Plumbing of Securities Markets: The Impact of Post-Trade Fees on Trading and Welfare¹

Hans Degryse², Mark Van Achter³ and Gunther Wuyts⁴

October 2020

¹We would like to thank the Editor (Gustavo Manso) an anonymous Associate Editor, and three anonymous referees for their thoughtful comments and suggestions that allowed to improve our paper. Furthermore, we 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. An earlier version of the paper circulated as 'Internalization, Clearing and Settlement, and Liquidity'.

²KU Leuven and CEPR, E-mail: hans.degryse@kuleuven.be

³The Liquid House and KU Leuven, E-mail: mark.vanachter@kuleuven.be

⁴KU Leuven, E-mail: gunther.wuyts@kuleuven.be

Abstract

We analyze how market design choices about counterparty information and costs related to the “plumbing” (i.e., clearing, settlement, and custody) of securities markets affect market quality and welfare. Our model compares two post-trade fee structures for allocating these costs. One charges a uniform fee for all trades, the other, marginal-cost-based structure a reduced fee for internalized trades (which, both traders being from the same broker, are less costly to process). Both market design and fee structure affect quote aggressiveness, trading volume and its composition. With marginal-cost-based fees and counterparty information being available, traders decide which counterparties to target through quote aggressiveness, trading off execution probability against fee. A social planner can maximize welfare by requiring marginal cost-based fees and providing traders the choice to disclose counterparty information.

JEL Codes: G10, G15

Keywords: clearing and settlement, market design, financial market infrastructure, internalization, welfare

1 Introduction

Financial markets deviate from the perfect-market ideal due to impediments to trade. This paper shows how a heretofore overlooked transaction cost, post-trade fees, determines market quality and welfare. This link may appear obvious – order processing cost models, for example, would suggest that larger post-trade fees lead to wider spreads – but this intuition may be misleading. The impact of post-trade fees depends on how the costs are allocated across counterparties, and thus the available information regarding counterparties’ identities. Our model shows that larger post-trade fees induce traders to post more aggressive quotes or submit illiquid limit orders with a small execution probability.

Post-trade fees are charged by the post-trade infrastructure to transfer securities after a buyer and seller agreed to trade. These fees cover the costs related to clearing, settlement, and custody. They are sizeable and influence traders’ actions. An Oliver Wyman report estimates that the post-trade and securities servicing fees are in the region of USD100 billion.¹ Oxera (2014) shows fees for institutional investors to range, on average, from 0.02 (United States) to 8.01 basis points (Spain), and for retail investors from 0.15 (United States) to 11.71 basis points (Spain). The “plumbing” of securities markets has in recent years become a topic of considerable interest to legislators and market operators, many of the latter currently in the process of modifying their fee structures.² Our model shows that such changes drive trader incentives and welfare, depending on the design of markets. We document these changes, the views of various involved parties, and the link to our model in more detail in Section 2.

We analyze the impact of post-trade fees by modeling the trading phase as a limit order market, and consider different market designs related to counterparty information. In line with practice, the post-trade infrastructure employs one of two fee structures. A first scheme charges a uniform fee which does not differ between internalized (i.e.,

¹<http://www.oliverwyman.com/content/dam/oliver-wyman/global/en/2016/feb/BlockChain-In-Capital-Markets.pdf>

²The terminology “plumbing” is often used in trading or regulatory environments (e.g., European Commission, (2016)).

buyer and seller employ the same broker) and non-internalized trades. A second scheme charges lower post-trade fees for internalized trades than for non-internalized trades. We label the latter “trade-specific fees,” and assume each trade to be charged at its marginal cost. To reflect the implications of the fee structure, we model each trader as linked to a broker and consider three market design choices regarding counterparty information, i.e., the broker affiliation of a standing limit order. In particular, we consider anonymity, transparency, and optional disclosure of broker affiliation. With trade-specific fees and information about counterparties, a trader endogenously chooses which counterparties to target, resulting in the following trade-off. Targeting all potential counterparties is associated with high execution probability, but necessitates an aggressive quote to offset the high post-trade fees that will be incurred by traders using other brokers. Moreover, she is likely to incur a high post-trade fee herself. Fees are lower and quotes can be less aggressive when only own-broker counterparties are targeted (i.e., only internalized trades are undertaken). With uniform post-trade fees or anonymity, the trade-off simplifies as traders then always target all potential counterparties, or abstain from trading.

Our model yields a number of novel insights. First, market design about counterparty information and the post-trade fee structure determine traders’ limit order (LO) submission strategies, and thus which traders are matched. With uniform fees or anonymity, traders have no incentive to be only matched with other traders who share the same broker. Conversely, with trade-specific fees and information about counterparty identity, LO traders may choose to target all possible counterparties when post-trade fees are low, or own-broker counterparties when fees are high. Anonymity thus prevents trade-specific fees to have a material impact on how traders are matched.

Second, we show that optional disclosure of broker ID combined with trade-specific fees delivers the socially preferred outcome. Market design and post-trade fees therefore should be investigated jointly. Welfare is driven by trading volume, and by the composition of this volume since internalized trades due to the lower post-trade costs produce higher welfare than non-internalized trades. We show that welfare in the absence of post-trade frictions, i.e., “benchmark welfare”, cannot be reached due to two mutually

exclusive inefficiencies resulting from these post-trade frictions. The first inefficiency is due to non-internalized trades generating post-trade costs (as opposed to internalized trades), the second stems from lower trading volume when foregoing non-internalized trades. When a social planner optimizes welfare by selecting the preferred market design and post-trade fee structure, it trades off these two inefficiencies.

We find that for low post-trade costs, the social planner prefers the market design and post-trade fee structure that has the highest trading volume. While non-internalized trades then do take place and generate post-trade costs, the beneficial effect of a higher volume is larger. For high post-trade costs, the social planner prefers a market design and post-trade fee structure that only leads to internalized trades. Trading volume is then lower, but the post-trade costs of non-internalized trades are avoided and the higher welfare produced by a higher volume of internalized trades dominates. Only optional disclosure of broker ID combined with trade-specific fees allows to align the LO traders' incentives with the socially preferred outcome over the entire post-trade cost range. We show that for low and intermediate post-trade costs, traders conceal their identity. They then prefer to target all potential counterparties which aligns with the social planner's preferences in this cost range. For high post-trade costs, traders are transparent about their identity and only internalized trades take place, as is socially preferred.

Third, and counterintuitively, higher post-trade fees distort traditional bid-ask spread measures. Traders submit limit orders at more aggressive quotes to partially compensate for the higher post-trade fees. A higher post-trade fee reduces the surplus to be shared by buyers and sellers. The observed bid-ask spread (computed from LO quotes) decreases and markets seem to become more liquid. Cum-fee bid-ask spreads (i.e., the spreads of market order (MO) traders after adding post-trade fees), however, increase in post-trade fees, as more aggressive quotes only partially compensate for higher post-trade fees. That MO payoffs are eroded despite more aggressive observable quotes is captured in the cum-fee spread.

Our analysis urges that the entire value chain of trading and post-trading should be accounted for when making assessments on the trading phase. Building on this insight,

we put forward a set of novel empirical and regulatory implications. On the empirical side, we show that trading volume not only hinges on market design choices about trading but also on how post-trade costs are allocated as reflected in the post-trade fee structure. We further argue that commonly-used market quality measures should account for post-trading fees to form a relevant measure of the true costs faced by traders. As highlighted in our model results, observed and cum-fee spreads may move in different directions upon post-trade fee changes, potentially inducing erroneous market quality assessments. Similarly, on the regulatory side, best-execution rules (such as the US Securities and Exchange Commission’s “trade-through” and “trade-at” rules) should be based on cum-fee quotes. Executing a trade based on the quote aggressiveness of the standing limit orders ignores that such a trade may be inferior to a similar trade at higher quotes but with substantially lower post-trade fees. In contrast, cum-fee quotes correctly incorporate the welfare effects of the endogenous adjustment of traders’ strategies to the way post-trade costs are allocated.

Finally, we discuss a number of extensions including fee designs that charge different post-trade fees to limit orders and market orders, reminiscent of make/take fees. A novel finding is that with trade-specific fees and information about counterparty identity, charging the entire post-trade fee on the limit order also implements the social planner’s preferred outcome.

