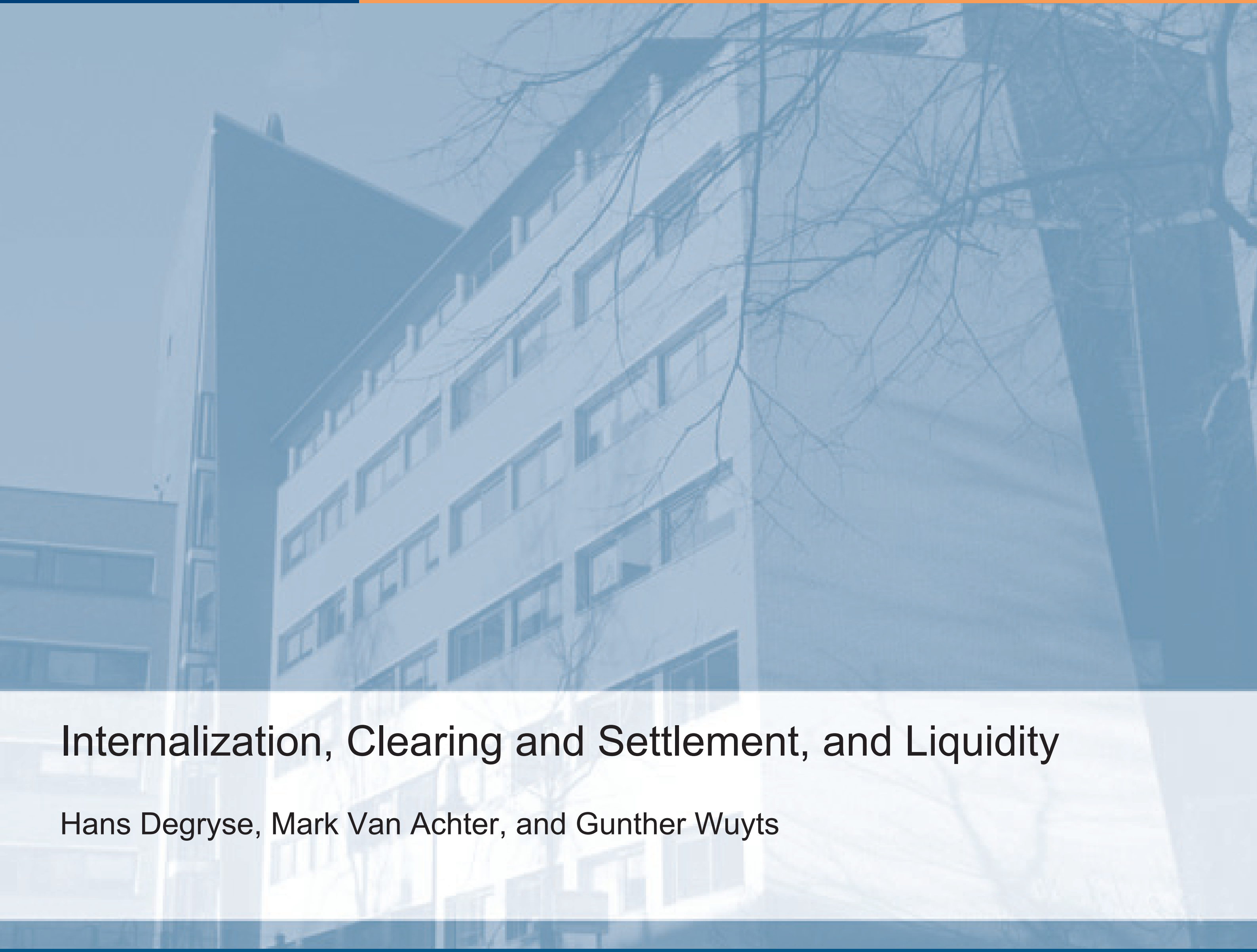
KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics



Internalization, Clearing and Settlement, and Liquidity

Hans Degryse, Mark Van Achter, and Gunther Wuyts

DEPARTMENT OF ACCOUNTANCY, FINANCE AND INSURANCE (AFI)

AFI_1262

Internalization, Clearing and Settlement, and Liquidity[1]

Hans Degryse[2], Mark Van Achter[3], and Gunther Wuyts[4]

December 2011

[1] We would like to thank Cecilia Caglio, Sarah Draus, Florian Heider, Jérémie Lefebvre, Sunny Li, Marco Pagano, Christine Parlour, Chester Spatt, Erik Theissen, Christian Voigt, Michel van der Wel, participants at the 2011 Central Bank Workshop on the Microstructure of Financial Markets (Stavanger), the 2011 European Retail Investment Conference (Stuttgart), the 2010 EFA meeting (Frankfurt), the 2010 CEPR European Summer Symposium in Financial Markets (Gerzensee), the 2010 Erasmus Liquidity Conference (Rotterdam), the 2010 Conference on the Industrial Organization of Securities Markets (Frankfurt), the 2010 AMF workshop (Paris), the 2010 AFM workshop (Amsterdam) and the 2010 KUL/UCL workshop (Brussels), as well as seminar participants at Banca d'Italia, Bank of England, BI Oslo, ECB, ESSEC Paris, NHH Bergen, Norges Bank, Sveriges Riksbank and at the Universities of Louvain, Mannheim, Milan, Naples, Stavanger and Tilburg for helpful comments and suggestions. Gunther Wuyts gratefully acknowledges financial assistance from FWO-Flanders under contract G.0567.10.

[2] Corresponding author: University of Leuven, Tilburg University and CEPR. Corresponding address: CentER - Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands. E-mail: *h.degryse@uvt.nl*.

[3] Rotterdam School of Management, Erasmus University, Department of Finance, Burgemeester Oudlaan 50, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. E-mail: *mvanachter@rsm.nl.*

[4] University of Leuven, Faculty of Business and Economics, Department of Accounting, Finance and Insurance, Naamsestraat 69, 3000 Leuven, Belgium. E-mail: *gunther.wuyts@econ.kuleuven.be*.

Abstract

We study the relation between liquidity in financial markets and post-trading fees (i.e. clearing and settlement fees). The clearing and settlement agent (CSD) faces different marginal costs for different types of transactions. Costs are lower for an internalized transaction, i.e. when buyer and seller originate from the same broker. We study two fee structures that the CSD applies to cover its costs. The first is a uniform fee on all trades (internalized and non-internalized) such that the CSD breaks even on average. Traders then maximize trading rates and higher post-trading fees increase observed liquidity in the market. The second fee structure features a CSD breaking even by charging the internalized and non-internalized trades their respective marginal cost. In this case, traders face the following trade-off: address all possible counterparties at the expense of considerable post-trading fees, or enjoy lower post-trading fees by targeting own-broker counterparties only. This difference in post-trading fees drives traders' strategies and thus liquidity. Furthermore, across the two fee structures, we find that observed liquidity may differ from cum-fee liquidity (which encompasses the post-trading fees). With trade-specific fees, the cum-fee spread depends on the interacting counterparties. Next, regulators can improve welfare by imposing a particular fee structure. The optimal fee structure hinges on the magnitude of the post-trading costs. Noteworthy, a fee structure yielding higher social welfare may in fact reduce observed liquidity. Finally, we consider a number of extensions including market power for the CSD, anonymous trading and differences in broker size.

JEL Codes: G10, G15

Keywords: transaction fees, internalization, clearing and settlement, liquidity, anonymity

1 Introduction

Trading in financial markets induces transaction costs (e.g. bid-ask spread, commissions, trading platform fees and post-trading fees), which are of considerable importance. Data from Elkins/McSherry, for example, show that explicit transaction costs constitute about three quarters of total transaction costs (see e.g. Domowitz and Steil (2002)). Further, according to a 2011 Oxera report, post-trading fees for European equities in 2009 are of equal importance as trading platform fees. While implicit transaction costs such as the bid-ask spread and market impact have been extensively studied in the finance literature[1], the impact of these explicit transaction costs on the organization of trading and market quality has largely been overlooked. Our paper makes a first step to fill this void by analyzing the impact of different post-trading fee schedules on market liquidity. As such, we investigate how the pricing of back office activities (i.e. post-trading fees) influences the front office (i.e. the organization of trading and market quality). Overall, our model features the trade-off between enhanced trading opportunities through accessing the broader market while facing considerable post-trading fees, or internalization of trading with lower execution probability combined with reduced post-trading fees. Of central importance throughout the analysis is the concept of "settlement internalization", which occurs when buyer and seller originate from the same broker or investment firm.

Our research approach is motivated by a number of recent events at the trading and the post-trading level. In the US, the Depositary Trust and Clearing Corporation (DTCC) which clears and settles trades of all exchanges observed that an increasing number of investment firms pre-netted their trades such that the order flow observed by the DTCC was not representative for the entire market. One of the actions taken by the DTCC was to reduce the post-trading fees for trades where buyer and seller originate from the same broker or investment firm (i.e. where settlement can be internalized) in order to reduce the economic incentive for using pre-netting (see e.g. DTCC (2003)). In Europe, with the implementation of the Markets in Financial Instruments Directive (MiFID), several regulated markets have introduced features allowing for settlement internalization. The London Stock Exchange for example started its SETS Internalizer in April 2007. This mechanism prevents on-book self-executions from passing through to clearing and settlement, thus minimizing post-trading fees. As a result, all order book executions where both sides of the trade originate from the same investment firm do not pass through to clearing and settlement. The tariff charged is 0.1 bp, which is 87.5% lower than the headline rate.[2] Similarly, Euronext has created an algorithm that induces buy and sell orders originating from the same investment firm to avoid post-

[1] See Madhavan (2000) and Biais, Glosten and Spatt (2005) for a survey.

[2] See page 8 on http://www.londonstockexchange.com/traders-and-brokers/rules-regulations/mifid/pre-trade.pdf

trading fees.[3] Our paper addresses how settlement internalization and the associated fee structure affects liquidity during trading in the financial market.

We model the trading phase as a limit order market. Each trader on this market is linked to a broker which leads to two potential types of trades: "internalized trades", where buyer and seller originate from the same broker, and "non-internalized trades", where buyer and seller are affiliated to different brokers. In turn, the post-trading infrastructure is modeled as a clearing and settlement agent. Throughout the paper, this entity is referred to as a Central Securities Depository (CSD), in line with the literature on clearing and settlement which is summarized later on in this introduction. We compare the impact on market quality of two different fee structures implemented by the CSD. Under the "uniform fee structure" the CSD aims to break even on average by charging the same fee for both internalized and non-internalized trades. In contrast, under the "trade-specific fee structure" the CSD breaks even by charging each transaction its individual marginal cost which implies internalized trades (which are easier to handle) are cheaper as compared to non-internalized trades. When determining their optimal order submission strategy traders take into account the (expected) post-trading fees resulting from the reigning fee structure.

Our main insights can be summarized as follows. First, explicit transaction costs such as post-trading fees affect traders' optimal order submission strategies, and thus liquidity observed in financial markets. In general, with uniform fees, traders always maximize the probability of finding a counterparty as targeting own counterparties only does not allow benefiting from lower post-trading fees. Higher post-trading fees then increase observed liquidity. The reasoning is that the resulting larger charged uniform post-trading fees lead to more aggressive limit order pricing to induce incoming counterparties to trade. This is in line with empirical evidence of Berkowitz, Logue and Noser (1988) who find that larger explicit costs decrease implicit transaction costs (be it non-commensurate). A higher degree of internalization stemming from a more concentrated broker industry reduces post-trading fees and observed liquidity. In turn, under the trade-specific fee structure, traders face a trade-off which hinges on the magnitude of the post-trading costs. With low post-trading costs (and thus low charged post-trading fees) for non-internalized trades, traders submit orders to maximize their probability of finding a counterparty taking into account that they may incur these post-trading fees. The trade-off tilts towards targeting own counterparties only when the post-trading cost (and thus the charged post-trading fee) for non-internalized trades becomes high or the broker industry becomes concentrated. Traders then prefer a higher surplus in case of execution (i.e. without post-trading fees as all trades are internalized) combined with a lower probability of execution. This shows that trading rates in the market are influenced by post-trading costs, the fee structure and the concentration in the broker industry.

[3] See page 40 on http://www.nyse.com/pdfs/NYSE_Euronext_%20Analyst_Presentation.pdf

Second, liquidity as observed in the market may differ from cum-fee liquidity (i.e. the liquidity at which a market order trades after adding the post-trading fees). With uniform fees, the cum-fee spread increases in post-trading fees. Interestingly, when the CSD implements the trade-specific fee schedule, the cum-fee spread hinges on the match between the two counterparties. The observed spread equals the cum-fee spread when counterparties of the same broker meet. In contrast, for non-internalized trades both buyer and seller incur the post-trading fee, implying the cum-fee spread is larger than the observed spread. Third, regulators can improve market liquidity by imposing a fee structure on the CSD. The liquidity-optimizing fee structure depends on the gains from trading, the marginal costs for the CSD and the concentration in the broker industry. Fourth, we perform a welfare analysis comparing the different settings employing an overall welfare measure (thus capturing all market participants). For high post-trading costs, the trade-specific fee structure maximizes welfare as it features only internalized trades. Within this cost range, the uniform fee structure induces a higher trading rate but also a large fraction of costly non-internalized trades. For intermediate levels of post-trading costs, a uniform fee structure is preferred by the social planner as it maximizes the trading rate and non-internalized trades contribute to welfare. In contrast, the trade-specific fee structure would induce limit order traders to submit orders targeting own-broker counterparties only. As a consequence, welfare-creating non-internalized trades would not take place. For low levels of post-trading costs, both fee structures yield the same welfare. Finally, our welfare results highlight an important trade-off for the social planner. For high post-trading costs, maximum observed liquidity is achieved under the uniform fee structure. However, uniform fees produce lowest welfare in this range of costs. In turn, for very low post-trading costs, while the uniform and trade-specific fee structures yield the same welfare, observed liquidity is higher under the trade-specific fee structure. As a consequence, a social planner potentially has to choose between liquidity and social welfare when setting its regulation for a fee structure to be implemented by the CSD: a fee structure implying higher market liquidity in fact may reduce social welfare.

