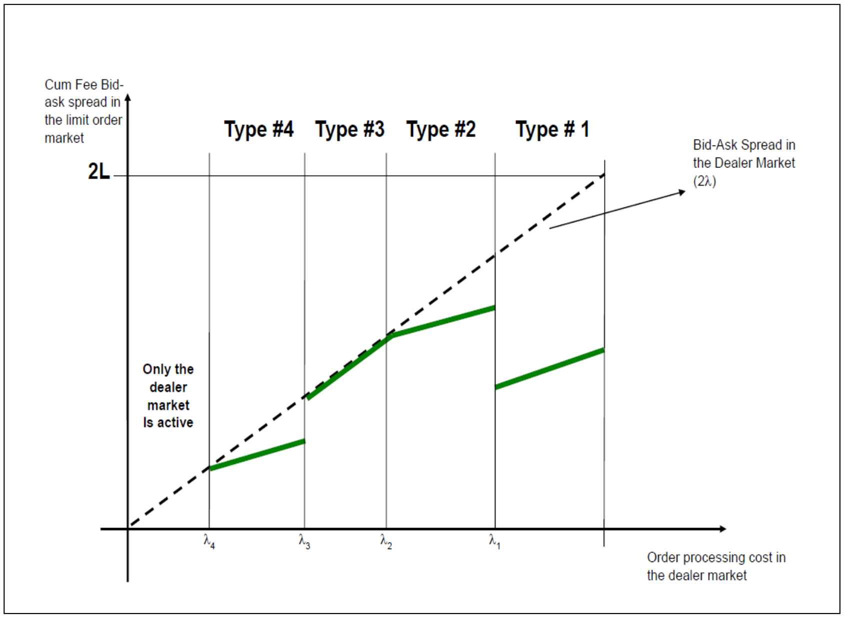

Figure 3: Cum Fee Bid-Ask Spread and Dealer Bid-ask Spread

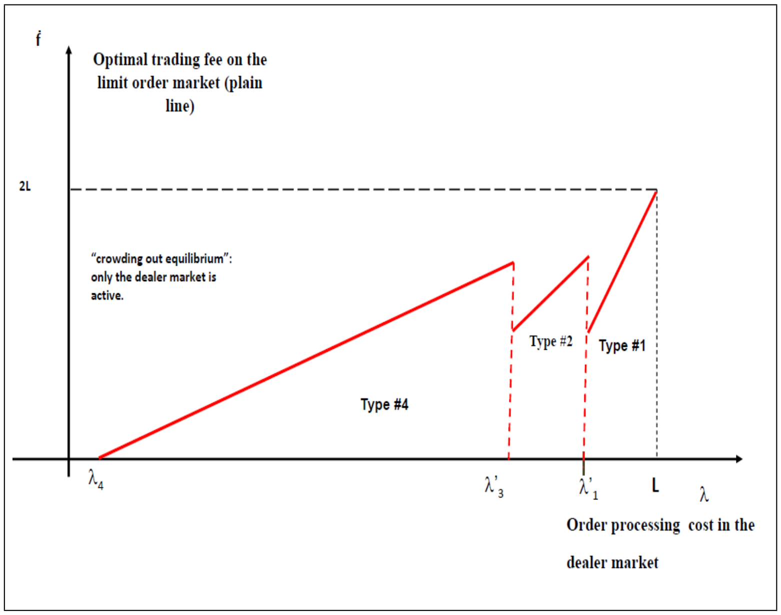

Figure 4: Optimal Trading Fee for a Monopolist Matchmaker