Our paper contributes to several strands in the literature. First, it bridges the formerly segregated literatures on market quality, market design, and organization of the post-trade infrastructure.³ Although the link between post-trade fees and market quality has been empirically observed (e.g., Foucault, Sraer, and Thesmar (2011) and Kadanakkam (2000)), ours is the first paper to provide theoretical guidance and to analyze the corresponding welfare effects. Second, our paper relates to the literature that assesses the impact of make/take fees on market quality. Theoretical analyses are provided by

³Analyzing the determinants of market quality in a wide variety of settings has been a focus of the broad market microstructure literature surveyed in Foucault, Pagano, and Roell (2013b). The more recently developed clearing and settlement literature includes contributions by Duffie and Zhu (2011), Fotak, Raman, and Yadav (2014), Holthausen and Tapking (2007), Koepl, Monnet, and Temzelides (2012), Loon and Zhong (2012), Merrick, Naik, and Yadav (2005), Rochet (2005), Tapking (2007), and Tapking and Yang (2006).

Brolley and Malinova (2017), Colliard and Foucault (2012) and Foucault, Kadan, and Kandel (2013a), while Malinova and Park (2015), Anand, Hua, and McCormick (2016) and Skjeltorp, Sojli, and Tham (2018) address empirically the effect of make/take fees. More broadly, we also relate to the literature on stamp duties and financial transaction taxes (e.g., Colliard and Hoffmann (2017)). Make/take fees and financial transaction taxes alter the net private valuation of a trader. Both are imposed on an individual trader and do not depend on the counterparty. In contrast, we analyze the impact of various post-trade fee structures established by the post-trade infrastructure (as opposed to fees applicable during trading). A key difference is that post-trade fees may hinge on the counterparties involved in the trade, which induces different trading strategies and welfare effects.

Third, we contribute to the literature that models dynamic trading in limit order books (e.g., Foucault, Kadan, and Kandel (2005), Goettler, Parlour, and Rajan (2005), and Roşu (2009)) by including the impact of (heterogeneity in) post-trade fees on limit order submission strategies as well as market design about counterparty information. Lastly, our work complements the theoretical literature that analyzes the impact of transparency on liquidity (e.g., Baruch (2005), Bloomfield and O'Hara (1999), Madhavan (1995), Naik, Neuberger, and Viswanathan (1999), Pagano and Röell (1996), and Rindi (2008)). We examine how information about a counterparty's identity affects liquidity. This topic is expected to come at the forefront as the degree of counterparty transparency is a critical design choice in the new Blockchain technology (Malinova and Park (2016)).

The remainder of the paper is structured as follows. Section 2 provides details of the post-trading infrastructure related to our model. Section 3 describes the model's set up. Section 4 presents our general solution approach and equilibrium definition. Section 5 shows the order submission strategies and corresponding equilibria for the three transparency regimes and two fee structures. Section 6 considers welfare. Section 7 discusses drivers of spreads and considers asymmetric fee allocation, while Section 8 provides the empirical and regulatory implications of our model. Robustness and extensions of the model are discussed in Section 9. Section 10 concludes. The proofs of

all propositions and corollary's are in Appendix A.

2 Post-Trade Infrastructure: Link to Practice

Post-trade costs are lower for within-broker (i.e., internalized) than for across-broker (i.e., non-internalized) trades because the former are less expensive to process and thus carry a lower marginal cost. Within-broker trades can bypass the settlement chain as they are netted out at the broker level in the settlement phase.⁴ In contrast, across-broker trades need to be processed by a number of institutions from execution on the exchange until settlement, such as the trade repository, custodian banks, or the central securities depository (CSD).

Modifications made by market operators and regulators to the handling and reporting of internalized trades further motivate our modeling choices. The implemented changes determine how post-trade costs are allocated, and are driven mainly by regulatory preferences regarding the monitoring of trade and post-trade activity. Mounting pressure in the United States for real-time trade reporting induced the NSCC to move, on July 1, 2006, from “*a clearing fee schedule that was primarily based on transaction count*” to a new “netting fee” model “*to address certain economic factors that it believes have influenced firms’ trade submission practices*” (SEC (2006)). The implementation of a “flat trade clearance fee” implied an economic incentive for brokers to pass on to the NSCC only more costly non-internalized trades, and process internalized (or netted) transactions themselves. The new “netting fee” factors in the benefits from a broker executing internalized trades. As such, the NSCC aims to incentivize brokers to directly send in all their transactions. As the NSCC is the only clearing and settlement agency for equities in the United States, this netting advantage applies to orders regardless of where (i.e., on-exchange, independent of trading venue, or off-exchange) the trade executes. Since

⁴In particular, this netting at broker level avoids book-entry movement in the settlement phase. Indeed, in the United States and most European countries, brokers are permitted to settle trades internally on their “omnibus accounts” at the Central Securities Depository (CSD), on which securities that belong to multiple end investors are held. Only a few countries impose a “segregated account” per end investor. An overview can be found in AFME (2012), AFME (2015) or Weinstein and Yekini (2015).

February 14, 2014 the real-time reporting requirement has become more strict as any “pre-netting” at the broker level is prohibited. Our model captures this evolution as a move from a uniform to a trade-specific fee structure.

Similarly in Europe, prior to the November 1, 2007 implementation of the Markets in Financial Instruments Directive (MiFID), several regulated markets (i.e., the former national stock exchanges) had begun to allow internalized trades at lower post-trade fees. Their underlying motivation was to prevent brokers from becoming “systematic internalizers” (a new type of trading platform allowed by MiFID), which would have drastically reduced trading on regulated markets. NYSE Euronext, for example, initiated an Internal Matching Service that fully exempted internalized order book executions on its platform from post-trade fees (which can represent up to 40% of total transaction costs, according to NYSE Euronext). Even before MiFID launched, a proportion of brokers corresponding to 70% of all trades had signed up for this service (NYSE Euronext (2008)).⁵ Further examples of fee reductions or exemptions for internalized trades include Deutsche Börse’s internal settlement processing and the London Stock Exchange’s SETS Internalizer. That other clearing houses apply uniform fees (e.g., EuroCCP) demonstrates that uniform and trade-specific fee schedules are implemented in practice. Furthermore, in line with the evolution in the United States, also in Europe there is an increasing regulatory interest to monitor internally settled volumes. This is for instance exemplified by the tightening of reporting requirements.⁶

⁵Trading fees may apply for trades executed on an on-exchange matching engine such as the Euronext Internal Matching Service. Off-exchange internalization would be an alternative. One reason why member firms might still want to execute on the order book is that the ‘Internal Matching Service will be charged on the basis of the Member Firm’s existing fee tiers, allowing users to benefit from the usual volume discounts as appropriate’. Often, marginal trading fees become zero when member firms execute quite some volume. Furthermore, subscribing members benefit from the liquidity of the central order book and save on the cost of building an IT system for order execution.

⁶The European Securities and Markets Authority (ESMA) published in March 2018 the final guidelines on how to report internalized settlement. This will enable ESMA to monitor growing internally settled volumes (ESMA (2018)).

3 Model Setup

3.1 Post-trade Infrastructure and Fee Structures

We develop an infinite horizon model of a limit order market that lists a single security. Post-trade fees set by the Post-Trade Service Provider (PTSP) prior to the start of trading are taken as given by traders during the subsequent trading game. Each period, $t = 0, 1, \dots, +\infty$, a single risk neutral trader arrives and maximizes the order's expected payoff. Buyers and sellers, each arriving with probability $1/2$, have a private valuation of the asset equal to V_h and V_l , respectively, where $V_h > V_l \geq 0$. This implies that trading gains could be realized in the absence of post-trade fees. Constant marginal utility is assumed. Traders decide on the choice between limit order and market order, and do not delegate this decision to their broker. In line with practice, arriving traders can observe whether a limit order is standing in the order book or not.

Each trader is linked to one of $N > 1$ brokers through which individual orders are sent to the market. Each broker has an equal share of affiliated traders. Broker affiliations are indexed by the superscript $j \in \{1, \dots, N\}$. Because $N > 1$, trades are either internalized (i.e., buyer and seller are affiliated with the same broker) or non-internalized (i.e., buyer and seller are affiliated with different brokers).

The PTSP handles clearing and settlement immediately after each transaction, is risk neutral, and has no fixed operating costs. It does, however, incur a marginal cost c per leg of the trade for non-internalized trades, and a lower marginal cost for internalized trades, which we normalize to zero. The PTSP aims to break even, on average, but does not necessarily charge the marginal cost for each transaction. We consider two fee structures: uniform and trade-specific. With uniform fees, the PTSP breaks even, on average, by charging the same fee for all trades. A PTSP that implements trade-specific fees breaks even by charging fees equal to the respective marginal cost of internalized and non-internalized trades. We assume throughout the analysis that all transactions are cleared and settled through the PTSP. Thus, a broker cannot establish its own clearing and settlement system to accommodate internalized trades. Relaxing this assumption

would imply that the setups of the uniform and trade-specific fee structures coincide since brokers only would pass on the non-internalized trades to the PTSP. Notice however that the PTSP enjoys the same technology as each broker would have for its within-broker trades. In particular, a within-broker trade does not involve a marginal cost. The PTSP is the only one who can settle across-broker trades; these involve a marginal cost of $2c$.