We also investigate a number of extensions to our model. First, we analyze a CSD having pricing power in setting the fee for non-internalized trades. The CSD then optimally sets this fee such that traders continue to maximize trading opportunities. Our findings indicate that observed liquidity increases compared to perfect competition because counterparties need to be compensated for the higher post-trading fee. This result further demonstrates that imperfect competition (market power) at the post-trading phase has an effect on liquidity during the trading phase. Overall welfare is not affected compared to the main model, but there is a redistribution from traders to the CSD. In a second extension we assume an arriving trader cannot observe the identity of the counterparty such that she cannot perfectly infer whether a trade would be internalized or not. In addition, we study an “in-between” setting where traders can choose whether or not to reveal their identity. We show that quotes and observed liquidity are different under each setting. Moreover, from a overall welfare perspective, a social planner always prefers the “in-between” setting over the full anonymity and the full transparency setting. In a third and last extension we allow brokers to be of different size. When the CSD imposes trade-specific fees, bid and ask quotes of traders then hinge upon their broker affiliation. As a result, traders from small brokers may submit different quotes than traders from large brokers. Markets could then become more or less liquid at points in time, depending on which group of traders (from the large or small broker) is active at that point.

To our knowledge no papers exist linking the organization of the post-trading infrastructure to stock market liquidity. Taking a wider perspective, our paper is related to different sets of literature.

First, it relates to the literature on order submission strategies in limit order markets such as Foucault (1999), Parlour (1998), Handa, Schwartz and Tiwari (2003), Foucault, Kadan and Kandel (2005), Goettler, Parlour and Rajan (2005), Roşu (2009) and Van Achter (2009). These papers model how traders choose between market orders and limit orders in different dynamic settings. We extend them by including the impact of heterogeneity in post-trading fees on the optimal quote setting behavior of traders linked to different brokers. Our paper also relates to the literature on make/take fees as modeled in Foucault, Kadan and Kandel (2011) and Colliard and Foucault (2011), and empirically analyzed in Malinova and Park (2011). In many markets, providers of liquidity receive a “make fee”, whereas consumers of liquidity pay a “take fee”. Foucault, Kadan and Kandel (2011) show this may induce liquidity cycles to arise, while Colliard and Foucault (2011) analyze how inter-market competition affects these make/take fees and ultimately trader behavior and liquidity. Our paper contributes to this literature by highlighting that outstanding quotes by one broker in the limit order book may induce asymmetries for traders affiliated to different brokers. When the transaction is internalized and implies no post-trading cost, the post-trading fee is low and it is as if the payable take fee is small. In contrast, when a trader of another broker is the counterparty, post-trading fees are high and it is as if the payable take fee is large. As such, our model generates different trading rates (i.e. targeting all counterparties or own-broker counterparties only) which stem from differences in post-trading fees and not from traders exhibiting different valuations for stocks and different degrees of impatience (as in Colliard and Foucault (2011)).

Second, our work contributes to the literature on clearing and settlement. Existing theoretical papers therein mostly deal with the optimal pricing strategies when central securities depositories (CSDs) interact, in order to explain the high markups for cross-border transfers of securities or the effects of different degrees of access to the CSDs (see e.g. Rochet (2005), Tapking and Yang (2006), Holthausen and Tapking (2007), Tapking (2007), and Koeppl, Monnet and Temzelides (2012)). We model how a cost-based post-trading infrastructure may affect liquidity in financial markets in two different ways. First, internalization of order flow reduces fees payable to the CSD and therefore changes the traders' aggressiveness in the stock market. Second, the way a cost-based fee structure is implemented by the CSD may lead to different stock market equilibria. In particular, a pricing strategy fully reflecting the CSD's marginal cost may lead to an equilibrium where traders opt to only address counterparties from the same broker. This reduces the total number of transactions and decreases liquidity. Further, the empirical papers on the post-trading infrastructure mainly investigate whether there are economies of scale and scope in the clearing and settlement industry (see e.g. Van Cayseele and Wuyts (2008)). Our paper shows that on average transactions may exhibit different degrees of difficulty (i.e. easier internalized clearing and settlement versus more difficult cross-broker clearing and settlement), hinging on the particular stock market equilibrium that is played.

Third, a few papers connect different phases of the trading process. Foucault and Parlour (2004) model how competition between stock exchanges links listing fees and transaction costs on those exchanges. They find that competing exchanges relax competition by choosing different trading technologies and listing fees. Ellul and Pagano (2006) link the IPO stage with trading in the after-market and show that IPO underpricing is larger when the after-market is expected to be less liquid. Our paper also links two phases of the trading cycle, i.e. trading and post-trading.

The remainder of this paper is structured as follows. Section 2 introduces the setup of our model. Section 3 presents two different fee structures implemented by the clearing and settlement agent, and the corresponding equilibria. Within Section 4, these equilibria are further analyzed and compared with respect to liquidity and trading rates, and a welfare analysis is provided. Section 5 presents a number of extensions to our main model. Finally, Section 6 concludes.

2 Setup

We develop an infinite horizon model to analyze a continuous limit order market listing a single security. Before trading starts, the clearing and settlement agent (from now on denoted as CSD) sets the fees of clearing and settlement. Traders take these post-trading fees as given during the subsequent trading game. Each period in time $t = 0, 1, ... + \infty$, a single trader arrives who is willing to trade one share of the asset. Traders are risk neutral, aim to maximize expected utility, and exhibit an exogenously determined trading orientation which makes them either a buyer or a seller. We assume that the likelihood of arrival of a buyer and a seller is identical.[4] Buyers have a private valuation for the asset equal to V_h, whereas sellers have a private valuation V_l. We assume both valuations are non-negative and $V_h - V_l > 0$, which implies there are always gains from trade between both parties in the absence of post-trading fees. These differences in valuation are an outcome of taxes, liquidity shocks, or other portfolio considerations such as differences in endowment, or in opinions on the expected value of the asset.[5] Each trader is linked to one of N possible brokers which means their individual orders are sent to the market through this particular broker. Brokers do not have any other role in our model. We assume $N \geq 2$ and that every broker has an equal share of affiliated traders (in an extension developed in Subsection 5.3 we consider a setup with two brokers where one broker is large and the other broker is small, allowing us to deal with brokers of different sizes). As such, $1/N$ can also be interpreted as a measure of concentration of the broker industry: the larger this fraction becomes, the more concentrated is the broker industry. Broker affiliations are indexed by superscript $j \in \{1, ..., N\}$. Hence, for a trader arriving in a random period t, with probability $1/(2N)$ it is or a buyer or a seller from broker j. In our base case setting, we assume that transparency holds such that broker affiliations are observable to traders (in an extension developed within Subsection 5.2 we relax this assumption). Since $N \geq 2$, two types of trades can then occur: internalized trades and non-internalized trades. We define internalized trades as trades where both buyer and seller are affiliated to the same broker. Non-internalized trades are then trades where buyer and seller stem from a different broker.

The CSD handles clearing and settlement immediately after each transaction, is risk neutral and has no fixed costs of operating. However, it has a marginal cost c per leg of the trade for non-internalized trades (i.e. more complex trades involving different brokers), and a lower marginal cost for internalized trades (i.e. trades involving the same broker) which we normalize to zero. In implementing its fee structure, the CSD aims to break even on average, but does not necessarily charge its true marginal cost on each individual transaction. Overall, depending on the sophistication of the fee structure, a CSD can charge different fees based on the type of transaction and thus differentiate between internalized and non-internalized trades. To properly account for this distinction, we consider two different fee structures: uniform and trade-specific. The uniform fee structure means that the CSD charges the same fee to internalized and non-internalized trades. Thus, the uniform fee is set such that the CSD breaks even on average. In turn, the trade-specific fee structure entails a CSD breaking even by charging internalized and non-internalized trades a fee equal to their respective individual marginal cost. Denote

[4] Our model is easily adjusted for the case where the likelihood of buyers and sellers arriving is different from 0.5; however it becomes slightly more complex since buyers and sellers no longer choose symmetric strategies. We prefer equal probabilities as this allows to more easily identify the impact of different fee structures implemented by the clearing and settlement agent.

[5] See Duffie et al. (2005) for further economic interpretations.

the post-trading fee by c^i, with superscript $i \in \{I, NI\}$ indicating the fee charged for internalized (I) and non-internalized (NI) trades.[6] The two different fee structures are then summarized as follows:

Fee Structure CSD	Uniform	$c^I = c^{NI}$
	Trade-Specific	$c^I < c^{NI}$

Both fee structures and their respective equilibrium fees and impact on quotes will be further analyzed in Section 3.

An arriving trader bases her order submission strategy on her observation of the standing limit order book (LOB). She may have two possibilities at her disposal to trade. On the one hand, she could post a quote by submitting a limit order (LO) which does not offer certainty of execution. Posted LOs stay in the market for one period and are thus take-or-leave offers for the next trader (see Foucault (1999) for a similar approach). On the other hand, she could submit a market order (MO) which guarantees immediate execution but at a less favorable execution price. Liquidity-demanding MOs execute against standing liquidity-supplying LOs, so they can only be submitted if a counterparty LO is already present in the LOB. Clearly, the LO's execution probability is endogenous in the model as it depends on other traders' order placement strategies. We will further discuss this issue at the end of this section. Orders are for one unit of the asset, and once submitted cannot be modified or cancelled. New in our model and a key contribution to the existing literature (such as Foucault (1999), Handa, Schwartz and Tiwari (2003), Van Achter (2009), and Colliard and Foucault (2011)) is that traders also account for the fee structure implemented by the CSD in choosing their optimal strategy. More specifically, conditional upon execution, the utility of trading the asset at price P for a buyer equals $V_h - P - c^I$ if the buyer and seller are of the same broker and $V_h - P - c^{NI}$ if the buyer and seller are affiliated to a different broker. A seller's utility is $P - V_l - c^I$when seller and buyer stem from the same broker and $P - V_l - c^{NI}$ when seller and buyer are of different brokers. Hence, as non-trading gains are normalized to zero, the fee-adjusted reservation price that buyers are willing to pay and that sellers are willing to receive for one share of the asset hinges on the trader's broker affiliation, the counterparty's broker affiliation and the implemented fee structure. Put differently, it is as if the "transaction tax" (i.e. the fee) on a particular trade hinges not only on whether there is a trade but also on whether the counterparty's broker is identical to the trader's one or not. This influences traders' strategies.

Traders aim to maximize the expected payoff of their order and therefore also need to account for its execution probability. In setting the optimal bid or ask quotes when submitting a LO, a trader in general has two possibilities. She could determine quotes

[6] We do not make a distinction between different fees for different brokers as brokers are identical in our setup.

that only attract counterparties from her own broker (we label this strategy "*own*") or she can opt for a quote that is attractive to all possible counterparties, i.e. traders from her own broker and from all other brokers (we label this strategy "*all*"). Do note that "attract" in this context means the targeted incoming trader is at least willing to hit the standing LO by submitting a MO. Thus, any trader submitting a LO needs to account for the MO strategy of the subsequently arriving trader.[7] Given that traders are linked to a broker j two possible strategies can be distinguished:

1. traders of broker j only aim to address counterparties of their own broker j: "*own*";
2. traders of broker j aim to address counterparties of all brokers: "*all*".

It can easily be shown that it is never optimal to target only a subset of other brokers. All parameters of the model, including V_h, V_l, and c^i are known to the investors and are constant over time. This allows to solve for a stationary equilibrium within each fee structure as in Foucault (1999), Van Achter (2009), or Colliard and Foucault (2011). More specifically, a stationary market equilibrium is defined as a set of mutual order submission strategies (specifying an optimal order type, quote and corresponding execution probability to each possible state of the LOB) such that each trader's strategy is optimal given the strategies of all other traders. Different types of stationary equilibria arise. The magnitude of the post-trading fees, the implemented fee schedule as well as the type of equilibrium influence stock market liquidity. In Section 3 we discuss the stationary equilibria corresponding to the different fee structures.

3 CSD Fee Structures

3.1 Uniform Fee Structure

Under the uniform fee structure, which is denoted by subscript U, the CSD charges a uniform fee to all orders upon execution and breaks even on average over all transactions.[8] Thus, it compensates the losses it makes on the complex (i.e. non-internalized) order flow stemming from different brokers with gains from the easy order flow stemming from trades that occur within the same broker (i.e. internalized). Denote this

[7] As such, the LO execution probabilities are endogenous, implying traders are in a game situation. In general, traders' optimal order submission strategies depend on their LO's probability of execution, which in turn is determined by their order submission strategies. To properly account for these endogenous linkages between the MO and the LO placement strategies, they will be determined simultaneously.

[8] Recall also that all transactions have to be cleared and settled through the CSD. Therefore, we now assume that a broker cannot set up its own clearing and settlement system to internalize trades between its own traders. In case it would, the equilibrium uniform fee would be c and the setup would coincide with the trade-specific fee structure as analyzed in Subsection 3.2.

break-even fee by c_U, this fee structure then implies that:

$$c^I = c^{NI} = c_U$$

Under this fee structure, it is clear that traders of each broker will always address all potential counterparties, and not restrict themselves to own-broker counterparties only. Put in other words: the "*all*" strategy dominates the "*own*" strategy. The reason is that as all traders face a uniform fee, it is impossible to set a quote only attractive to traders of one particular broker.[9] Therefore, when analyzing the equilibrium we only consider the "*all*" strategy.

We now turn to the determination of the equilibrium quotes and the optimal fee under a uniform CSD fee structure. We solve the model backwards. First, for a given fee c_U we derive traders' order placement strategies in equilibrium. Second, we solve for the optimal fee c_U^*. In determining its fee, the CSD rationally anticipates how this fee affects traders' order submission behavior.

How do traders set their quotes, taking c_U as given? Given that the "*all*" strategy prevails and that costs and gains are identical for traders of all brokers, it must hold that bid and ask quotes set by traders of all brokers are identical. We denote this as follows:

$$A^j_{U,all} = A_{U,all} \quad \text{and} \quad B^j_{U,all} = B_{U,all}, \quad \forall j$$

where $A^j_{U,all}$ refers to the ask quote (A) set by a trader from broker j (superscript j) with uniform fees by the CSD (subscript U) and under the "*all*" sub-equilibrium (subscript *all*) which prevails here. The bid quote has a similar notation.

Suppose now a buyer arrives in the market. She will set the bid quote of her LO such that the next incoming seller is indifferent between hitting the LO (by submitting a sell MO) or submitting a sell LO herself. This implies that the expected payoff for the incoming seller of submitting a MO or a LO must be the same. The following equation shows this indifference condition:

$$B_{U,all} - V_l - c_U = \frac{1}{2}\left[A_{U,all} - V_l - c_U\right]$$

The left hand side of this equation presents the gain from a sell MO, given the bid quote set by the buyer in the previous period. The right hand side is the expected gain of a sell LO, which is the execution probability of this order (i.e. 1/2 or the probability that the next arriving trader is a buyer who will hit the standing sell LO since the seller optimally also sets her ask quote to make the next arriving buyer indifferent) multiplied by the payoff upon execution of her order corrected for the appropriate post-trading fee.

[9] Do note that if playing the "*own*"-strategy would be possible, this would still be a sub-optimal strategy as it only reduces execution probabilities without inducing any quote advantage.

Thus, the idea here is that $B_{U,all}$ is chosen at the lowest level at which the subsequently arriving seller is just willing to submit a MO, while both are accounting for the post-trading fee c_U. In other words, $B_{U,all}$ equals the seller's cutoff price and renders this seller indifferent between hitting the standing LO at $B_{U,all}$ and submitting her own LO at $A_{U,all}$. Submitting a LO at all other quotes is easily proven to be sub-optimal for this buyer.

Similarly an arriving seller sets her LO quote in order to make a subsequently arriving buyer indifferent between submitting a buy MO at $A_{U,all}$ or a buy LO at $B_{U,all}$:

$$V_h - A_{U,all} - c_U = \frac{1}{2}\left[V_h - B_{U,all} - c_U\right]$$

Solving the system of indifference equations yields the quotes for a given uniform fee c_U. Proposition 1 presents the optimal fee and resulting equilibrium quotes and bid-ask spread ($S_{U,all}$) for the uniform fee structure.

Proposition 1 *When the CSD applies a uniform fee, the optimal fee announced by the CSD is:*

$$c_U^* = \left(\frac{N-1}{N}\right)c$$

Traders always play the "all" sub-equilibrium. The optimal quotes and resulting spread are:[10]

$$\begin{aligned} A_{U,all}^* &= \frac{2V_h + V_l}{3} - \frac{(N-1)\,c}{3N} \\ B_{U,all}^* &= \frac{V_h + 2V_l}{3} + \frac{(N-1)\,c}{3N} \\ S_{U,all}^* &= \frac{V_h - V_l}{3} - \frac{2\,(N-1)}{3N}c \end{aligned}$$

Proof. See Appendix A. ■

Under the "*all*" strategy which is played within this fee structure, the fee c_U^* at which the CSD breaks even over all transactions is equal to $\left(\frac{N-1}{N}\right)c$. Intuitively, this expression can be seen to capture the costly non-internalized match between counterparties from the $N-1$ other brokers and each broker j. By charging c_U^* on both legs of every transaction (internalized and non-internalized), the CSD on average breaks even: it gains on internalized trades for which it does not face marginal costs and loses on non-internalized trades as active clearing and settlement takes place. By charging a uniform fee on all transactions, the CSD removes the advantage for the trader of being matched with a

[10] This computed spread is actually never observed in the market at a single point in time. We merely use it as a proxy for average liquidity during trading. Further, the computed spread is only negative when $V_h - V_l < 2\frac{N-1}{N}c$. That is, when the potential trading gains of a transaction are fully annihilated by the charged post-trading fee $\left(\frac{N-1}{N}\right)c$. Evidently, under these circumstances, there will be no trading at all.

counterparty of the own broker and the disadvantage of being matched with a counterparty of another broker. Furthermore, we observe that, for $N \geq 2$, the ask decreases in c, while the bid increases in c. Consequently, the resulting spread decreases. Thus, all other things equal, a larger c leading to a higher uniform fee $(\left(\frac{N-1}{N}\right) c)$ induces more liquid quote-setting behavior and improves stock market liquidity. The reasoning behind this remarkable result is that traders submit more aggressive LOs in order to induce the counterparty to submit a MO (which incurs the post-trading fee with certainty). That is, it is as if the counterparty now has a lower willingness to trade resulting from the increase in the post-trading fee. Moreover, holding c constant, when the broker industry concentration $1/N$ increases due to a lower number of brokers N, the spread increases (i.e. the ask increases and the bid drops). The reasoning is that a lower N leads to a smaller fraction of non-internalized trades and as a result a lower uniform fee. This in turn induces a less aggressive pricing strategy.

Next, we distinguish between "observed liquidity" and "cum-fee liquidity". Definition 1 provides the definition of both concepts.

Definition 1 ***Observed liquidity*** *corresponds to the bid-ask spread or the quotes as observable in the market.* ***Cum-fee liquidity*** *corresponds to the bid-ask spread or the quotes from the point of view of the market order trader after adding the relevant post-trading fees.*

When the concept "liquidity" is mentioned without any further detail, this always corresponds to observed liquidity. When referring to cum-fee liquidity, this is always stated explicitly.

The cum-fee ask and bid quote for a market order trader are

$$A^*_{U,all} + c^*_U \text{ and } B^*_{U,all} - c^*_U \tag{1}$$

and thus the cum-fee spread becomes

$$S^*_{U,all} + 2c^*_U. \tag{2}$$

Clearly, cum-fee liquidity is lower than observed liquidity. Moreover, and in contrast to observed liquidity, cum-fee liquidity decreases in c and increases in broker industry concentration $1/N$. There is a less than complete pass-through of c^*_U as the observed liquidity partially absorbs increases of c^*_U. Corollary 1 summarizes the empirical implications of the equilibrium under uniform CSD fees.

Corollary 1 *Under uniform fees by the CSD and ceteris paribus:*

- *Observed liquidity increases with the post-trading fee* c_U^**;*
- *Observed liquidity decreases with broker industry concentration* $1/N$*;*
- *Cum-fee liquidity is lower than observed liquidity;*
- *Cum-fee liquidity decreases with the post-trading fee* c_U^**;*
- *Cum-fee liquidity increases with broker industry concentration* $1/N$*.*

3.2 Trade-Specific Fee Structure

Under the trade-specific fee structure, denoted by subscript TS, we assume the CSD breaks even by pricing according to the marginal costs that are associated with individual transactions. That is, post-trading fees are set to zero for internalized trades, and amount to c for non-internalized trades. As argued before, note that the zero cost attributed to internalized trades merely represents a normalization. More generally, as long as internalized trades imply lower marginal costs than non-internalized trades, all results obtained below hold. In terms of the notation introduced in Section 2, we have:

$$\begin{aligned} c^I &= 0 \\ c^{NI} &= c \end{aligned}$$

A novel implication of these different fees is that the quoting behavior of traders linked to a specific broker j may be to target only traders affiliated to the same broker j (the "*own*" strategy), or to target traders of all brokers (the "*all*" strategy). Consider the following example to illustrate this point. Assume a buyer linked to broker j arrives in the market. On the one hand, she could submit a LO. Her quote choice allows her to choose which counterparties she wants to address: (i) by posting a lower bid, she only attracts counterparties from the same broker implying a higher payoff with a lower execution probability, whereas (ii) by posting a higher bid, she also attracts counterparties from the other brokers implying a lower payoff with a higher execution probability. Do note $B^j_{TS,own}$ is the lowest bid quote at which an incoming seller from the same broker is willing to submit a MO (while accounting for the according zero post-trading fee and her own LO strategy quoting $A^j_{TS,own}$, and given that traders from another broker play an "*own*" strategy). In turn, $B^j_{TS,all}$ is the lowest bid quote at which an incoming seller from another broker $j' \neq j$ is willing to submit a MO (while accounting for the according higher post-trading fee c and her own LO strategy quoting $A^j_{TS,all}$, and given that traders from broker j' play an "*all*" strategy). Submitting a LO at any other quote is easily proven to be sub-optimal for this buyer.[11] On the other hand, given the

[11] That is, higher bid quotes do not increase the execution probability yielding lower expected payoffs.

availability of a standing sell LO which is attractive to her, she could also submit a MO. A buyer affiliated to broker j' faces a similar trade-off. Further, as the proportion of buyers and sellers in the trader population is equal, the actions of sellers linked to all brokers are completely symmetric, and is derived in a similar way. As we will see below, the choice between these quotes hinges on market parameters such as the number of brokers with its associated execution probability, and the post-trading fee. For both "*all*" and "*own*" strategy, we will now determine the according equilibrium quotes set by traders at both brokers. The fee structure of the CSD (i.e. zero fee for internalized trades, c for non-internalized trades), is again taken as given by the traders.

Starting with the "*all*" strategy, traders at each broker j set their quote to keep the marginal other-broker trader indifferent as they want to address all traders.[12] Thus, they account for the post-trading fee c. So for buyers and sellers from each broker j, we respectively have:

$$\begin{aligned} B^{j}_{TS,all} - V_l - c &= \frac{1}{2}\left[A^{j'\neq j}_{TS,all} - V_l - \frac{(N-1)}{N}c\right] \\ V_h - A^{j}_{TS,all} - c &= \frac{1}{2}\left[V_h - B^{j'\neq j}_{TS,all} - \frac{(N-1)}{N}c\right] \end{aligned}$$

Thus, within the first indifference condition for instance, the incoming seller from any other broker $j' \neq j$ is kept indifferent between hitting the standing quote $B^{j}_{TS,all}$ by submitting a sell MO (accounting for the appropriate post-trading fee c) and submitting her own sell LO (of which the execution probability, the quote and the post-trading fee c correctly correspond to the "*all*" strategy this seller is playing herself). Similar indifference conditions apply for traders from all other brokers $j' \neq j$. As all brokers are symmetric, all bid and ask quotes will be identical, i.e. $A^{j'\neq j}_{TS,all} = A^{j}_{TS,all}$ and $B^{j'\neq j}_{TS,all} = B^{j}_{TS,all}$.

Next, within the "*own*" strategy, all traders only keep potential counterparties of their own broker indifferent. Hence, all trades are internalized and thus incur a zero post-trading fee. The indifference equations for buyer and seller from any broker then become:

$$\begin{aligned} B^{j}_{TS,own} - V_l &= \frac{1}{2N}\left[A^{j}_{TS,own} - V_l\right] \\ V_h - A^{j}_{TS,own} &= \frac{1}{2N}\left[V_h - B^{j}_{TS,own}\right] \end{aligned}$$

Thus, within the first indifference condition for instance, the incoming seller from broker j is kept indifferent between hitting the standing quote $B^{j}_{TS,own}$ by submitting a sell MO (accounting for the appropriate zero post-trading fee) and submitting her own sell LO

[12] Evidently, traders from the same broker always accept this quote as they face no fees.

(of which the execution probability, the quote and the zero post-trading fee correctly correspond to the "*own*" strategy this seller is playing herself). At these quotes, only traders from broker j are indifferent. For traders originating from other brokers $j' \neq j$ trading at these quotes is too costly given their higher post-trading fee c. Therefore, the execution probabilities are only related to those of the own broker j. Similar equations apply for all N brokers.

Solving the above systems of indifference conditions renders the equilibrium quotes and thus the two distinct "*all*" and "*own*" sub-equilibria. Comparing expected payoffs for each of the sub-equilibria, we are able to determine when each of the sub-equilibria is valid. All these elements are shown in the equilibrium presented in Proposition 2.

Proposition 2 *With a CSD applying trade-specific (marginal cost-based) fees, traders at every broker play the following LO strategies hinging on the value of c:*

- *For low values of c, i.e.* $c \leq \frac{2N(N-1)(V_h - V_l)}{(2N+1)(2N-1)}$*, traders from every broker target counterparties of all brokers, thus the "all" sub-equilibrium is played. The equilibrium quotes and resulting spread are:*[13]

$$\begin{aligned} A^*_{TS,all} &= \frac{2V_h + V_l}{3} - \frac{(N+1)}{3N}c \\ B^*_{TS,all} &= \frac{V_h + 2V_l}{3} + \frac{(N+1)}{3N}c \\ S^*_{TS,all} &= \frac{V_h - V_l}{3} - \frac{2\,(N+1)}{3N}c \end{aligned}$$

- *For high values of c, i.e.* $c > \frac{2N(N-1)(V_h - V_l)}{(2N+1)(2N-1)}$*, traders from each broker only target own counterparties, thus the "own" sub-equilibrium is played. The equilibrium quotes and resulting spread are:*[14]

$$\begin{aligned} A^*_{TS,own} &= \frac{2V_h N + V_l}{2N+1} \\ B^*_{TS,own} &= \frac{V_h + 2V_l N}{2N+1} \\ S^*_{TS,own} &= \frac{2\,(N-1)}{2N+1}\,(V_h - V_l) \end{aligned}$$

Proof. See Appendix A. ■

First, for low values of c the "*all*" sub-equilibrium holds and traders at each broker target counterparties at all brokers by quoting relatively liquid prices. The quotes depend on the fee c for non-internalized trades. Holding N fixed, we find that observed quotes

[13] This computed spread is never negative under the "*all*" sub-equilibrium as $c \leq \frac{2N(N-1)(V_h - V_l)}{(2N+1)(2N-1)}$ and $\frac{V_h - V_l}{3} - 2\frac{N+1}{3N}c < 0$ can never be jointly satisfied.

[14] Do note that this computed spread is always positive for $N \geq 2$.

(and thus also the resulting spread computed from these quotes) become more liquid if c increases. Hence, observed market liquidity improves with higher post-trading fees within this c range. Next, at given c it can easily be seen that the quotes become more illiquid when N becomes larger. The reasoning is as follows. If a trader places a limit order, with probability $1/(2N)$ this order will be hit by a trader from her own broker. In that case, both traders receive a "bonus" as they both do not have to pay c. If the number of brokers N is large (i.e. a less concentrated broker industry) the probability of this bonus is small making the threat of the arriving trader to submit a limit order herself less present. Thus, quotes are set at a less liquid level.

Second, within the other sub-equilibrium holding for higher c, the "*own*" strategy is played and relatively illiquid prices are quoted. As all trades are internalized (implying zero post-trading fees are charged), evidently quoted prices are independent of c. In turn, the liquidity of the observed quotes increases with broker industry concentration $1/N$: the outside option of the next trader (submitting a LO instead of hitting a standing LO order) becomes more attractive since - under the "*own*" strategy - execution probabilities increase with lower N.

Furthermore, from Proposition 2, it also follows directly that $S^*_{TS,all} < S^*_{TS,own}$ (given that $N \geq 2$). In other words, the observed spread under the "*all*" strategy is always smaller - or liquidity is higher - than the observed spread under the "*own*" strategy.