3.2 Market Design Choices

Traders are linked to a broker. Different market design choices for the limit order market about whether this broker affiliation is transparent or not can be made. We consider three market design choices: anonymous trading, transparency about broker-IDs, and optional disclosure of broker-ID. In the first regime, we assume anonymous trading, which is the standard in many markets. Important in our setting is that this implies that the broker ID cannot be observed (nor revealed).⁷ A second regime would be “transparency”, i.e., the broker affiliation is always shown with the order and thus observable to counterparties. While this may not directly apply to order-driven markets, this transparency setting serves as a benchmark. A third market design choice is to allow the limit order trader to decide on whether or not she wants to reveal her broker ID. We call this regime “optional broker disclosure”. Market participants’ identity can be optionally disclosed in some markets (e.g., NASDAQ) and is concealed in others (e.g., NYSE Euronext). Also other trading venues have altered the degree of counterparty transparency over the past two decades.⁸

⁷One could argue that a broker knows whether he has outstanding client orders in the limit order book, and could use that information when trade-specific fees apply. In that event, the anonymity setup would coincide with optional disclosure of broker-ID when (i) traders would delegate the choice between MO or LO to their broker, and (ii) the broker acts in the interest of its clients. The outcomes of optional disclosure with trade-specific fees are shown in Subsection 5.3

⁸Gozluklu (2013) provides an overview of transparency and disclosure in different trading venues. See Foucault, Moinas, and Theissen (2007) or Rindi (2008) for analyses of transparency.

4 Order Submission Strategy and Equilibrium Definition

In our model, all parameters of the model are known to the traders and constant over time. This enables us to solve for Markov perfect equilibria in which traders' strategies hinge only on the market state upon arrival, not on the history of the trading game or time of arrival. This also allows us to suppress the time dimension t when presenting the equilibria.⁹ We are considering stationary order placement strategies meaning that independent of time every trader of the same type chooses the same order placement strategy when faced with the same state of the LOB. The subgame perfect equilibrium of the market is an order submission strategy that maximizes the expected gains of trade of the arriving trader whose beliefs are that other traders follow strategies on the same equilibrium path.¹⁰

A trader's order submission strategy depends on the state of the limit order book (LOB) at time of arrival t . The state of the order book contains the best quotes as well as information on whether a limit order in the book was sent by the same broker as the current investor or not. Formally, the state $s_t = (A, B, id)$, where A is the current ask ($A = +\infty$ if there is no ask in the book), B is the current bid ($-\infty$ if no bid in the book), and $id \in \{own, other, either\}$ indicates the information investor arriving at $t + 1$ on the broker's identity of investor submitting a LO at time t . Broker identity id can thus be the *own* broker, *other* broker, or *either* of them.

There are three components to a trader's order placement decision: the order type choice (MO or LO), and when submitting a LO, the bid price B or ask price A of that LO and whether or not to reveal her broker-id. A LO remains in the book for one period, and are thus take-or-leave offers for the next trader (e.g., Foucault (1999))

⁹The pure strategy equilibria derived for this model are unique. A mixed strategy equilibrium in which traders always randomize between submitting a LO and a MO is not deviation-proof. Breaking the indifference by improving the LO quote by a tiny fraction epsilon is always preferred as it increases the execution probability significantly. The derived pure strategies equilibria could be seen as the limit of this epsilon-argument. This is in line with the related literature (e.g., Colliard and Foucault (2012), Foucault (1999), Goettler et al. (2005), Hoffmann (2014), Van Achter (2010)).

¹⁰Deviations by a trader imply also beliefs that other traders follow suit with similar deviations.

takes a similar approach).¹¹ Alternatively, when an interesting LO is outstanding, the trader can submit a MO for which execution is certain. The LO's execution probability is uncertain and endogenous as it depends on the targeted traders' order placement strategies. Orders are for one unit of the asset, and non-execution of LOs implies no cost. The expected profit of a buy LO at price B is $\gamma[V_h - B - E(\text{fee}(\text{LO}))]$, with γ the endogeneously determined execution probability, E the expectations operator, and $\text{fee}(\text{LO})$ the post-trade fee incurred for a LO. The expected profit of a buy MO becomes $V_h - A - E(\text{fee}(\text{MO}))$. We need the expectation operator as at the time of order submission there may still be uncertainty about the fee the trader will incur. Similar expressions apply for a sell LO and sell MO.

A trader's order placement strategy thus is a mapping $O(\cdot)$ from the set of possible values for the state of the LOB to $\{\text{MO to buy, MO to sell, LO with broker-id, LO without broker-id}\} \times \mathbb{R}^2$. For each possible state of the market, the strategy thus specifies the order type (MO or LO), including the decision to disclose the broker-id or not, and the quotes (A, B) associated to the placement of limit orders. In some market designs, the decision to disclose the broker-id or not becomes irrelevant.

In general, we show that our model has two candidate equilibrium strategies, of which one will prevail depending upon market design and parameter conditions. More specifically, in setting optimal bid or ask quotes when submitting a LO, a trader generally has two possibilities, (1) establish a quote that attracts only counterparties from the same broker (which strategy we term "*own*"), or (2) opt for a quote attractive to all possible counterparties (which strategy we term "*all*"). A trader indifferent between the "*all*" and "*own*" strategies is assumed to attract the broadest set of counterparties (i.e., the "*all*" strategy). Moreover, traders are assumed to choose a MO when they are indifferent to submitting a new LO. A stationary market equilibrium is defined as a set of mutual order submission strategies (specifying an optimal order type, quote and corresponding

¹¹Our model restricts the lifetime of limit orders to one period. This makes submitting limit orders less attractive than when they would have a longer lifetime. Extending the lifetime of limit orders to more periods (e.g., two) considerably complicates the analysis. However, as it affects all our market design choices (opacity, optional disclosure and transparency) symmetrically, we do not expect such longer life time to qualitatively impact our findings. We refer to Degryse and Karagiannis (2019) for a model with a longer life time of limit orders in a context of priority rules.

execution probability to each possible state of the LOB) such that each trader's strategy is optimal given the strategies of all other traders.

We now explain our general solution approach. Traders submitting LOs set their quotes to satisfy the following system of indifference equations:

$$\begin{aligned} B - V_l - E(fee(MO)) &= \gamma [A - V_l - E(fee(LO))], \\ V_h - A - E(fee(MO)) &= \gamma [V_h - B - E(fee(LO))]. \end{aligned} \quad (1)$$

The first equation considers the buyer's decision. The buyer will optimally set the bid quote B at the lowest level at which the subsequently arriving marginal seller is just willing to submit a MO. The buyer thus renders the seller indifferent between hitting the standing LO (by submitting a sell MO) or submitting a sell LO. The left-hand side of the equation presents the gain from a sell MO, given the standing bid quote B set by the buyer in the previous period. These gains, being expressed in cum-fee terms, take into account the expected post-trade fee when submitting a MO ($E[fee(MO)]$). The right-hand side of the equation is the expected gain of a sell LO submitted at the ask quote A . This is the endogenous execution probability of the LO (denoted γ), multiplied by the payoff upon execution of the order, corrected for the appropriate expected post-trade fee ($E(fee(LO))$). The interpretation of the second equation, which considers the seller's decision, is symmetric.

The general solution of the system of indifference equations (1) yields the following ask, bid quotes, and observed spread (* indicates equilibrium):

$$\begin{aligned} A^* &= \frac{V_h + \gamma(V_l + E[fee(LO)]) - E[fee(MO)]}{1 + \gamma}, \\ B^* &= \frac{V_l + \gamma(V_h - E[fee(LO)]) + E[fee(MO)]}{1 + \gamma}, \\ S^* &= \frac{(V_h - V_l)(1 - \gamma) - 2E[fee(MO)] + 2\gamma E[fee(LO)]}{1 + \gamma}. \end{aligned} \quad (2)$$

As we can see, γ and the expected fees on LO and MO are important determinants of ask, bid and observed spread S . Notice that the impact of $E[fee(LO)]$ depends on γ whereas this is not the case for $E[fee(MO)]$. The system of indifference equations

(1) and the resulting outcomes as shown in (2) differ between market design choices and fee structures. Before going to the different setups, we spell out the possible values that $E[fee(LO)]$ and $E[fee(MO)]$ can take. Consider first a LO. If the LO trader is targeting all counterparties, then the expected fee this trader will incur is $\left(\frac{N-1}{N}\right)c$, as this reflects the costs the PTSP incurs from the across broker trades. If the LO trader targets only *own* broker counterparties, then the fee is 0. For MO traders, the $E[fee(MO)]$ either equals $\left(\frac{N-1}{N}\right)c$ when it is uninformed about the counterparty's broker or when the PTSP charges uniform fees. The MO trader incurs a fee of c when it is informed about the counterparty using another broker and the PTSP charges trade-specific fees. Furthermore, in Equation (1), we have that $\gamma = 1/2$ when the “*all*” strategy prevails as any arriving interested counterparty will submit a MO. In contrast, $\gamma = 1/2N$ when the “*own*” strategy prevails as only interested counterparties of the *own* broker are willing to submit a MO against the standing quote.