Next, observed liquidity can again be contrasted with cum-fee liquidity. Compared to the uniform fee structure equilibrium, however, we now no longer have one but two cases: the "*all*" and the "*own*" sub-equilibrium. Within the "*own*" sub-equilibrium, evidently observed quotes are identical to cum-fee quotes (as all trades are internalized implying zero post-trading fees are charged), and thus all results on increases in c and N derived for the observable quotes continue to hold. In contrast, within the "*all*" sub-equilibrium, traders stemming from different brokers may trade resulting in the following cum-fee ask and bid quotes, respectively: $A^*_{TS,all} + c$ and $B^*_{TS,all} - c$. These quotes imply a cum-fee spread of $S^*_{TS,all} + 2c$. Transacting buyer and seller may, however, also be affiliated to the same broker. The trade is then internalized and cum-fee ask and bid quotes are $A^*_{TS,all} + 0$ and $B^*_{TS,all} - 0$, respectively. These quotes imply a cum-fee spread equal to the (observed) spread $S^*_{TS,all}$. Do note that these two possibilities within the "*all*" sub-equilibrium point to a key difference between post-trading fees and make/take fees as modeled in Colliard and Foucault (2011). Make/take fees are charged to everyone in each trade. In contrast, in our setting post-trading fees are trade-dependent: non-internalized trades are charged a high fee c, while internalized trades are charged a zero fee. The size of c may lead to the "*all*" sub-equilibrium or the "*own*" sub-equilibrium being played, implying distinct trading intensities. Weighing each cum-fee quote by its average rate of occurrence within the "*all*" sub-equilibrium allows us to compute the average cum-fee ask and bid quotes

$$\begin{aligned} &\frac{N-1}{N}\left(A^*_{TS,all}+c\right)+\frac{1}{N}\left(A^*_{TS,all}+0\right) \\ &\frac{N-1}{N}\left(B^*_{TS,all}-c\right)+\frac{1}{N}\left(B^*_{TS,all}+0\right) \end{aligned} \tag{3}$$

and the average cum-fee spread

$$S^*_{TS,all}+2\frac{(N-1)}{N}c \tag{4}$$

Clearly, within the "*all*" sub-equilibrium, cum-fee liquidity is lower than liquidity observed in the market. Moreover, in contrast to observed liquidity, cum-fee liquidity decreases in post-trading fees (for $N > 2$) and decreases with broker industry concentration. Corollary 2 summarizes the main empirical implications of the trade-specific fee structure.

Corollary 2 *Under trade-specific fees by the CSD and ceteris paribus:*

- *Observed liquidity increases (under the "all" sub-equilibrium) or remains constant (under the "own" sub-equilibrium) with the post-trading fee for non-internalized trades c;*
- *Observed liquidity increases with broker industry concentration $1/N$;*
- *Observed liquidity in the "all" sub-equilibrium is higher than in the "own" sub-equilibrium;*
- *Cum-fee liquidity is lower than (under the "all" sub-equilibrium) or equal to (under the "own" sub-equilibrium) observed liquidity;*
- *Cum-fee liquidity decreases (under the "all" sub-equilibrium) or remains constant (under the "own" sub-equilibrium) with the post-trading fee for non-internalized trades c;*
- *Cum-fee liquidity decreases (under the "all" sub-equilibrium) or increases (under the "own" sub-equilibrium) with broker industry concentration $1/N$.*

4 Market Quality and Welfare: Comparison of CSD Fee Structures

In the first subsection we compare the implications of the different fee structures set by the CSD on market quality. We do so by investigating liquidity as measured by the bid-ask spread, and by focusing on trading volume. In a next subsection we discuss the impact of CSD fees on welfare. To highlight our main points, we illustrate the results of our model using the following parameter values: $V_h = 20$, $V_l = 0$, $N = 5$ and c varies in the interval $[0, 20]$. Important to stress is that all results are general, and do not depend on these specific parameter values and ranges.

4.1 Market Quality

4.1.1 Liquidity: Bid-Ask Spreads

We start with observed liquidity as derived from quotes and spreads observable in the markets. Using the spreads presented in Propositions 1 and 2, it is easy to show that when the "*all*" sub-equilibrium under trade-specific fees holds, spreads are lower for the trade-specific fee structure. The reasoning is that the arriving market order trader that is kept indifferent incurs a higher post-trading fee under the trade-specific fee structure than under the uniform fee structure. In contrast, when the "*own*" sub-equilibrium under trade-specific fees holds, the spread is more liquid under uniform fees. Thus, which CSD fee structure maximizes market liquidity depends on the parameters driving the cutoff value between the "*own*" and the "*all*" sub-equilibrium: V_h, V_l, c and N. This result brings us to Corollary 3.

Corollary 3 *Regulators can improve market liquidity by imposing a fee structure on the CSD. The liquidity-optimizing fee structure depends on the trading gains level $V_h - V_l$, the magnitude of c and the concentration in the broker industry $1/N$.*

Next, we consider cum-fee liquidity as measured by the cum-fee spread, i.e. the observed bid-ask spread adjusted for post-trading fees. Cum-fee spreads were computed in Equation (2) for uniform fees and Equation (4) for trade-specific fees. From these equations, it could be derived that cum-fee and observed liquidity may provide opposite results. For low c, when the "*all*" sub-equilibrium is played under trade-specific fees, uniform fees lead to lower liquidity, both observed and cum-fee. However, when c is sufficiently high, and we are already in the "*own*" sub-equilibrium under trade-specific fees, observed liquidity is higher under uniform fees while cum-fee liquidity is higher under trade-specific fees.

The above-mentioned results are illustrated in Figure 1, where in Panel A we depict the observed bid-ask spreads as a function of c for our two CSD fee structures. The line with stars "$*$" represents spreads for the uniform fee structure (computed as in Proposition 1), and the lines with squares "$\Box$" indicates spreads corresponding to the two sub-equilibria within the trade-specific fee structure (computed as in Proposition 2). In turn, Panel B shows average cum-fee bid-ask spreads, computed using Equations (2) and (4). Note that if c becomes extremely high (i.e. if c is higher than the gains that can be obtained from trading) the market shuts down.

Please insert Figure 1 around here.

4.1.2 Trading Volume

In the previous subsection, we focused on liquidity as measured by bid-ask spreads. Now, we turn to trading volume, another measure for market liquidity often used in the literature and by practitioners. In doing so, we follow a similar approach as Colliard and Foucault (2011) by focusing on trading rates per period. In each period, we observe one out of the following three possible states (i.e. actions): (1) a trader submits a limit order; (2) a trader submits a market order that is internalized (i.e. a trader hits a standing limit order submitted by another trader of the same broker); (3) a trader submits a market order that is not internalized (i.e. a trader hits a limit order submitted by a trader from another broker). We do not need to make a distinction between buyers and sellers because both sides of the market are symmetric in our model. For each fee structure of the CSD $k \in \{U, TS\}$, denote the stationary probabilities of each of these three possible states under a given sub-equilibrium $s \in S_k$ as $\varphi_{k,s} = \left\{\varphi^1_{k,s}, \varphi^2_{k,s}, \varphi^3_{k,s}\right\}$. S_k denotes the set of all possible sub-equilibria under fee structure k. Hence, for the uniform fee structure $S_U = \{all\}$, while under trade-specific fees $S_{TS} = \{all, own\}$. In Appendix B, we derive the exact expressions for the various $\varphi_{k,s}$. Based on these stationary probabilities, we can now easily develop measures for trading volume (i.e. the trading rate) and the degree of internalization (i.e. the internalization rate) for each sub-equilibrium within each fee structure.

The trading rate TR in a period under sub-equilibrium s of fee structure k is the likelihood of a market order initiating a trade in a given period. Clearly, this occurs in states 2 and 3 mentioned above, thus:

$$TR_{k,s} = \varphi^2_{k,s} + \varphi^3_{k,s}$$

In turn, the internalization rate is the likelihood of a market order initiating an internalized trade (possibility 2) divided by the trading rate :

$$IR_{k,s} = \frac{\varphi^2_{k,s}}{\varphi^2_{k,s} + \varphi^3_{k,s}}$$

and can be seen as the percentage of trades that is internalized. The non-internalization rate is then equal to $1 - IR_{k,s}$.

Proposition 3 presents the trading rates and internalization rates for the different fee structures.

Proposition 3 *Trading rates and internalization rates for the different CSD fee structures are as follows:*

- *Under uniform fees:*

$$\begin{aligned} TR^*_{U,all} &= \frac{1}{3} \\ IR^*_{U,all} &= \frac{1}{N} \end{aligned}$$

- *Under trade-specific fees for the different sub-equilibria:*

$$\begin{aligned} TR^*_{TS,all} &= \frac{1}{3} \\ TR^*_{TS,own} &= \frac{1}{2N+1} \\ IR^*_{TS,all} &= \frac{1}{N} \\ IR^*_{TS,own} &= 1 \end{aligned}$$

Proof. See Appendix A. ■

Since $N \geq 2$, we obtain that $TR^*_{U,all} = TR^*_{TS,all} \geq TR^*_{TS,own}$ implying that trading volume is highest when the CSD applies uniform fees, or when c is small such that the "*all*" sub-equilibrium applies under trade-specific fees. For the internalization rates, we have that $IR^*_{U,all} = IR^*_{TS,all} \leq IR^*_{U,own}$. The latter rate is obviously equal to one since in the "*own*" sub-equilibrium all trades are internalized. Further, do note that trading volume is not directly related to market liquidity as measured by quote aggressiveness. This can be seen most easily from the fact that $TR^*_{U,all} = TR^*_{TS,all}$, while the aggressiveness of ask quotes for the according cases as derived in the previous subsection is different. In the next subsection, we use these trading and internalization rates to derive welfare implications of the different fee structures.

4.2 Overall Welfare

In this subsection, we characterize welfare for the two CSD fee structures. Our welfare measure builds on rational trader behavior and is therefore identical to the "mean" realized ex post welfare per period. We focus on overall welfare (OW), i.e. the sum of all agents' expected utilities from trading (see Glosten (1998), Goettler, Parlour and Rajan (2005), Hollifield, Miller, Sandås and Slive (2006), and Degryse, Van Achter and Wuyts (2009) for a similar approach in quantifying welfare). As the CSD always breaks even in expected terms, in our model OW coincides with trader welfare.

Welfare is realized when a trade occurs. An internalized trade generates $V_h - V_l$, whereas a non-internalized trade produces $V_h - V_l - 2c$. Thus, when non-internalized trades occur in equilibrium, an increase in c reduces the surplus to be split between buyer and seller involved in the transaction. In turn, the prices at which trades occur are not relevant for welfare as they only represent a redistribution between buyer and seller. Expected overall welfare per period within fee structure k and sub-equilibrium s is computed by multiplying the trading rate $TR_{k,s}$ with the welfare realized in the occurrence of a trade (appropriately weighing internalized and non-internalized trades):

$$OW_{k,s} = TR_{k,s}\left[V_h - V_l - 2c\left(1 - IR_{k,s}\right)\right].$$

Proposition 4 summarizes our main results regarding overall per-period welfare for both CSD fee structures.

Proposition 4 *Expected overall welfare per period depends on the CSD fee structure.*

- *Under uniform fees, it equals:*

$$OW^{*}_{U,all} = \frac{1}{3}\left[V_h - V_l - 2c\left(\frac{N-1}{N}\right)\right]$$

- *Under trade-specific fees, it hinges on the sub-equilibria:*
  - *For low values of c, i.e. $c \leq \frac{2N(N-1)(V_h-V_l)}{(2N+1)(2N-1)}$ or the "all" sub-equilibrium, it equals:*

$$OW^{*}_{TS,all} = \frac{1}{3}\left[V_h - V_l - 2c\left(\frac{N-1}{N}\right)\right]$$

  - *For high values of c, i.e. $c > \frac{2N(N-1)(V_h-V_l)}{(2N+1)(2N-1)}$, or the "own" sub-equilibrium, it equals:*

$$OW^{*}_{TS,own} = \frac{1}{2N+1}\left[V_h - V_l\right]$$

Proof. See Appendix A. ■

Next, we determine which fee structure a social planner prefers depending upon the magnitude of c.[15] In fact, the social planner faces the following trade-off. On the one hand, it wants to maximize the trading rate as this increases trading gains. On the other hand, it prefers internalized trades above non-internalized trades as the former do not generate post-trading costs. Therefore it also cares about the internalization rate. Corollary 4 presents welfare rankings for the entire range of c.

Corollary 4 *Expected per-period overall welfare ranking for the entire range of c.*

- *For low values of c, i.e. $c \leq \frac{N(V_h-V_l)}{2N+1}$, we find that $OW^{*}_{U,all} = OW^{*}_{TS,all} \geq OW^{*}_{TS,own}$. Only the uniform fee structure achieves the "all" sub-equilibrium for the entire range of $c \leq \frac{N(V_h-V_l)}{2N+1}$, the trade-specific fee structure achieves the "all" sub-equilibrium for $c \leq \frac{2N(N-1)(V_h-V_l)}{(2N+1)(2N-1)}$;*

[15] We assume the social planner can only impose the fee structure it prefers and does not intervene in the trading and post-trading phase.

- *For high values of c, i.e.* $c > \frac{N(V_h - V_l)}{2N+1}$*, we find that* $OW^*_{TS,own} > OW^*_{TS,all} = OW^*_{U,all}$*. Only the trade-specific fee structure achieves the "own" sub-equilibrium for* $c > \frac{N(V_h - V_l)}{2N+1}$*.*

Hence when the cost of a non-internalized trade is low (i.e. when $c \leq \frac{N(V_h - V_l)}{2N+1}$), the social planner will choose to maximize the trading rate through the "*all*" sub-equilibrium. It could do so by imposing uniform fees. Imposing trade-specific fees only yields the socially optimal "*all*" sub-equilibrium for $c \leq \frac{2N(N-1)(V_h - V_l)}{(2N+1)(2N-1)}$. However, when the cost of a non-internalized transaction becomes too large (i.e. when $c > \frac{N(V_h - V_l)}{2N+1}$), the social planner wants to prevent expensive non-internalized trades from occurring as these are welfare-reducing and prefers the "*own*" sub-equilibrium, thus aiming to maximize the internalization rate. Only the trade-specific fee structure succeeds in producing this outcome. Furthermore, with extremely high values of c (or extremely low gains from trade), trade-specific fees allow to create a market for internalized trades only, whereas markets would collapse under uniform fees since this yields zero profits and welfare.

We illustrate Proposition 4 and Corollary 4 graphically in Figure 2 using the same parameter values as before. The "$*$" line represents overall welfare for the uniform fee structure, and the "□" lines for the trade-specific fee structure.

Please insert Figure 2 around here.

Our welfare results show that maximizing welfare may conflict with maximizing liquidity. Recall from Corollary 3 that the maximum observed liquidity for high values of c is achieved under the uniform fee structure. However, uniform fees produces lowest welfare in this range of c. As a consequence, a social planner potentially has to choose between liquidity and social welfare when setting its regulation for a fee structure to be implemented by the CSD: a fee structure implying higher market liquidity in fact reduces social welfare. Moreover, for very low c, both fee structures yield the same welfare, although observed liquidity differs under each fee structure. This leads to the following result in Corollary 5.

Corollary 5 *The observed bid-ask spread is not always an appropriate measure for welfare.*

5 Extensions

In the main model, we have made a number of assumptions regarding the market power of the CSD, the transparency of the trader's identity, and the individual broker size. In this section, we discuss three extensions relaxing each of these assumptions. As such, we investigate how these extensions affect our main results obtained above, and delineate the yielded additional insights.

5.1 Extension 1: Market Power for CSD

Thus far, we assumed that the CSD, when implementing its fee structure, aims to break even on average. In other words, the CSD always behaves perfectly competitive. In this subsection, we alter this assumption and allow the CSD to possess market power in setting the fee it charges to its customers. More specifically, we assume that the CSD has monopoly power when setting the fee charged to non-internalized trades, but has no pricing power for internalized trades. This assumption is realistic since brokers can always set-up an own clearing and settlement system for internalized trades if the CSD does not charge a competitive fee for this type of trade. For non-internalized trades, establishing a clearing and settlement system is much more difficult, and therefore the CSD may have pricing power for such trades. Given that the CSD differentiates its fees between types of trades, we only analyze the trade-specific fee structure.

The CSD will now set a fee of $c^m_{TS,s}$ for non-internalized trades (as opposed to c under perfect competition) and a zero fee for internalized trades. $c^m_{TS,s}$ denotes the CSD fee under monopoly m, trade-specific fee structure TS and sub-equilibrium $s \in S_{TS}$ with S_{TS} the set of all possible sub-equilibria under trade-specific fees: $S_{TS} = \{all, own\}$. For all relevant variables in this subsection, the superscript m thus refers to the model featuring monopoly power for the CSD. The CSD wants to set $c^m_{TS,s}$ to maximize expected profits per period. In doing so, it faces a trade-off: by setting $c^m_{TS,s}$ too high, the CSD may alter the equilibrium played by traders. For instance, if it sets $c^m_{TS,s}$ very high, the consequence will be that the "*own*" sub-equilibrium prevails, driving expected profits of the CSD to zero since it has no pricing power for internalized trades. Therefore, the CSD aims to maximize expected profits by setting $c^m_{TS,s}$ at a maximum, within the "*all*" sub-equilibrium. Otherwise the CSD has zero expected profits as the "*own*" strategy is played. Thus, the maximization problem of the CSD is:

$$\max_{c^m_{TS,s}} 2TR_{k,s}\left(1 - IR_{k,s}\right)\left(c^m_{TS,s} - c\right) \tag{5}$$

$$s.t.\ V_h - V_l \geq 2c^m_{TS,s}$$

In the objective function $c^m_{TS,s} - c$ represents the CSD's mark-up above the marginal cost of the trade, $TR_{k,s}$ is the trading rate and $1 - IR_{k,s}$ is the proportion of non-internalized trades. We multiply by two because the CSD charges the fee (and thus obtains the profit) on each leg of the trade. The constraint implies that the monopoly fee can only be as high as the total gains of trade.

Using the trading rates and the internalization rates from Proposition 3, expected per-period profits of the CSD under the different sub-equilibria are as given in Proposition 5.

Proposition 5 *If a CSD applies trade-specific fees and has monopoly pricing power on non-internalized trades, traders play the following strategies hinging on c:*

- *For low values of c, i.e.* $c \leq \frac{2N(N-1)(V_h-V_l)}{(2N+1)(2N-1)}$*, traders play the "all" sub-equilibrium. The optimal fee for non-internalized trades set by the CSD is*

$$c^{m,*}_{TS,all} = \frac{2N(N-1)\left(V_h - V_l\right)}{\left(2N+1\right)\left(2N-1\right)}$$

and the expected per-period profit of the CSD is

$$\frac{1}{3}\left[\frac{N-1}{N}\right] 2\left(c^{m,*}_{TS,all} - c\right).$$

The equilibrium quotes and spread are:

$$\begin{aligned} A^{m,*}_{TS,all} &= \frac{2V_h + V_l}{3} - \left(\frac{N+1}{3N}\right) c^{m,*}_{TS,all} \\ B^{m,*}_{TS,all} &= \frac{V_h + 2V_l}{3} + \left(\frac{N+1}{3N}\right) c^{m,*}_{TS,all} \\ S^{m,*}_{TS,all} &= \frac{V_h - V_l}{\left(2N+1\right)\left(2N-1\right)} \end{aligned}$$

- *For high values of c, i.e.* $c > \frac{2N(N-1)(V_h-V_l)}{(2N+1)(2N-1)}$*, traders of every broker only target own counterparties, thus the "own" sub-equilibrium is played. The equilibrium quotes and spread are:*

$$\begin{aligned} A^{m,*}_{TS,own} &= \frac{2V_hN + V_l}{2N+1} \\ B^{m,*}_{TS,own} &= \frac{V_h + 2V_lN}{2N+1} \\ S^{m,*}_{TS,own} &= \frac{\left(2N-1\right)\left(V_h - V_l\right)}{\left(2N+1\right)} \end{aligned}$$

The per-period profit of the CSD is zero.

Proof. See Appendix A. ■

Next, we compare this setting with the perfect competition case of our main model. In fact, granting monopoly power to the CSD does not influence the equilibrium played. The only difference compared to the perfect competition case is that the CSD charges a fee such that all extra rents from the "*all*" sub-equilibrium compared to the "*own*" sub-equilibrium are expropriated from the trader and now flow to the CSD. More specifically, the CSD optimally charges $c^{m,*}_{TS,all}$: at this fee, the traders are indifferent between the "*all*" strategy and the "*own*" strategy. Charging a higher fee is not optimal as the traders then will adopt the "*own*" strategy. When c is larger than the cut-off value $\frac{2N(N-1)(V_h-V_l)}{(2N+1)(2N-1)}$, the traders play "*own*" and the CSD can never make a profit. Furthermore, in comparison with the competitive case, the equilibrium quotes in the "*all*" sub-equilibrium are different, and in general more aggressive. The reason is that counterparties need to be compensated for the higher post-trading fees. These results demonstrate that imperfect competition (market power) at the post trading phase has an effect on market liquidity during the trading phase.

As a final step in this extension, we investigate the welfare implications of having a monopolist CSD. Notice that overall welfare is not affected in the monopoly setting. Indeed, there is only a redistribution of welfare among market participants. While in the competitive case, all welfare is realized by traders, under monopoly the CSD skims some of the welfare from the traders. Proposition 6 provides the precise welfare distribution between the CSD and traders.

Proposition 6 *If a CSD applies trade-specific fees and has monopoly pricing power on non-internalized trades, expected per-period welfare realized by the CSD (CSDW) and by traders (TW) under the "all" and "own" case are:*

$$\begin{aligned} CSDW^{m,*}_{TS,all} &= \frac{1}{3}\left[\left(\frac{N-1}{N}\right)\right] 2\left(c^{m,*}_{TS,all} - c\right) \\ CSDW^{m,*}_{TS,own} &= 0 \end{aligned}$$

and

$$\begin{aligned} TW^{m,*}_{TS,all} &= \frac{1}{3}\left[V_h - V_l - 2\left(\frac{N-1}{N}\right) c^{m,*}_{TS,all}\right] \\ TW^{m,*}_{TS,own} &= \frac{1}{2N+1}\left[V_h - V_l\right] \end{aligned}$$

with the optimal fee that the monopolist CSD charges equal to:

$$c^{m,*}_{TS,all} = \frac{2N(N-1)\left(V_h - V_l\right)}{(2N+1)\left(2N-1\right)}$$

Proof. See Appendix A. ■

From Proposition 6 it is easily seen that CSD welfare is weakly increasing in N: the less concentrated the broker industry under the "*all*" sub-equilibrium, the higher the expected profits the CSD obtains. In contrast, under the "*own*" sub-equilibrium, CSD welfare is unaffected by N. Trader welfare on the other hand is decreasing in N in both sub-equilibria.

5.2 Extension 2: Anonymous Trading

In a second extension, we relax the assumption of transparency in the trading process (while again assuming a perfectly competitive CSD).[16] In particular, we assume that an arriving trader cannot observe the identity behind the counterparty's quote. Consequently, a trader cannot observe whether her counterparty stems from the same broker and thus whether a trade would be internalized or not. We implicitly assume that if the identity of a trader is observed, so is the identity of the trader's broker (which is the relevant feature in our model). Motivated by real-world financial markets, we distinguish between two anonymity settings. The first is a setting where the trader has no means to reveal her identity and trading is completely anonymous. The second is a case where the trader can choose to reveal her identity by attaching a so-called "flag" to her limit order. This corresponds to common practice in some markets where limit order submitters have the option to reveal their identity. For instance, since 2005 the Toronto Stock Exchange offers "attribution choices" to its traders/members on an order-by-order basis.[17] Attributing a limit order entails a unique broker identifier visible to all market participants that is attached to this order. Comerton-Forde, Putnins and Tang (2011) provide an empirical analysis of this setting. In general, do note our anonymity extension is only relevant under a trade-specific fee structure, since under uniform fees, no distinction in fees exists between internalized and non-internalized trades.

Under "transparency", the trader's identity is always revealed. The quotes in this equilibrium have been shown in Proposition 2. In contrast, with "full anonymity", there are no means to reveal the identity of the trader. Therefore, an arriving trader does not know whether a standing limit order stems from a trader of her own broker or from another broker.[18] Consequently, she is uncertain about the CSD fees she will incur when submitting a market order. Therefore, the market order trader will account for expected post-trading fees $\left(\frac{N-1}{N}\right) c$, as shown in the proof of Proposition 7. Thus, the full anonymity setting is shown to perfectly coincide with the uniform fee case shown in Proposition 1. Next, we focus on the setting where traders have the choice to reveal their identity using a flag, which we label "anonymity with flag". In fact, the equilibrium played is fully determined by investors submitting limit orders, rather than by market order traders. The reasoning is that limit orders are submitted before market orders.[19] First, consider the "*all*" strategy. Comparing the respective payoffs of "using a flag" versus "not using a flag", we show in the proof of Proposition 7 that limit order traders always opt not to reveal their identity under the "*all*" strategy as quotes are more liquid

[16] For a discussion of anonymity, see e.g. Foucault, Moinas and Theissen (2007) or Rindi (2008).

[17] See http://www.tmx.com/en/listings/newsletters/article_5.html for some descriptive information.

[18] Evidently, the arriving trader also accounts for the fact she is not able to reveal her identity when submitting a limit order herself.

[19] An alternative setup would be that traders jointly decide on the transparency regime before the trading day starts.

when using a flag. In contrast, when setting a quote according to the "*own*" strategy, traders do opt to reveal their identity and attach a flag to their order.[20] Overall, as long as c is small enough (i.e. $c \leq \frac{N(V_h - V_l)}{2N+1}$), payoffs under "*all*" are larger than payoffs under "*own*" (see proof). Otherwise, the "*own*" strategy yields the limit order submitters a greater payoff. In sum, when limit order traders can endogenously determine the transparency regime themselves by having the choice to reveal their identity, markets will be fully anonymous for $c \leq \frac{N(V_h - V_l)}{2N+1}$. Otherwise, an identity-revealing flag is attached to limit orders. This result shows that the preferred microstructure of a trading venue may depend in part on the post-trading phase. Proposition 7 presents the equilibrium quotes and spreads for the anonymity and anonymity with flag cases, which are denoted by superscript a and a^f, respectively.

Proposition 7 *With a CSD applying trade-specific fees, traders at all brokers play the following LO strategies:*

- *If traders cannot reveal their identity (i.e. anonymity), then they always play the "all" sub-equilibrium which results in the following optimal quotes and spread:*

$$\begin{aligned} A^{a,*}_{TS,all} &= \frac{2V_h + V_l}{3} - \frac{(N-1)}{3N}c \\ B^{a,*}_{TS,all} &= \frac{V_h + 2V_l}{3} + \frac{(N-1)}{3N}c \\ S^{a,*}_{TS,all} &= \frac{V_h - V_l}{3} - \frac{2(N-1)}{3N}c \end{aligned}$$

- *If traders can reveal their identity through the use of a "flag" (i.e. anonymity with flag). The equilibrium then hinges on the value of c:*

  - *For low values of c, i.e. $c \leq \frac{N(V_h - V_l)}{2N+1}$, traders from each broker target counterparties of all brokers, thus the "all" sub-equilibrium is played, no traders reveal their identity. The equilibrium quotes and spread are:*

$$\begin{aligned} A^{a^f,*}_{TS,all} &= \frac{2V_h + V_l}{3} - \frac{(N-1)}{3N}c \\ B^{a^f,*}_{TS,all} &= \frac{V_h + 2V_l}{3} + \frac{(N-1)}{3N}c \\ S^{a^f,*}_{TS,all} &= \frac{V_h - V_l}{3} - \frac{2(N-1)}{3N}c \end{aligned}$$

  - *For high values of c, i.e. $c > \frac{N(V_h - V_l)}{2N+1}$, traders from each broker only target own counterparties by revealing their identity, thus the "own" sub-equilibrium*

[20] An alternative interpretation is when traders are able to submit quotes on an own broker-crossing network (where only clients can post limit orders).

is played. The equilibrium quotes and spread are:

$$\begin{aligned}
A^{a^f,*}_{TS,own} &= \frac{2V_hN+V_l}{2N+1} \\
B^{a^f,*}_{TS,own} &= \frac{V_h+2V_lN}{2N+1} \\
S^{a^f,*}_{TS,own} &= \frac{(2N-1)}{(2N+1)}(V_h-V_l)
\end{aligned}$$

Proof. See Appendix A. ■

Furthermore, we investigate which setting – anonymity, anonymity with flag, or transparency – a social planner prefers in case it could impose one. Recall from Corollary 4 that the social planner prefers the "*all*" sub-equilibrium when $c \leq \frac{N(V_h-V_l)}{2N+1}$ and the "*own*" sub-equilibrium for larger c. As this outcome exactly coincides with the anonymity with flag case (which has the same cutoff point), the social planner always prefers this "in-between" setting over the anonymity and the transparency settings. Consequently, within Figure 2, the anonymity with flag case corresponds to the maximum welfare curve. Thus the anonymity with flag case perfectly reflects the socially optimal balance between the trading rate and the internalization rate in view of c. This is because it offers the unique combination of having traders to account for the correct uniform break even fee when they play the "*all*" strategy, while still allowing them to reveal their identity and play a zero-fee "*own*" strategy when c becomes high.