5 Trading under Different Market Design Choices and Fee Structures

We now consider our three market design choices (anonymous trading, transparency, and optional disclosure of broker ID), where we each time discuss the different post-trade fee structures.

5.1 Anonymous Trading

With anonymous trading, a trader cannot condition her strategy on the standing or arriving counterparty. A LO trader therefore optimally chooses for the “*all*” strategy, independent of whether uniform or trade-specific fees are charged by the PTSP. Of course, if fees are too large, traders may abstain from trading.

With uniform fees, the execution probability γ equals 0.5, and the PTSP charges the equilibrium fee $\left(\frac{N-1}{N}\right)c$ to all transactions. By charging this fee to both legs of all transactions, the PTSP, on average, breaks even: internalized trades with zero marginal

costs cross-subsidize non-internalized trades with high marginal costs.

With trade-specific fees, an arriving trader does not know whether a standing limit order stems from a trader of her own broker or from another broker. Consequently, she is uncertain about the post-trade fees she will incur when submitting a market order. Similarly, at the time of submitting a LO, a trader does not know whether the arriving counterparty will use her own or another broker. In sum, the execution probability γ equals 0.5, and both the market order trader and limit order trader will account for expected post-trade fee $\left(\frac{N-1}{N}\right)c$, since a fraction $(N-1)/N$ of the trades is non-internalized.

The implication is that with anonymous trading, the equilibrium for trade-specific fees and uniform fees perfectly coincide. For both fee structures, traders abstain from submitting LOs when their expected profit becomes negative.

Define c_3 as¹²

$$c_3 \equiv \frac{N(V_h - V_l)}{2(N-1)}. \quad (3)$$

Proposition 1 shows the equilibrium under anonymous trading.

Proposition 1. *When trading is anonymous, results for the uniform and trade-specific fee structure of the PTSP coincide. In equilibrium:*

- for $c \leq c_3$, traders play the “all” strategy, resulting in the ask and bid quotes $A_{2,all}^*$ and $B_{2,all}^*$, observed spread $S_{2,all}^*$ and the cum-fee spread $S_{2,all}^{*cum}$, where

$$\begin{aligned} A_{2,all}^* &= \frac{2V_h + V_l}{3} - \frac{(N-1)}{3N}c, \quad B_{2,all}^* = \frac{V_h + 2V_l}{3} + \frac{(N-1)}{3N}c, \\ S_{2,all}^* &= \frac{V_h - V_l}{3} - \frac{2(N-1)}{3N}c, \quad S_{2,all}^{*cum} = S_{2,all}^* + 2\left(\frac{N-1}{N}\right)c \end{aligned} \quad (4)$$

- for $c > c_3$, traders refrain from trading.

Superscript $*$ refers to the equilibrium, and subscript *all* to the “all” strategy of traders. Subscript 2 is based upon the ranking of observed spreads (for a given c) we will encounter in the paper.

¹²We label it c_3 as later in the paper we will have two critical values of c that are smaller than c_3 .

5.2 Transparency

Transparency implies that broker-IDs are observable. With uniform fees, this information is irrelevant for a trader's strategy as the charged fee does not discriminate between possible counterparties (internalized or non-internalized). The outcome with uniform fees therefore coincides with the one presented in anonymous trading. With trade-specific fees, transparency of broker-ids allows LO traders to choose for either the “all” strategy, or the “own” strategy. Define c_1 as

$$c_1 \equiv \frac{2N(N-1)(V_h - V_l)}{(2N+1)(2N-1)}. \quad (5)$$

Proposition 2 shows the equilibrium.

Proposition 2. *When trading is transparent and the PTSP applies a trade-specific fee structure:*

- for $c \leq c_1$, traders play the “all” strategy, resulting in the ask and bid quotes $A_{1,all}^*$ and $B_{1,all}^*$, observed spread $S_{1,all}^*$ and the cum-fee spreads $S_{1,all}^{*cum,j' \neq j}$, $S_{1,all}^{*cum,j=j}$, and $S_{1,all}^{*cum,avg}$;

$$\begin{aligned} A_{1,all}^* &= \frac{2V_h + V_l}{3} - \frac{(N+1)}{3N}c, \quad B_{1,all}^* = \frac{V_h + 2V_l}{3} + \frac{(N+1)}{3N}c, \\ S_{1,all}^* &= \frac{V_h - V_l}{3} - \frac{2(N+1)}{3N}c, \quad S_{1,all}^{*cum,j=j} = S_{1,all}^*, \quad S_{1,all}^{*cum,j' \neq j} = S_{1,all}^* + 2c, \\ S_{1,all}^{*cum,avg} &= S_{1,all}^* + 2\frac{(N-1)}{N}c. \end{aligned} \quad (6)$$

- for $c > c_1$, traders play the “own” strategy. The ask and bid quotes $A_{3,own}^*$ and $B_{3,own}^*$, observed spread $S_{3,own}^*$, which is identical to the cum-fee spread $S_{3,own}^{*cum}$.

$$\begin{aligned} A_{3,own}^* &= \frac{2V_h N + V_l}{2N+1}, \quad B_{3,own}^* = \frac{V_h + 2V_l N}{2N+1}, \\ S_{3,own}^* &= \frac{2N-1}{2N+1}(V_h - V_l), \quad S_{3,own}^{*cum} = S_{3,own}^*. \end{aligned} \quad (7)$$

When the PTSP applies a uniform fee structure, the equilibrium coincides with the one under anonymous trading.

Superscript $*$ refers to the equilibrium, $j' \neq j$, $j = j$, and *avg* to other broker, same broker, and average, respectively. Subscript 1 and 3 refer to the ranking of the observed spreads, and subscript *all* and *own* to the “*all*” and “*own*” strategies of traders, respectively.

We now discuss in more detail the insights for trade-specific fees. We first analyze the “*all*” strategy. Traders at each broker j then set their quotes to keep any marginal other-broker trader $j' \neq j$ that pays a fee c indifferent between submitting a MO, or submitting a LO. Own-broker counterparties always accept this quote, entailing a zero fee. A submitted LO that also targets other-broker counterparties incurs the post-trade fee only with probability $\frac{(N-1)}{N}$, yielding an expected post-trade fee of $\frac{(N-1)}{N}c$.

The average cum-fee spread $S_{1,all}^{*cum,avg}$ is obtained by weighing each cum-fee spread by its average rate of occurrence. The $j' \neq j$ and $j = j$ possibilities, moreover, emphasize a key difference between post-trade fees and make/take fees as modeled in Colliard and Foucault (2012). Whereas make/take fees are charged to everyone in each trade, in our setting, post-trade fees are counterparty-dependent; non-internalized trades are charged a high fee c , internalized trades a zero fee.

Secondly, we examine the “*own*” strategy. Because only counterparties from the same broker are targeted, all trades are internalized and post-trade fees are equal to zero. The execution probability of a LO then reduces to $1/(2N)$, as it only relates to arrivals of own-broker counterparties.

5.3 Optional Disclosure of Broker ID

We now turn to the optional disclosure regime where a LO trader can choose whether or not to reveal her broker’s identity. Under uniform fees, information about broker affiliation is irrelevant as arriving market order traders pay the same fee irrespective the counterparty’s broker. We therefore have the same result as with anonymous trading.

For optional disclosure to have any effect, the PTSP needs to apply a trade-specific fee structure. Under this fee structure, the PTSP charges each transaction its marginal cost, zero for internalized and c for non-internalized trades. The endogenously chosen

disclosure hinges on c . Define c_2 as

$$c_2 \equiv \frac{N(V_h - V_l)}{2N + 1}. \quad (8)$$

Proposition 3 shows the equilibrium disclosure, trading strategies and resulting spreads.