5.3 Extension 3: Different Broker Sizes

In a final extension, we consider asymmetries across brokers. Thus far, all brokers had an equal share of affiliated traders. We now introduce large and small brokers, and investigate the impact of this setting on our findings. To conserve space, in this subsection we only present the general setup of the altered model, and its main conclusions hereby focusing on additional insights with respect to the ones presented in the main model. The formal development of the altered model and proofs of all propositions and corollaries is omitted for brevity and can be found in a supplementary appendix to this paper. Furthermore, to focus on the main ideas and intuitions and to keep the exposition tractable, we fix the number of brokers at two (i.e. $N = 2$) and assume that the market shares of both brokers are different. More specifically, a fraction γ (with $\gamma > \frac{1}{2}$) of the total trader population is linked to the "large" broker, and a complementary fraction $1-\gamma$ is linked to the "small" one. Broker affiliations are indexed by subscript $j \in \{large, small\}$. Because traders from different brokers now may possibly choose different strategies, we need to introduce some additional notation. Let $\{all, all\}$ now denote the combination of strategies where traders of the large broker (first element) play the "*all*" strategy and traders from the small broker (second element) also play the "*all*" strategy; $\{own, all\}$ and $\{own, own\}$ then have a similar interpretation.[21]

Under uniform fees, the $\{all, all\}$ combination of strategies is still the only equilibrium. Moreover, the empirical implications of Corollary 1 remain valid: observed market liquidity is increasing in c, and decreasing in γ (i.e. the larger the market share of the large broker γ, the lower observed liquidity becomes). Next, under trade-specific fees, introducing different broker sizes does lead to a number of additional insights, compared to our baseline model. Interestingly, now three possible sub-equilibria exist. Moreover, the quoting behavior of traders is no longer always identical: traders stemming from the large and small broker may quote different prices for their LO. We now briefly discuss the three distinct sub-equilibria. First, for low values of c, traders at both brokers target counterparties at all brokers by quoting relatively liquid prices. Consequently, the $\{all, all\}$ sub-equilibrium applies. Still, an interesting divergence arises compared to our main model. Traders from the small broker have to quote more liquid prices as compared to traders linked to the large broker. The reason is that they need to convince traders from the large broker (who face the opportunity to submit a LO featuring lower expected post-trading fees) to accept their LO. Do note that given this quote setting behavior, in case a counterparty from the same broker hits a standing quote, both traders involved in the trade receive a "bonus" as they both do not have to pay c. An increase in the large broker's market share γ evidently induces traders from the large broker to quote relatively less liquid prices, whereas traders from the small broker are obliged to quote relatively more liquid prices to remain attractive to the traders from the large broker. Second, for sufficiently large post-trading costs (inducing larger cost savings from internalization), both traders from the large and the small broker only address own-broker counterparties such that the $\{own, own\}$ sub-equilibrium applies. Compared to the $\{all, all\}$ sub-equilibrium relatively illiquid prices are quoted, and now the quotes from the small broker are more illiquid as they face a lower execution probability. All quoted prices are also independent of the post-trading fees as these strategies aim at targeting own-broker counterparties only. Third, and this is a new sub-equilibrium, for an intermediate range of post-trading costs traders from the large broker still prefer to target counterparties at their own broker only. They thus set a more illiquid quote since compensating the post-trading fee c a potentially arriving counterparty from the small broker would face, is no longer necessary. In contrast, traders at the small broker alter their strategy and submit relatively liquid quotes targeting all traders. They do so because the gain from increased matching probabilities still outweighs the concessions in terms of aggressive pricing. Hence, within this intermediate post-trading costs range an $\{own, all\}$ sub-equilibrium is played. Corollary 6 presents the general additional result

[21] Recall that the notation in the main model, e.g. the "*all*" strategy, could accordingly be read as $\{all, all, ..., all\}$.

this asymmetric broker size equilibrium brings.

Corollary 6 *When brokers differ in size and a trade-specific fee structure holds, bid and ask quotes of traders hinge upon their broker affiliation due to post-trading fees.*

A direct consequence of this corollary is that markets can become more or less liquid at points in time, depending on which group of traders (from the large or small broker) dominates the trader population at that point.

We illustrate the ask quotes under the uniform and trade-specific fee structures when brokers differ in size in Figure 3. The "$*$" lines correspond to the uniform fee structure, whereas the "$\square$" lines represent the trade-specific fee structure.[22] For the trade-specific fee structure, the different parts correspond to the three distinct sub-equilibria. Panel A draws the ask quotes for traders stemming from the large (full lines) and small broker (dotted lines). In Panel B, we present the "average" ask quote by taking a weighted average of the quotes of large and small broker's traders, using the market share of the respective broker (i.e. γ and $1-\gamma$) as weights. Prices are reported instead of spreads, as there no longer is a unique bid-ask spread due to the fact that traders from large and small brokers quote different prices. Within Panel A we observe, as already mentioned in the discussion above, that traders from the large and the small broker may quote different prices because of differences in post-trading fees. Next, Panel B indicates that the CSD fee structure as well as the level of c influence the average observed liquidity. For low levels of c, the average ask quote under the trade-specific fee structure is most liquid. In contrast, for intermediate and high levels of c the market is most liquid under uniform fees. This finding has policy implications for a regulator who wants to maximize observed liquidity. Technological progress, lowering c, may induce a regulator to implement trade-specific fees instead of uniform fees. Therefore, as in our main model, regulators can improve market liquidity by imposing a fee structure on CSD. The optimal fee structure depends on c.

Please insert Figure 3 around here.

Welfare results with different broker sizes are illustrated in Figure 4. The "$*$" line represents welfare for the uniform fee structure, and the "$\square$" lines for the trade-specific fee structure. As before, the figure highlights a trade-off for the social planner. Recall that the maximum liquidity for high values of c is achieved under the uniform fee structure. However, uniform fees produce the lowest welfare in this range of c. As a consequence, a social planner potentially has to choose between liquidity and social welfare when setting its regulation for a fee structure to be implemented by the CSD: a fee

[22] The exact formulas underlying this figure can be found in the supplementary appendix, which can be found on http://www.econ.kuleuven.be/public/ndaaf41/Files/Internet-Appendix.pdf.

structure implying higher market liquidity in fact reduces social welfare. Moreover, for very low c, both fee structures yield the same welfare, although observed liquidity differs under each fee structure. This leads to the following result, also obtained in the main model: the bid-ask spread is not always an appropriate measure for welfare.

Please insert Figure 4 around here.

6 Concluding Remarks

Explicit transaction costs such as the fees related to clearing and settlement are of considerable importance in today's financial markets. Both in the US and Europe, policies have been implemented in order to reduce post-trading fees. In this paper, we model how internalization of clearing and settlement affects liquidity in financial markets. We find that explicit transaction costs (such as clearing and settlement fees) affect liquidity in financial markets. Two distinct fee structures are analyzed. First, under a uniform fee structure, higher post-trading fees tend to increase observed liquidity. The reasoning is that larger post-trading fees induce more aggressive limit order pricing to convince counterparties to trade. Moreover, the concentration of the broker industry is important: if there is more concentration in the broker industry, this allows the clearing and settlement agent to reduce post-trading fees due to increased internalization opportunities, which in turn induces a decrease in observed liquidity. Second, under a trade-specific fee structure, traders face the following trade-off hinging on the level of the post-trading fees. With low post-trading fees for non-internalized trades, they submit orders to maximize their probability of finding a counterparty. In contrast, with high post-trading fees for non-internalized trades the trade-off tilts towards targeting own counterparties only which implies a higher surplus in case of execution at the expense of a lower probability of execution. Under this fee structure, observed liquidity weakly increases with higher post-trading fees. An increase in the broker industry concentration also increases observed liquidity.

Furthermore, our findings also bear regulatory implications. More specifically, it is shown that liquidity can be improved by imposing a fee structure on the CSD. The liquidity-optimizing fee structure depends on the trading gains level, the marginal costs for the CSD and the concentration in the broker industry. Moreover, our welfare analysis highlights an important trade-off for the social planner: a fee structure implying higher market liquidity may in fact turn out to be detrimental to social welfare. Therefore, liquidity measures do not necessarily constitute good proxies for welfare. In addition, initiatives affecting the marginal cost c in our model (such as TARGET2-Securities (T2S) which aims, among other things, to deliver domestic European-wide settlement at low cost) could impact the equilibrium played and therefore affect liquidity during trading. It is important to recognize and account for this effect when designing legislation regarding the post-trading phase. Noteworthy, all results are robust to extensions of the main model (such as market power for the CSD, anonymous trading and different broker sizes).

In general, our model deals with traders who endogenously decide to opt for enhanced trading opportunities by targeting all counterparties but with considerable post-trading fees, or to opt for the low post-trading fees / low execution probability strategy by addressing own-broker counterparties only. In fact, this endogenous trade-off between the probability of matching and post-trading transaction fees extends to many other situations where transaction fees are important. We consider two closely related examples. First, consumers willing to trade real estate often employ a real estate broker. The real estate broker may search only within its existing customer base, implying a lower matching probability combined with low transaction fees. Alternatively, the real estate broker may enhance matching opportunities by advertising broadly or contacting other real estate brokers leading to higher transaction fees. Second, traders may dramatically increase trading opportunities by making their quotes attractive to cross-border traders. Cross-border trades typically carry large transaction fees. As an alternative, traders may aim at local counterparties only leading to low execution probabilities and lower transaction fees. In future research it would be interesting to apply our theoretical approach to these or other applications where transaction fees differ in the type of the matched counterparty.

Appendix A: Proofs

Proof of Proposition 1.

The equilibrium bid and ask quotes follow immediately from solving the system of indifference conditions delineated in the main text.

Next, we derive the fee strategy for which the CSD breaks even when it charges a uniform per-transaction fee, while accounting for the fact that internalized order flow does not imply costs. Within Proposition 3 we compute the internalization rate under uniform fees as $IR_{U,all} = 1/N$. As $IR_{U,all}$ represents the percentage of trades out of total order flow that is internalized in a given period, its complement stands for the percentage of non-internalized trades, i.e. $1 - IR_{U,all} = (N-1)/N$. Clearly, only the fraction $1 - IR_{U,all}$ induces positive marginal costs for the CSD. As the CSD is active on both sides of the market in each transaction, it should charge a fee to both legs of the trade. More specifically, a CSD charging

$$c_U^* = \left(1 - IR_{U,all}\right) c = \left(\frac{N-1}{N}\right) c$$

on both legs of every transaction (internalized and non-internalized) on average breaks even: it gains on transactions for which it does not face marginal costs and loses on transactions where active clearing and settlement takes place.

Q.e.d. ■

Proof of Proposition 2.

Solving the systems of indifference equations delineated in the main text, taking post-trading fees as given, results immediately in the quotes for the sub-equilibria. We thus only need to prove existence.

Thus, we now investigate under which conditions the different possible strategies correspond to a sub-equilibrium. First, the expected limit order payoffs are computed for the different strategies. Next, we will demonstrate under which conditions the different sub-equilibria will hold. Two distinct possibilities for a sub-equilibrium arise, which one is played depends on the level of c. As in the main text, we assume the proportion of buyers and sellers in the trader population to be equal. This will imply we only have to analyze the expected payoffs of one market side as quotes and expected payoffs of the other market side are completely symmetric. We first compute the limit order payoffs under the two possible strategies:

- "*all*":

The expected payoff of a buyer linked to any broker submitting $B_{TS,all}$ under this strategy is:

$$\pi_{TS,all} = \frac{1}{2}\left[V_h - \left(\frac{V_h + 2V_l}{3} + \left(\frac{N+1}{3N}\right)c\right) - \left(\frac{N-1}{N}\right)c\right]$$

- "*own*":

The expected payoff of a buyer linked to any broker submitting $B_{TS,own}$ under this combination of strategies is:

$$\pi_{TS,own} = \frac{1}{2N}\left[V_h - \left(\frac{V_h + 2V_l N}{2N+1}\right)\right]$$

We now derive under which conditions the different sub-equilibria apply:

1. Sub-equilibrium "*all*" applies when traders at each broker should have no incentives to deviate to the "*own*" strategy This applies when:

$$\pi_{TS,all} > \pi_{TS,own}, \text{ or } c < \frac{2N(N-1)\left(V_h - V_l\right)}{\left(2N+1\right)\left(2N-1\right)}$$

2. Sub-equilibrium "*own*" applies (using similar reasoning) when:

$$\pi_{TS,own} > \pi_{TS,all}, \text{ or } c > \frac{2N(N-1)\left(V_h - V_l\right)}{\left(2N+1\right)\left(2N-1\right)}$$

Q.e.d. ■

Proof of Proposition 3.

The proof is immediate by filling in the stationary probability distribution results of Appendix B in the definition of trading rate and internalization rate.

Q.e.d. ■

Proof of Proposition 4.

The proof is immediate by filling in the computed trading rates and internalization rates (see Proposition 3) in the overall welfare definition.

Q.e.d. ■

Proof of Proposition 5.

Consider first the "*all*" sub-equilibrium. Traders solve a similar system of indifference equations as in the proof of Proposition 2 but now account for the monopoly fee instead of c. For buyers and sellers from each broker j, we respectively have:

$$\begin{aligned}
B^{j}_{TS,all} - V_l - c &= \frac{1}{2}\left[A^{j'\neq j}_{TS,all} - V_l - \frac{(N-1)}{N}c^{m,*}_{TS,all}\right] \\
V_h - A^{j}_{TS,all} - c &= \frac{1}{2}\left[V_h - B^{j'\neq j}_{TS,all} - \frac{(N-1)}{N}c^{m,*}_{TS,all}\right]
\end{aligned}$$

Noting that the quotes are the same across all brokers j, and solving the system of equations results in the equilibrium quotes and spread.

Next, under the "*own*" sub-equilibrium, the proof is identical to the one of Proposition 2 since all trades are internalized and the CSD receives only these trades.

The proof of existence is identical to the proof of Proposition 2 but we have to use $c^{m,*}_{TS,all}$ instead of c in the derivation.

Finally, it is easy to see that, when the CSD has monopoly power, it will set its fee as high as possible within the "*all*" sub-equilibrium. The highest possible fee such that traders do not switch to "*own*" strategies is

$$c^{m,*}_{TS,all} = \frac{2N(N-1)\,(V_h - V_l)}{(2N+1)\,(2N-1)}$$

For the computation of the expected per-period profit for the CSD under the "*all*" and "*own*" sub-equilibrium, we refer to the proof of Proposition 6.

Q.e.d. ■

Proof of Proposition 6.

Under the "*all*" sub-equilibrium, expected per-period trader welfare $TW^{m,*}_{TS,all}$ is computed in the same way is in Proposition 4, the only difference being that $c^{m,*}_{TS,all}$ instead of c is used. The expected per-period welfare of the CSD is the expected per-period profit it obtains, which is

$$\frac{1}{3}\left[\frac{N-1}{N}\right] 2\left(c^{m,*}_{TS,all} - c\right)$$

The expected per-period profit of the CSD is the product of (i) the mark-up $c^{m,*}_{TS,all} - c$ which is charged on each leg of the trade (hence 2 times the mark-up is received per trade); (ii) the fraction of non-internalized trades $\frac{N-1}{N}$; and (iii) the trading rate under "*all*" which is 1/3.

Under the "*own*" sub-equilibrium, the CSD has zero profits (i.e. $CSDW^{m,*}_{TS,own} = 0$) as all trades are internalized and fees are zero. Expected per-period trader welfare $TW^{m,*}_{TS,own}$ is then identical to the result in our base model and is computed in Proposition 4.

Q.e.d. ■

Proof of Proposition 7.

With "full anonymity", an arriving trader does not know whether a standing limit order stems from a trader of her own broker or from another broker. Thus, the market order trader faces expected post-trading fees $\left(\frac{N-1}{N}\right) c$. Consequently, the indifference conditions under the "*all*" sub-equilibrium (which evidently is the only relevant one within this full anonymity setting) for any trader from any broker are:

$$\begin{aligned} B^a_{TS,all} - V_l - \left(\frac{N-1}{N}\right) c &= \frac{1}{2}\left[A^a_{TS,all} - V_l - \left(\frac{N-1}{N}\right) c\right] \\ V_h - A^a_{TS,all} - \left(\frac{N-1}{N}\right) c &= \frac{1}{2}\left[V_h - B^a_{TS,all} - \left(\frac{N-1}{N}\right) c\right] \end{aligned}$$

Solving these indifference equations leads directly to the quotes and spread in the proposition.

Next, within the "anonymity with flag" setting, traders are offered the choice to reveal their identity. As argued in the main text, the equilibrium played is fully determined by investors submitting limit orders. First, consider the "*all*" strategy. We analyze whether traders prefer to reveal their identity by comparing the respective payoffs of both options. The payoffs if traders of each broker do not reveal their identity (left-hand-side) are larger than if they do reveal their identity by adding a flag to their quote (right-hand-side) if:

$$\begin{aligned} & \frac{1}{2}\left[V_h - \left(\frac{V_h + 2V_l}{3} + \frac{(N-1)}{3N}c\right) - \left(\frac{N-1}{N}\right) c\right] \\ > \; & \frac{1}{2}\left[V_h - \left(\frac{V_h + 2V_l}{3} + \frac{(N+1)}{3N}c\right) - \left(\frac{N-1}{N}\right) c\right] \end{aligned}$$

Clearly, this always holds as the equilibrium quotes when not revealing (see the first part of the proposition) are less liquid than when identities are revealed. Therefore, limit order traders always deliberately opt not to reveal their identity when they choose the "*all*" strategy. Second, when is this equilibrium played? Deviating from it is possible by setting a quote according to the "*own*" strategy, and simultaneously reveal their identity using a flag. Limit order traders prefer hiding their identity over revealing as long as the payoff of submitting a limit order under hiding (i.e. $\frac{1}{2}\left[V_h - \left(\frac{V_h+2V_l}{3} + \frac{c(N-1)}{3N}\right) - \left(\frac{N-1}{N}\right) c\right]$) is larger than the payoff of submitting a limit order under revealing (i.e. $\frac{1}{2N}\left[V_h - \left(\frac{V_h+2V_l N}{2N+1}\right)\right]$), which holds as long as $c \leq \frac{N(V_h - V_l)}{2N+1}$. In other words, for this range of c payoffs under "*all*" are larger than under "*own*". Otherwise, the "*own*" strategy yields the limit order submitters a greater payoff.

Q.e.d. ■

Appendix B: Infinite Markov chain in this model

At any given discrete point in time t, the market can be in three possible states (i.e. actions): (1) a trader submits a limit order; (2) a trader submits an internalized market order; (3) a trader submits a non-internalized market order.[23] These three states form a finite state space. For each possible sub-equilibrium s corresponding to fee structure k, a Markov chain (with the property that the next state depends only on the current state) could be constructed with transition matrix $\widehat{M}_{k,s}$, which is a 3×3 matrix capturing all transitions from one state to another (see Colliard and Foucault (2011) for a similar approach). These matrices reflect the transition probabilities corresponding to the equilibrium decisions under the considered sub-equilibrium, and could be written as follows:

$$\widehat{M}_{U,all} = \widehat{M}_{TS,all} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2N} & \frac{N-1}{2N} \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix};$$

$$\widehat{M}_{TS,own} = \begin{bmatrix} \frac{1}{2} + \frac{N-1}{2N} & \frac{1}{2N} & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix};$$

From each of these right stochastic transition matrices, in which each row sums to one and all elements are non-negative, it is possible to derive the stationary probability distribution over all states. More specifically, the stationary distribution $\varphi_{k,s}$ is a row vector satisfying $\varphi_{k,s} = \varphi_{k,s}.\widehat{M}_{k,s}$, i.e. $\varphi_{k,s}$ is a normalized left eigenvector of $\widehat{M}_{k,s}$ associated with the eigenvalue 1. Do note that as this Markov chain is irreducible and aperiodic, the stationary distribution $\varphi_{k,s}$ is unique. Let $\varphi^1_{k,s}$, $\varphi^2_{k,s}$ and $\varphi^3_{k,s}$ be the stationary probability of occurrence of states 1, 2 and 3 under the considered sub-equilibrium. Then the stationary probability distribution could be denoted as $\varphi_{k,s} = \left(\varphi^1_{k,s}, \varphi^2_{k,s}, \varphi^3_{k,s}\right)$. This distribution $\varphi_{k,s}$ could be derived for each of the sub-equilibria as:

$$\begin{aligned} \varphi_{U,all} &= \varphi_{TS,all} = (\frac{2}{3}, \frac{1}{3N}, \frac{N-1}{3N}) \\ \varphi_{TS,own} &= (\frac{2N}{2N+1}, \frac{1}{2N+1}, 0) \end{aligned}$$

and could also be seen as the proportion of time spent in each state within the considered sub-equilibrium.

[23] We do not need to make a distinction between buyers and sellers because both sides of the market are symmetric in our model.

References

Berkowitz, S., Logue, D. and E. Noser, 1988, The Total Cost of Transactions on the NYSE, *Journal of Finance* 32, pp. 159-163.

Biais, B., Glosten, L. and C. Spatt, 2005. Market Microstructure: a Survey of Microfoundations, Empirical Results, and Policy Implications, *Journal of Financial Markets* 8, pp. 217-264.

Colliard, J.-E. and T. Foucault, 2011, Trading Fees and Efficiency in Limit Order Markets, working paper.

Comerton-Forde, C., Putnins, T. and K.M. Tang, 2011, Why Do Traders Choose to Trade Anonymously?, forthcoming *Journal of Financial and Quantitative Analysis.*

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, pp. 319-338.

Domowitz, I. and B. Steil, 2002, Innovation in Equity Trading Systems: the Impact on Trading Costs and the Cost of Equity Capital, in Steil, Benn, David G. Victor, and Richard R. Nelson (eds.), Technological Innovation and Economic Performance, Princeton: Princeton University Press.

DTCC, 2003, Managing Risk in Today's Equity Market: a White Paper on New Trade Submission Safeguards, Depository Trust & Clearing Corporation report.

Duffie, D., Gârleanu, N. and L. Pedersen, 2005, Over-the-Counter Markets, *Econometrica* 73, pp. 1815-1847.

Ellul, A. and M. Pagano, 2006, IPO Underpricing and After-Market Liquidity, *Review of Financial Studies* 19, pp. 381-421.

Foucault, T., 1999, Order Flow Composition and Trading Costs in a Dynamic Limit Order Market, *Journal of Financial Markets* 2, pp. 99-134.

Foucault, T. and C. Parlour, 2004 , Competition for Listings, *RAND Journal of Economics* 34, pp. 328-355.

Foucault, T., Kadan, O. and E. Kandel, 2005, Limit Order Book as a Market for Liquidity, *Review of Financial Studies* 18, pp. 1171-1217.

Foucault, T., Kadan, O. and E. Kandel, 2011, Liquidity Cycles, and Make/Take Fees in Electronic Markets, working paper.

Foucault, T., Moinas, S. and E. Theissen, 2007, Does Anonymity Matter in Electronic Limit Order Markets?, *Review of Financial Studies*, 20, pp. 1707 - 1747.

Glosten, L., 1998, Competition, design of exchanges and welfare, *Unpublished manuscript*, Columbia University.

Goettler, R., Parlour, C. and U. Rajan, 2005, Equilibrium in a Dynamic Limit Order Market, *Journal of Finance* 60, pp. 2149-2192.

Handa, P., Schwartz, R. and A. Tiwari, 2003, Quote Setting and Price Formation in an Order Driven Market, *Journal of Financial Markets* 6, pp.461-489.

Hollifield, B., Miller, R., Sandås, P. and J. Slive, 2006, Estimating the Gains from Trade in Limit Order Markets, *Journal of Finance* 61, pp. 2753-2804.

Holthausen, C. and J. Tapking, 2007, Raising Rival's Costs in the Securities Settlement Industry, *Journal of Financial Intermediation* 16, pp. 91-116.

Koeppl, T., Monnet, C. and T. Temzelides, 2012, Optimal Clearing Arrangements for Financial Trades, *Journal of Financial Economics* 103, pp. 189-203.

Madhavan, A., 2000, Market Microstructure: a Survey, *Journal of Financial Markets* 3, pp. 205-258.

Malinova, K. and A. Park, 2011, Subsidizing Liquidity: The Impact of Make/Take Fees on Market Quality, working paper University of Toronto.

Oxera, 2009, Monitoring Prices, Costs and Volumes of Trading and Post-Trading Services, Oxera report.

Oxera, 2011, Monitoring Prices, Costs and Volumes of Trading and Post-Trading Services, Oxera report.

Parlour, C., 1998, Price Dynamics in Limit Order Markets, *Review of Financial Studies* 11, pp. 789-816.

Rindi, B., 2008, Informed Traders as Liquidity Providers: Transparency, Liquidity and Price Formation, *Review of Finance* 12, pp. 497-532.

Rochet, J.-C., 2005, The Welfare Effects of Vertical Integration in the Securities Clearing and Settlement Industry, working paper.

Roşu, I., 2009, A Dynamic Model of the Limit Order Book, *Review of Financial Studies* 22, pp. 4601-4641.

Tapking, J., 2007, Pricing of Settlement Link Services and Mergers of Central Securities Depositories, working paper.

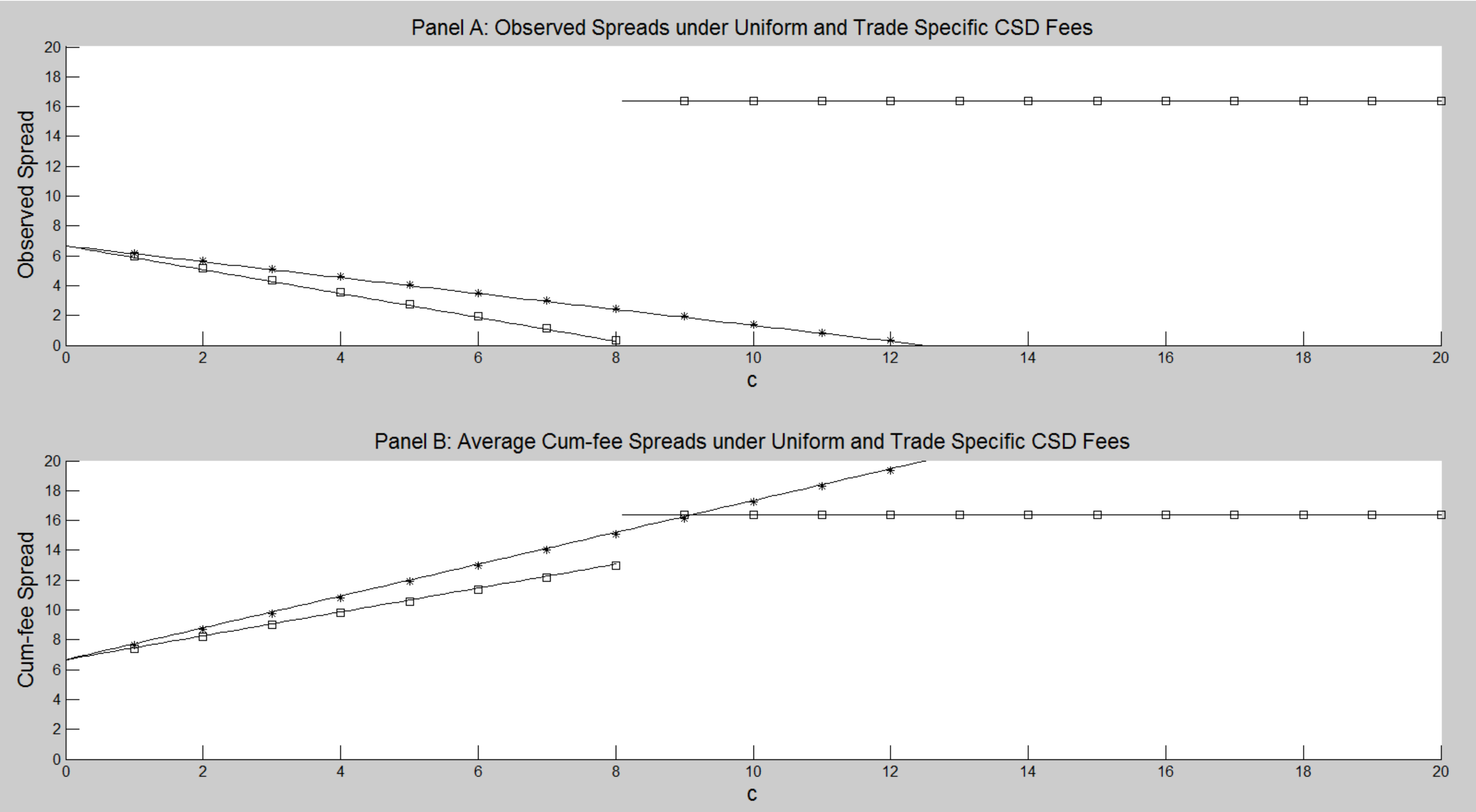
Tapking, J. and J. Yang, 2006, Horizontal and Vertical Integration in Securities Trading and Settlement, *Journal of Money, Credit and Banking* 38, pp. 1765-1795.