Proposition 3. *When traders have optional disclosure, i.e., can choose whether to reveal their broker affiliation or not, and the PTSP applies a trade-specific fee structure:*

- *for $c \leq c_2$, traders will play the “all” strategy and conceal their broker affiliation, resulting in the ask and bid quotes $A_{2,all}^*$ and $B_{2,all}^*$, observed spread $S_{2,all}^*$ and cum-fee spread $S_{2,all}^{*cum}$;*

$$\begin{aligned} A_{2,all}^* &= \frac{2V_h + V_l}{3} - \frac{(N-1)}{3N}c, \quad B_{2,all}^* = \frac{V_h + 2V_l}{3} + \frac{(N-1)}{3N}c, \\ S_{2,all}^* &= \frac{V_h - V_l}{3} - \frac{2(N-1)}{3N}c, \quad S_{2,all}^{*cum} = S_{2,all}^* + 2\left(\frac{N-1}{N}\right)c \end{aligned} \quad (9)$$

- *for $c > c_2$, traders will play the “own” strategy and disclose their broker affiliation, resulting in the ask and bid quotes $A_{3,own}^*$ and $B_{3,own}^*$, observed spread $S_{3,own}^*$ and cum-fee spread $S_{3,own}^{*cum}$.*

$$\begin{aligned} A_{3,own}^* &= \frac{2V_h N + V_l}{2N + 1}, \quad B_{3,own}^* = \frac{V_h + 2V_l N}{2N + 1}, \\ S_{3,own}^* &= \frac{2N - 1}{2N + 1}(V_h - V_l), \quad S_{3,own}^{*cum} = S_{3,own}^*. \end{aligned} \quad (10)$$

When the PTSP applies a uniform fee structure, traders are indifferent between disclosing or revealing their broker affiliation. Results coincide with those under anonymous trading.

Superscript $*$ refers to the equilibrium, subscript 2 and 3 to the ranking of observed spreads, and subscript *all* and *own* to the “all” and “own” strategies of traders, respectively.

With trade-specific fees, LO traders opt to conceal their broker identity if the “all” strategy is played (i.e., for $c \leq c_2$). Quotes can be less aggressive and the probability

of execution is not compromised by remaining anonymous. A MO trader then accounts for the expected post-trade fee $\left(\frac{N-1}{N}\right)c$, such that this outcome perfectly coincides with uniform fees. At c_2 , LO traders would make identical profits either by playing the “*all*” strategy while preserving anonymity, or the “*own*” strategy when revealing broker affiliation.¹³ When $c > c_2$, traders will reveal their broker identity in order to opt for the “*own*” strategy. Remaining anonymous would force them to post highly liquid quotes resulting in low profits, whereas now they prefer illiquid quotes with a lower associated execution probability.

5.4 Summary

Table 1 summarizes our results so far.

Table 1: Market Design Choices and Fee Structures

Design \ Fee		Uniform		Trade-specific	
Anonymous	strategy	all if $c \leq c_3$	no trading if $c > c_3$	all if $c \leq c_3$	no trading if $c > c_3$
	spread	$S_{2,all}^*; S_{2,all}^{*cum}$	-	$S_{2,all}^*; S_{2,all}^{*cum}$	-
Transparent	strategy	all if $c \leq c_3$	no trading if $c > c_3$	all if $c \leq c_1$	own if $c > c_1$
	spread	$S_{2,all}^*; S_{2,all}^{*cum}$	-	$S_{1,all}^*; S_{1,all}^{*cum,avg}$	$S_{3,own}^*; S_{3,own}^{*cum}$
Optional Discl.	strategy	all if $c \leq c_3$	no trading if $c > c_3$	all if $c \leq c_2$	own if $c > c_2$
	spread	$S_{2,all}^*; S_{2,all}^{*cum}$	-	$S_{2,all}^*; S_{2,all}^{*cum}$	$S_{3,own}^*; S_{3,own}^{*cum}$

Note: c_1 , c_2 , and c_3 are defined in Equations (5), (8) and (3), respectively. The observed and cum-fee spreads are defined in Propositions 1 to 3.

It is important to note that $c_1 \leq c_2 \leq c_3$. This implies that the level of c where the optimal strategy changes from “*all*” to “*own*” is not the same under optional disclosure versus transparency.

¹³Technically, at c_2 the system of Indifference Equations (1) is simultaneously satisfied for the uniform fee’s “*all*” and trade-specific fee’s “*own*” strategy.

Corollary 1. *With trade-specific fees, under optional disclosure traders play the “all” strategy for higher levels of c , compared to transparency.*

In particular, when $c_1 < c \leq c_2$, with optional disclosure, LO traders prefer to conceal their broker affiliation and play the “all” strategy, even though revealing it would enable them to play the “own” strategy, as is the case for transparency for this range of c . Under optimal disclosure, staying anonymous enables them to enjoy high execution probabilities for relatively illiquid quotes, which are preferred to low execution probabilities with highly illiquid quotes. With transparency, where broker-IDs are revealed, LO traders need to quote more aggressively when playing the “all” strategy, implying that they prefer the *own* strategy already at lower levels of c . This observation will have important consequences for welfare discussed in our next Section.

6 Welfare

6.1 Welfare Definition and Inefficiencies

We define welfare as the sum of all agents’ expected utilities from trading.¹⁴ As the PTSP breaks even in expected terms, welfare coincides with trader welfare. Welfare is realized when trades produce trading gains net of post-trade costs. An internalized trade generates $V_h - V_l$, a non-internalized trade $V_h - V_l - 2c$. Expected welfare per period (W) is determined by the expected trading rate (TR , i.e., expected trading volume per period) and expected trade composition (i.e., the expected internalization rate IR , which is the fraction of trades that are expected to be internalized). The prices at which trades occur are not relevant to welfare, as they represent only a redistribution between buyer and seller. Generally, expected welfare per period then equals:

$$W = TR [V_h - V_l - 2c (1 - IR)], \quad (11)$$

¹⁴We refer to Degryse, Van Achter, and Wuyts (2009), Glosten (1998), Goettler et al. (2005), and Hollifield, Miller, Sandås, and Slive (2006) for a similar approach to quantifying welfare.

with TR the expected trading rate per period and $[V_h - V_l - 2c(1 - IR)]$ the expected net trading gains per trade.

To understand the drivers of welfare, we first consider the benchmark welfare per period, $W^{\max}$, obtainable in the absence of post-trade frictions, and taking as given the design of the limit order market with sequentially arriving traders. This could be seen as a world where post-trade costs are zero, implying that also market design about counterparty information is not relevant. The benchmark expected welfare per period in such a world becomes:

$$W^{bench} = \frac{1}{3} [V_h - V_l], \quad (12)$$

where the trading rate equals $\frac{1}{3}$ and the trading gains per trade are $V_h - V_l$.¹⁵ Yet, the social planner may not be able to reach W^{bench} due to inefficiencies stemming from post-trade frictions. The magnitude of these inefficiencies will guide the social planner in implementing its preferred post-trade fee structure and market design. This choice will depend on the post-trade cost c .

Before going into more detail on the potential inefficiencies the social planner faces, we first discuss in Proposition 4 the equilibrium trading and internalization rates per period for the different market design choices and strategies.

Proposition 4. *When trading is anonymous (A) and both with uniform and trade-specific fee structures*

$$TR_{A,all}^* = \frac{1}{3}, \quad IR_{A,all}^* = \frac{1}{N}. \quad (13)$$

When traders have optional disclosure or when trading is transparent ($k \in \{OD, T\}$), and the PTSP applies a trade-specific fee structure:

$$TR_{k,all}^* = \frac{1}{3}, \quad IR_{k,all}^* = \frac{1}{N}, \quad TR_{k,own}^* = \frac{1}{2N+1}, \quad IR_{k,own}^* = 1. \quad (14)$$

When traders have optional disclosure or when trading is transparent, and the PTSP

¹⁵In this benchmark, the “all” strategy is played. The trading rate of $\frac{1}{3}$ then results from the two possible actions an arriving trader can take. With probability $\frac{2}{3}$, an arriving trader will submit a LO (and no trade results) and with probability $\frac{1}{3}$ an arriving trader will submit a MO. More details on this computation are provided in Appendix B.

applies a uniform fee structure, TR and IR coincide with anonymous trading.