Van Achter, M., 2009, A Dynamic Limit Order Market with Diversity in Trading Horizons, working paper.

Van Cayseele, P. and C. Wuyts, 2008, Cost Efficiency in the European Securities Settlement and Depository Industry, *Journal of Banking and Finance* 31, pp. 3058 - 3079.

Figure 1: Spreads under Different CSD Fee Structures

Note: This figure illustrates the results of our main model using the following parameter values: $V_h = 20$, $V_l = 0$, $N = 5$, and c, shown on the x-axis, varies in the interval $[0, 20]$. Panel A presents observed spreads under the uniform $(*)$ and trade-specific $(\square)$ CSD fee structures, as computed in Propositions 1 and 2. Panel B shows average cum-fee spreads under both fee structures as computed in Equations 2 and 4.


Panel A: Observed Spreads under Uniform and Trade Specific CSD Fees

Panel B: Average Cum-fee Spreads under Uniform and Trade Specific CSD Fees

Figure 2: Overall Welfare Per Period

Note: This figure illustrates the results of our main model using the following parameter values: $V_h = 20$, $V_l = 0$, $N = 5$, and c, shown on the x-axis, varies in the interval $[0, 20]$. Overall welfare is shown under the uniform ($*$) and trade-specific ($\square$) CSD fee structures, as computed in Proposition 4

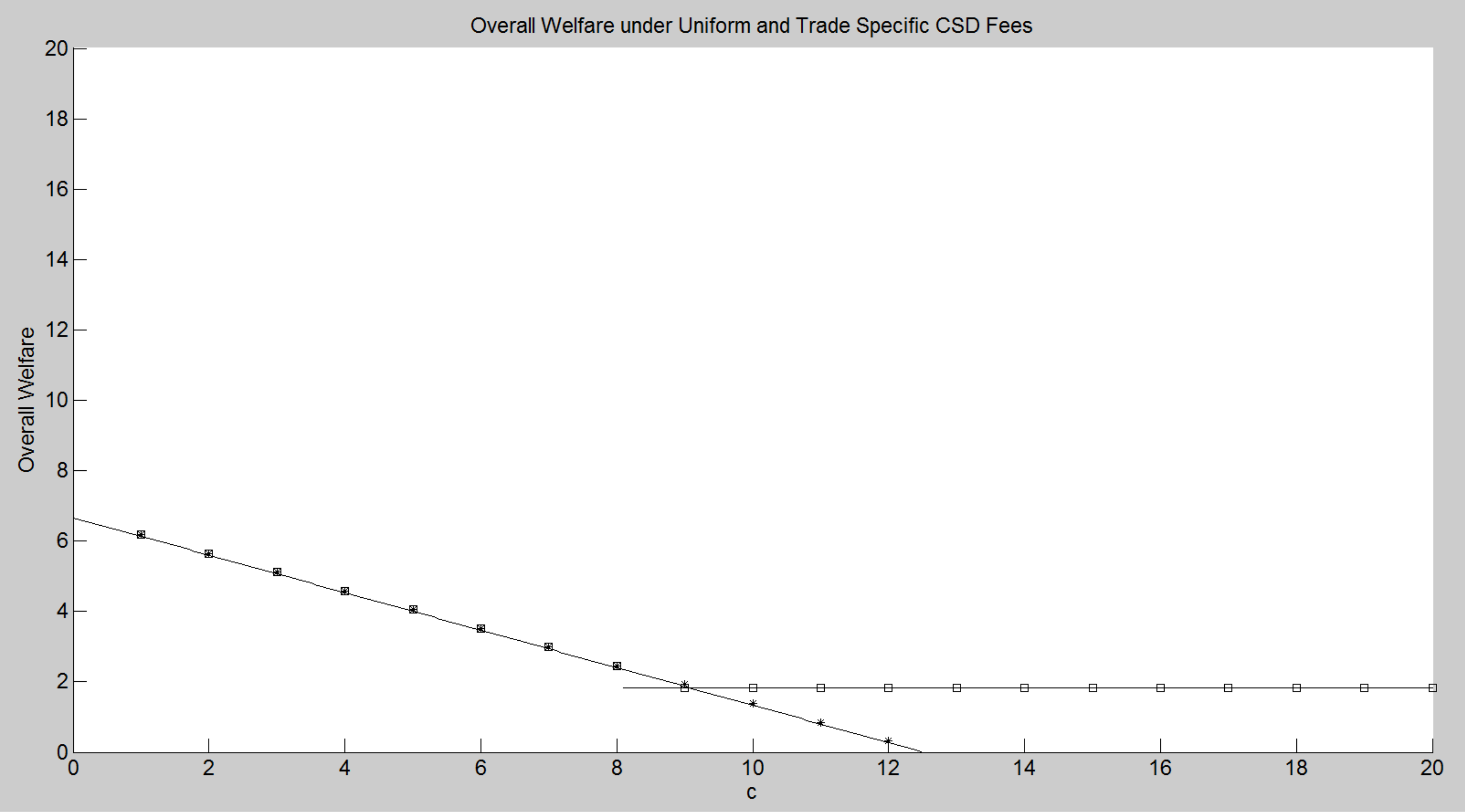
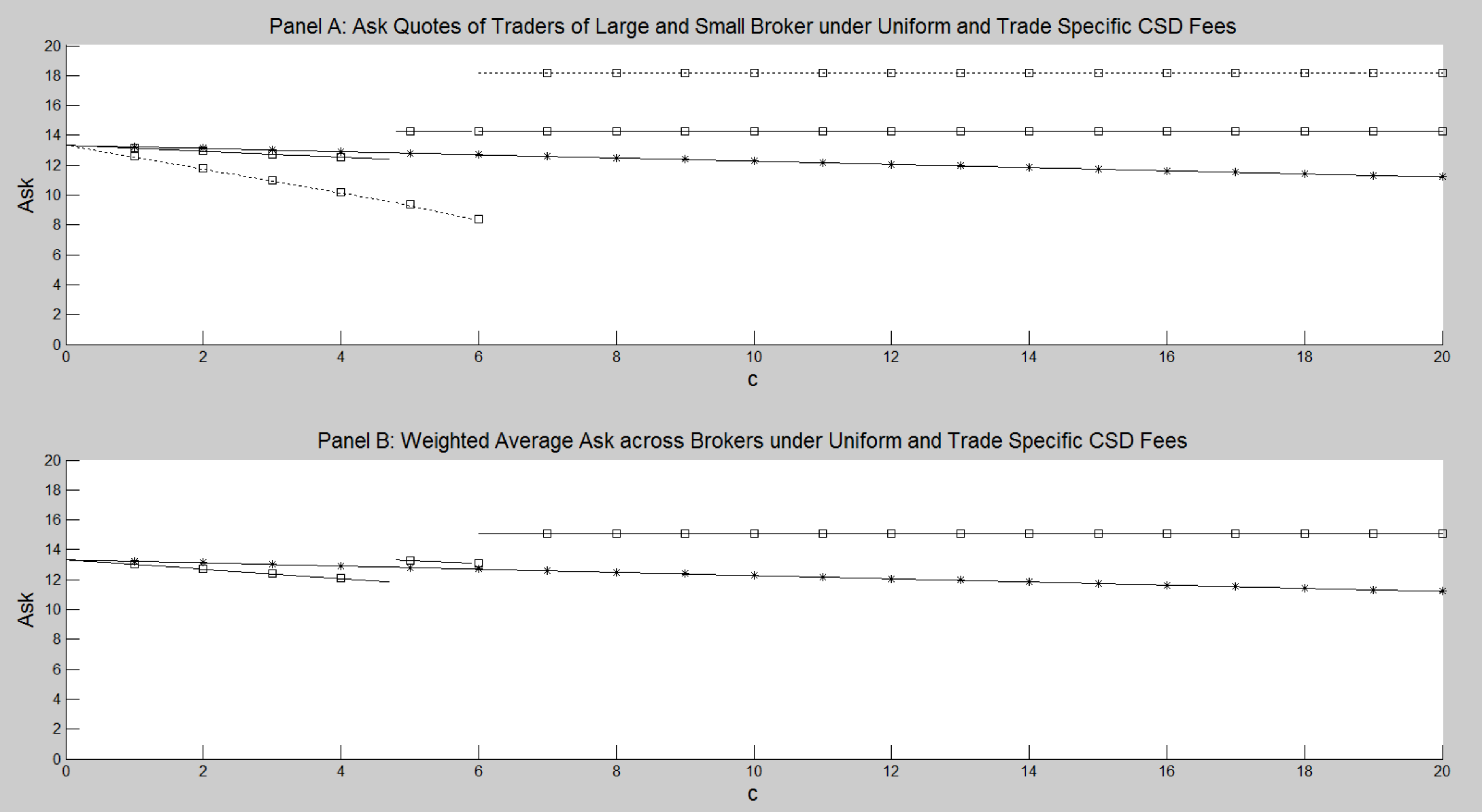


Figure 3: Ask Quotes Under Different CSD Fee Structures with Differences in Broker Size

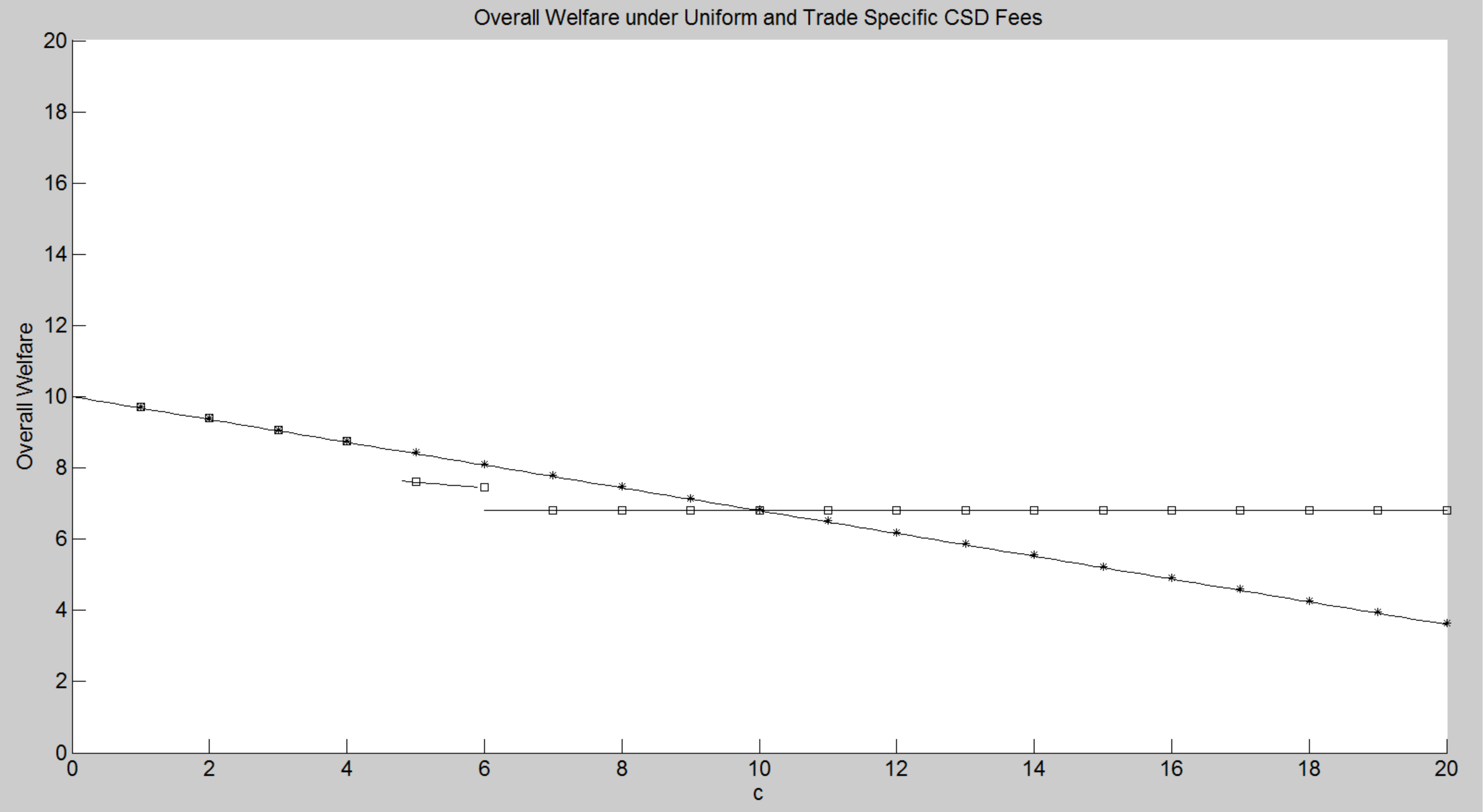
Note: This figure illustrates the results of our model for the different broker sizes extension using the following parameter values: $V_h = 20$, $V_l = 0$, $N = 5$, and c, shown on the x-axis, varies in the interval $[0, 20]$. Panel A presents ask quotes of traders of the large broker (full lines) and small broker (dotted lines) under the uniform ($*$) and trade-specific ($\Box$) CSD fee structures. Panel B shows the weighted average ask quotes across traders of the large and small brokers, with the weights being the market shares of the respective brokers (γ and $1 - \gamma$). The ask quotes are computed in the supplementary appendix to this paper.


Panel A: Ask Quotes of Traders of Large and Small Broker under Uniform and Trade Specific CSD Fees

Panel B: Weighted Average Ask across Brokers under Uniform and Trade Specific CSD Fees

Figure 4: Overall Welfare Per Period with Differences in Broker Size

Note: This figure illustrates the results of our model for the different broker sizes extension using the following parameter values: $V_h = 20$, $V_l = 0$, $N = 5$, and c, shown on the x-axis, varies in the interval $[0, 20]$. The figure presents overall welfare under the uniform $(*)$ and trade-specific $(\Box)$ CSD fee structures as computed in the supplementary appendix to this paper.


Overall Welfare under Uniform and Trade Specific CSD Fees
Overall Welfare
c