The trading and internalization rates depend on the trading strategy played (and therefore on the market design choice and post-trade fee structure). Since $N > 1$, the “*all*” strategy yields a larger trading rate (i.e., $TR_{k,all}^* > TR_{k,own}^*$) but a smaller internalization rate (i.e., $IR_{k,all}^* < IR_{k,own}^*$) than the “*own*” strategy. The “*all*” strategy has a larger trading rate as matches are generated between counterparties of all brokers. Its internalization rate is lower as the “*own*” strategy only results in internalized trades. In turn, the “*own*” strategy induces a larger trading rate of internalized trades (i.e., $\frac{1}{3N} = TR_{k,all}^* IR_{k,all}^* < \frac{1}{2N+1} = TR_{k,own}^* IR_{k,own}^*$). To provide intuition for the latter result, consider a case with transparent trading with a standing sell LO, and a buyer arriving at t through another broker. When the “*own*” strategy prevails, the other-broker buyer will forego the standing LO but submit a buy LO herself that may result in an internalized trade at $t + 1$. That is, the LOB is immediately refilled. In contrast, with the “*all*” strategy, the standing sell LO results in a non-internalized trade at t and an empty LOB at $t + 1$. This example highlights that each strategy entails different preemption effects. The “*own*” strategy completely preempts non-internalized trades, which happen with a trading rate of $\frac{N-1}{3N}$ under the “*all*” strategy. The “*all*” strategy leads to a lower trading rate of internalized trades compared to the “*own*” strategy (i.e., $\frac{-(N-1)}{3N(2N+1)} = \frac{1}{3N} - \frac{1}{2N+1}$). The “*all*” strategy therefore preempts some internalized trades.

Filling in Equations (13) and (14) into Equation (11) yields the expected welfare per period for the various market design choices ($k \in \{A, OD, T\}$) with the trade-specific fee structure when the *all* and *own* strategy are played, respectively.

$$W_{k,all}^* = \frac{1}{3} \left[V_h - V_l - 2c \left(\frac{N-1}{N} \right) \right] \text{ for market design choice } k \in \{A, OD, T\}$$

$$W_{k,own}^* = \frac{1}{2N+1} [V_h - V_l] \text{ for market design choice } k \in \{OD, T\}$$

For uniform fees, we have the same $W_{k,all}^*$ as above. Furthermore, when traders abstain from trading, welfare is zero.

The social planner may not be able to reach $W^{\max}$ due to inefficiencies stemming

from post-trade costs. Weighing off the economic relevance of these inefficiencies will guide the social planner in implementing its preferred market design and post-trade fee structure. The first inefficiency stems from non-internalized trades generating post-trade costs. This inefficiency occurs with the “*all*” strategy and equals $2c \left\lceil \frac{N-1}{3N} \right\rceil$; that is, the trading rate of non-internalized trades (i.e., $TR_{k,all}^* (1 - IR_{k,all}^*) = \frac{N-1}{3N}$) times the associated post-trade costs $2c$. We label this first inefficiency as “post-trade cost inefficiency”. The second inefficiency stems from completely foregoing non-internalized trades. This inefficiency occurs with the “*own*” strategy and equals $\frac{2(N-1)}{3(2N+1)} [V_h - V_l]$, i.e., the difference between the maximum trading rate and the trading rate of internalized trades (i.e., $\frac{2(N-1)}{3(2N+1)} = \frac{1}{3} - \frac{1}{2N+1}$), times the trading gains per trade $(V_h - V_l)$. We label this second inefficiency as “low trading rate inefficiency.”

6.2 Welfare Analysis

The social planner weighs off these two mutually exclusive inefficiencies when choosing the market design and post-trade fee structure that determines traders’ strategies. We have that both inefficiencies are equally important when

$$c = c_2 \equiv \frac{N(V_h - V_l)}{2N + 1}. \quad (15)$$

For $c \leq c_2$, the social planner prefers the “post-trade cost inefficiency” over the “low trading rate inefficiency.” The social planner then will choose a post-trade fee structure and market design that maximizes the trading rate but entails post-trade costs of non-internalized trades. In other words, it will choose a post-trade fee structure and market design where traders play the “*all*” strategy. When $c > c_2$, the “low trading rate inefficiency” is less important than the “post-trade cost inefficiency.” The planner then prefers to have internalized trades only and to forego non-internalized trades. It will choose a post-trade fee structure and market design where traders play the “*own*” strategy. Proposition 5 shows which strategies the social planner prefers for the entire range of c , as well as which market design choices and post-trade fee structures achieve

this.

Proposition 5. *Expected per-period welfare is highest:*

- *for low values of c (i.e., $c \leq c_1$) when the “all” strategy is played. All market design choices and post-trade fee structures achieve this;¹⁶*
- *for intermediate values of c (i.e., $c_1 < c \leq c_2$) when the “all” strategy is played. Anonymous trading and Optional Disclosure achieve this with both post-trade fee structures. Transparency achieves this only with the uniform fee structure;*
- *for high values of c (i.e., $c > c_2$) when the “own” strategy is played. Optional Disclosure and Transparency achieve this only with the trade-specific fee structure. Anonymous trading does not achieve this with either fee structure.*

Proposition 5 shows that the post-trade fee structure affects welfare. But there is also an interaction with the market design choice. In particular, optional disclosure combined with trade-specific fees is the only combination of market design choice and fee structure that is able to achieve the highest welfare (and thus achieve the preferences of the social planner) over the entire range of c . In other words, optional disclosure combined with trade specific fees implies that the resulting equilibrium strategies perfectly align with the socially preferred outcome. This alignment is not straightforward as every LO trader makes her decision individually and does not necessarily maximize the welfare of all traders jointly. We now discuss why the other combinations are not able to do this.

First, anonymous trading implies that the “all” strategy prevails for all post-trade cost levels, both with the uniform and trade-specific fee structures. Anonymous trading prevents the preemption of non-internalized trades and thus avoids the socially costly “post-trade cost inefficiency.” This is also the socially preferred outcome when $c \leq c_2$. When $c > c_2$, anonymity leads to the “post-trade cost inefficiency”. This is socially not the preferred outcome: the social planner prefers the “low trading rate inefficiency”

¹⁶The equivalence between trade-specific and uniform fees when $c \leq c_1$ may change when decreasing marginal utility is assumed. The trade-specific fee structure then becomes socially more desirable, as it yields more equal sharing of net trading gains. However, it also entails greater variability in ex-post net trading gains than uniform fees. Thus, a novel welfare trade-off arises between the level and variability of net trading gains.

over the “post-trade cost inefficiency” when $c > c_2$. Implementing an anonymous setting does not allow traders to individually internalize the lower post-trade cost associated with within-broker trades. Remarkably, allowing a LO trader to endogenously reveal her broker identity induces that traders implement the strategy leading to the lowest inefficiency. The optional disclosure combined with trade-specific fees therefore induces perfect alignment between private and social incentives.

Second, uniform fees make the market design choice redundant. LO traders then always opt for the “*all*” strategy implying that the “post-trade cost inefficiency” is realized. From a social welfare perspective, this is not optimal for $c > c_2$.

Third, with transparent trading and trade-specific fees, the LO trader switches from the “*all*” to the “*own*” strategy at c_1 , whereas the socially-preferred cut-off point c_2 is larger than c_1 . Traders then individually fail to fully internalize the negative impact on welfare stemming from the large “low trading rate inefficiency.” Hence, with transparent trading, the trade-specific fee structure underperforms for the intermediate cost range (i.e., $c_1 < c \leq c_2$) because it induces traders to implement the “*own*” strategy and not the socially-preferred “*all*” strategy. In that cost range, traders’ individual behavior results in the “low trading rate inefficiency” that prevents socially beneficial non-internalized orders to take place. Traders therefore fail to fully internalize the negative impacts this “*own*” strategy generates on the preemption of non-internalized trades. LO traders are put off by the substantial post-trade fee c entailed by non-internalized trades and corresponding quote aggressiveness needed to attract other-broker traders. Thus, they individually prefer to avoid non-internalized trades, whereas the social planner prefers the “post-trade cost inefficiency” over the “low trading rate inefficiency.”

Our analysis thus shows that allowing for transparency improves social welfare outcomes when it is important to induce the *own* equilibrium. One may question why in reality we see so little transparency and which market failure makes that our proposed market design differs from observed ones. One explanation for the limited observed transparency may have to do with asymmetric information (which we do not model) or the behavior of for-profit exchanges. If exchanges have incentives to maximize trading

volume, they may prevent the *own* equilibrium from occurring. Exchanges may do so by limiting transparency (see subsection 9.1 for further discussion). Regulatory intervention then is required.

7 Spreads and Asymmetric Fee Allocation

7.1 Spread Determinants

We now discuss how observed and cum-fee spreads are affected by fees and number of brokers across the various market design choices. Corollary 2 summarizes our findings.

Corollary 2. *Observed and cum-fee spreads behave as follows:*

- *when the “all” strategy prevails:*
 - *observed spreads decrease and cum-fee spreads increase in post-trade fees ,*
 - *observed spreads $S_{1,all}^*$ are smaller than observed spreads $S_{2,all}^*$,*
 - *$S_{1,all}^*$ decreases whereas $S_{2,all}^*$ increases in the number of brokers N . Cum-fee spreads increase with N .*
- *under the “own” strategy:*
 - *the observed and cum-fee spreads coincide and are independent of the post-trade fee*
 - *observed and cum-fee spreads increase in N .*

When the “all” strategy prevails, a larger post-trade fee induces more aggressive quote-setting behavior as a consequence of reduced trading gains. LO traders submit more aggressive quotes to induce the counterparty to submit a MO. That is, it is as if the counterparty now has a lower willingness to trade resulting from the increase in the (expected) post-trade fee. The resulting observed spreads decrease in the fee. In contrast, a larger fee induces a higher cum-fee spread. Hence, despite the availability of more liquid quotes, higher post-trade fees erode MO payoffs. Also the market design

choice for transparency and post-trade fee structure interact: given trade-specific fees, observed spreads are lower with transparency than with optional disclosure or anonymity. With transparency, a LO trader needs to keep another broker trading indifferent between hitting her LO and submitting a MO. Hitting that LO implies a higher fee than with optional disclosure or anonymity inducing the LO submitter to post more liquid quotes. $S_{1,all}^*$ decreases in the number of brokers N . The average fee charged by the PTSP increases with N as less trades will be internalized. As a result, observed quotes, set by LOs, will have to be more liquid to convince a counterparty to submit a MO. In contrast, $S_{2,all}^*$ increases in N . Intuitively, a higher N makes the outside option of an arriving trader less attractive as she is more likely to incur a post-trade fee. The LO trader therefore can set less attractive quotes and thus observed spreads increase.

When the “own” strategy prevails, fees are irrelevant for the spreads as by assumption no fees apply on internalized trades. Spreads increase in N as the outside option of an arriving trader to submit a LO herself becomes less attractive.

Our analysis on spreads allows to answer whether welfare is highest when liquidity is highest. Our main result is established in the following proposition.

Proposition 6. *A comparison of fee structures and market design choices shows that, for given c , the lowest average cum-fee spread yields the highest welfare. In contrast, neither a low observed spread nor high trading volume necessarily produce the highest welfare.*

The lowest observed spread does not always yield the highest welfare. For instance, when $c > c_2$, observed spreads are smallest under uniform fees (with any market design), whereas trade-specific fees (with optional disclosure or transparency) yield the highest welfare. The intuition is that uniform fees induce non-internalized trades to still occur, cross-subsidized by cheaper internalized trades. From a welfare perspective, non-internalized trades should be prevented as the “post-trade cost inefficiency” is more important than the “low trading rate inefficiency.” Minimizing cum-fee spreads, however, always yields the highest welfare. For $c > c_2$, the average cum-fee spread is lowest and welfare highest under trade-specific fees (with optional disclosure or transparency) when

the “*own*” strategy is played. For $c \leq c_2$, welfare is highest when the “*all*” strategy is played. The average cum-fee spread is lowest under trade-specific fees for $c \leq c_1$ (with optional disclosure or transparency), and under uniform fees for $c_1 < c \leq c_2$, and both fee structures deliver the “*all*” strategy for this range of c . Intuitively, traders who submit MOs on the basis of the cum-fee spread explicitly take post-trade fees into account. In our setting, post-trade fees are cost-based, and decisions that minimize the average cum-fee spread also yield the highest welfare.

7.2 Asymmetric Fee Allocation between LO and MO

Our main analysis assumed that, conditional on a trade, LOs and MOs are charged an identical post-trade fee. We now study the following economic question: would an unequal split of post-trade fees between LO and MOs allow to obtain the socially preferred outcome? This economic question relates to the make/take fee literature where makers and takers are provided different incentives by splitting the trading fee among them. Colliard and Foucault (2012) for example show that such a split is neutral as it does not affect trading incentives. Novel to the literature, we show non-neutrality holds under asymmetric split when fees are trade-specific and transparency.

We study the impact of an unequal split on the equilibrium outcomes for each market design and on the welfare analysis. We allocate a fraction α (with $0 \leq \alpha \leq 1$) of the total expected post-trade fee to the MO and the complementary fraction $1 - \alpha$ to the LO. Our main analysis then coincides with $\alpha = 0.5$.¹⁷ With unequal splits, a uniform fee structure implies there is no fee differential *within* LOs and *within* MOs. The fee however is strictly different *between* LOs and MOs when $\alpha \neq 0.5$. Trade-specific fees further imply that *within* LOs and MOs, the fee depends upon whether it is a within-broker or across-broker trade.

With asymmetric fee allocation, the system of indifference equations (1) applies where $E(\text{fee}(\text{MO}))$ and $E(\text{fee}(\text{LO}))$ depend on how the post-trade costs are split. For example, with transparency and uniform fees, the total post-trade fee equals $2 \frac{N-1}{N} c$,

¹⁷With make/take fees, takers typically subsidize makers. In our setting this would imply that $\alpha > 1$, i.e., limit orders receive a fee bonus fully compensated by the market orders.

of which a fraction α is charged to the MO and $1-\alpha$ to the LO. This implies that $fee(MO) = 2\alpha\frac{N-1}{N}c$ and $fee(LO) = 2(1-\alpha)\frac{N-1}{N}c$. In Appendix C we provide details on the values of $E(fee(MO))$ and $E(fee(LO))$ for all possible cases, and derive the equilibria for all settings and the associated observed and cum-fee spreads for unequal splits of post-trade costs. We summarize our main findings in Table 2.

Table 2: Unequal split of post-trade fees

Design \ Fee		Uniform		Trade-specific	
Anonymous	strategy	all if $c \leq c_3$	no trading if $c > c_3$	all if $c \leq c_3$	no trading if $c > c_3$
	spread	$S_{\alpha 2, all}^*; S_{2, all}^{*cum}$	-	$S_{\alpha 2, all}^*; S_{2, all}^{*cum}$	-
Transparent	strategy	all if $c \leq c_3$	no trading if $c > c_3$	all if $c \leq c_\alpha$	own if $c > c_\alpha$
	spread	$S_{\alpha 2, all}^*; S_{2, all}^{*cum}$	-	$S_{\alpha 1, all}^*; S_{\alpha 1, all}^{*cum}$	$S_{3, own}^*; S_{3, own}^{*cum}$
Optional Discl.	strategy	all if $c \leq c_3$	no trading if $c > c_3$	all if $c \leq c_2$	own if $c > c_2$
	spread	$S_{\alpha 2, all}^*; S_{2, all}^{*cum}$	-	$S_{\alpha 2, all}^*; S_{2, all}^{*cum}$	$S_{3, own}^*; S_{3, own}^{*cum}$

Note: c_2 , and c_3 are defined above, and $c_\alpha = \frac{N(V_h - V_l)(N-1)}{(N+\alpha-1)(2N+1)}$. $S_{\alpha 2, all}^* = \frac{N(V_h - V_l) - 4c(N-1)(3\alpha-1)}{3N}$; $S_{\alpha 1, all}^* = \frac{N(V_h - V_l) - 4c(N(3\alpha-1) - \alpha + 1)}{3N}c$; $S_{\alpha 1, all}^{*cum} = \frac{4c(N-2\alpha-1) + N(V_h - V_l)}{3N}$. All other observed and cum-fee spreads are defined in Propositions 1 to 3.

What are the implications from an uneven allocation in the post-trade fee on equilibrium outcomes and welfare? First, for uniform fees (independent of market design) and trade-specific fees (optional disclosure and anonymity), a higher allocation to the MO, i.e., a higher α , leads to a narrower observed spread $S_{\alpha 2, all}^*$ but leaves the cum-fee spread $S_{2, all}^{*cum}$ unaffected. The expected trading gains of a LO and MO thus remain unaffected by modifying the split. This result is reminiscent of Colliard and Foucault (2012) who show that only the *size* of the total trading fee is important but not its allocation between makers and takers.

Second, with transparency and trade-specific fees, both the observed ($S_{\alpha 1, all}^*$) and cum-fee ($S_{\alpha 1, all}^{*cum}$) spread decrease in α . The fact that the cum-fee spread is affected by modifying the split shows its non-neutrality and distinguishes our approach from the

make/take fee literature. To understand this, consider the following thought experiment: increase α such that the fee for the MO trader increases by 1 cent and the fee for the LO decreases by 1 cent, leaving the total fee unchanged. To keep attracting MOs coming from a different broker, the LO submitter (i.e., the maker) needs to decrease the bid price B by 1 cent, to compensate the increase that a taker faces. With regular make/take fees the maker would earn 1 cent less from his order, but this would be exactly compensated by the decrease in his fee, hence the neutrality. In our setting, this does not apply, because the fee is paid by the LO trader only with probability $(N-1)/N$, so that the decrease in fee gives the maker less than 1 cent. Hence, the split is not neutral. The maker therefore only wants to decrease B with less than one cent leading to an increase in the cum-fee spread. Counterintuitively, LO traders actually benefit from lower take fees more than from lower make fees. The non-neutrality result holds when fees are counterparty dependent and information about the counterparty is revealed.

Third, allocating the entire fee on the LO (i.e., $\alpha = 0$) implies that the observed and cum-fee spreads coincide.

Fourth, with optional disclosure, an uneven split does not impact on the choice between the “all” or “own” strategy. Recall that with optional disclosure, as long as the “all” strategy is preferred, a LO trader chooses to remain anonymous. With anonymity, the split does not affect the profitability of a LO, as a change in fee is completely compensated by a similar change in the quote. An unequal split also leaves the “own” strategy unaffected since there are no fees charged. As a result, an uneven split has no effect on a LO trader’s choice between the “all” and “own” strategy. In contrast, with transparency and trade-specific fees, a split more tilted towards a MO (i.e., a higher α) induces LO traders to implement the “own” strategy for lower values of c as c_α decreases in α . Intuitively, and as discussed above, a higher α reduces the profits of LO traders under the *all* strategy. Therefore, for some values of c , it now becomes attractive to adopt the “own” strategy.

Finally, optional disclosure combined with trade-specific fees still coincides with the social planner’s preferred outcome independent of the breakdown in total post-trade

fee. Intuitively, the breakdown does not affect a LO-trader's profit and also the social planner does not care about how post-trade costs are allocated across LO and MO trader. Therefore the outcome is independent from α . However, with trade-specific fees and transparency, the social planner could also reach its preferred outcome by allocating the complete post-trade fee to the LO trader, i.e., $\alpha = 0$. Put differently, imposing no fee on the MO trader allows LO traders to price less aggressively and make greater expected profits. Compared to the equal split, $\alpha = 0$ induces LO traders to play the *all* strategy for values of c in between c_1 and c_2 . In this way, setting $\alpha = 0$ allows to minimize the two inefficiencies. A very important practical advantage to the social planner is that this allocation gives an additional tool to achieve the optimal outcome when the planner does not have complete information on the magnitude of c . We therefore establish the next corollary.

Corollary 3. *With trade-specific fees and transparency, a social planner can maximize welfare by allocating the entire post-trade cost on LOs.*

8 Empirical and Regulatory Implications

Our analysis urges that the entire value chain of trading and post-trading is accounted for when making assessments about financial markets. In particular, the market design choices and the post-trade fee structure cannot be evaluated independently from each other.

Market design and the post-trade fee structure determine (i) trading volume and (ii) whether a market exists or not. Trading volume is determined by traders' strategies. Our analysis shows that these strategies hinge on both the market design about counterparty information (i.e., broker-IDs) and the post-trade fee structure. Keeping the underlying traders' intentions fixed, we show that differences in trading volume across countries or trading venues may stem from differences in market design and the way how post-trade costs are allocated. Also, allowing for differences in fees depending on buyer-seller matches may create a market that with uniform fees would break down. The post-trade

fee structure in combination with market design may thus explain why a market exists in one environment or breaks down in another.

Furthermore, our analysis provides new insights on the components of the bid-ask spread. Traditionally, the literature has identified three reasons for the existence of a bid-ask spread: (i) order processing costs (including rents from imperfect competition), (ii) inventory costs, and (iii) asymmetric information (Huang and Stoll (1997)). Our baseline setup only features order processing costs as the spread results from imperfect competition and strategic behavior.¹⁸ In line with other commissions or fees, post-trade fees are naturally incorporated in this order processing component and the post-trade fee structure determines its size. As before, average cum-fee spreads are required for an unbiased identification of the order processing cost component of the bid-ask spread. Using observed spreads leads to systematic underestimation of the order processing cost component. To see this, recall that a post-trade fee increase narrows the observed bid-ask spread in case the “*all*” strategy is played. This incorrectly would be interpreted as a cost decrease whereas there was an actual cost increase.

Our analysis also generates interesting policy implications for the organization of both the trade and post-trade infrastructure. First, regulators currently employ observed, not cum-fee, quotes for rules on best execution. For example, the SEC’s “trade-through” and “trade-at” rules forbid execution at quoted prices worse than the best observed quotes. We show such rules to yield best execution with uniform, but not necessarily with trade-specific fees. For example, a trade at an aggressive quoted price with a high post-trading fee is consistent with best-execution policy. However, such a trade could be detrimental to net trading gains compared to a within-broker trade at an inferior quote but without post-trade fee. This would happen when the quote difference stemming from the tick size for example is lower than the difference in post-trade fees. This is not unlikely to happen when comparing on-exchange versus off-exchange trading. Menkveld (2013), for example, shows that post-trade fees are sizeable as cum-fee spreads (stemming from

¹⁸The LO trader faces only indirect intertemporal competition from traders at the same market side through the outside option of the incoming counterparty. This reason for the spread is in line with models such as Foucault (1999) and Van Achter (2010).

post-trade fees) are about 20% higher than observed spreads for his sample of Dutch stocks.

Second, at the post-trade infrastructure level, the pre-netting prohibition imposed in the United States reduces overall netting efficiency. In our model, this could be captured by an increase in N , which is expected to decrease both observed and cum-fee liquidity as trade-specific fees are charged in the United States. LO traders quote less aggressively, as an arriving trader now has a higher expected fee (when the “*all*” strategy prevails), or a lower likelihood of encountering a same-broker trader (when the “*own*” strategy prevails). Interestingly, with uniform fees, lower netting efficiency leads to improved observed (but lowered cum-fee) liquidity. Third, in Europe, the TARGET2-securities (T2S) settlement platform applies uniform fees, which, in practice, might disincentivize the routing of cheaper (domestic) internally settleable trades to that platform. Our analysis indicates that this could be resolved by imposing a trade-specific fee structure.

9 Robustness and Extensions

9.1 Multiple Brokers and Incentives of Exchanges

Our model assumes that there are $N > 1$ brokers and that traders are affiliated to one broker only. An objection one could make is that traders could be affiliated with multiple (if not all) brokers in order to improve their outcomes. Of course, such a strategy may entail (unmodelled) access costs of establishing multiple relations with brokers. Furthermore, as trading opportunities are short-lived, the time needed to transfer securities or cash from one broker to another when seeing a LO may make such a strategy practically impossible for most traders. That is switching costs would continue to exist even when affiliated to multiple brokers. Also, $N = 1$ would remove post-trading costs and the associated frictions. However, we think $N = 1$ is unrealistic as we clearly want to make the distinction between cheaper within-broker trades and more expensive across-broker trades. With $N = 1$, this distinction is removed; a monopolistic broker may generate other inefficiencies that lead to post-trade costs as well as the potential to exploit its

market power. Our modeling choice implicitly assumes that with $N = 2$, the inefficiencies are minimized and competition between brokers prevents market power to be an issue. We leave for future research the endogenization of the brokers' market structure.

Our model made abstraction from the incentives of exchanges. These incentives may explain why market designs that appear more efficient in our analysis (such as optional disclosure) are less observed in reality, and why regulatory intervention may be warranted to implement optional disclosure. For-profit exchanges typically only charge trading fees when trades occur. Exchanges may aim to maximize the trading rate by either adjusting the market design in terms of transparency, or vertically integrating with the PTSP. Maximizing the trading rate happens when the *all*-strategy prevails as long as the market does not break down (i.e., when $c \leq c_3$). Exchanges then may prevent the “own” equilibrium from happening as in the “own” equilibrium the trading volume is by definition lower than with the “all” equilibrium, since traders with different brokers do not trade with each other. For $c \in [c_2, c_3]$, an implication is that when the PTSP uses the trade-specific fee structure, the exchange would choose an anonymous market design to prevent the *own* equilibrium from occurring. Regulatory intervention requiring transparency then might be required. In contrast, when $c > c_3$, an exchange prefers the *own* strategy as otherwise the market breaks down. An exchange then would prefer optional disclosure or transparency when $c > c_3$, and the PTSP to implement trade-specific fees. This again shows that trading and post-trading cannot be analyzed independently.

The incentives of exchanges further influences how the industrial organization of trading and post-trading could be organized. Vertical integration of an exchange and a PTSP could simultaneously decide on both market design and the fee structure. In this way, such a vertically integrated silo could easily reach the objective of maximizing the trading rate. We leave the topic of competition between separated structures and vertical silos, and the optimal industrial organization as a topic for further research (see Cespa and Foucault (2014) for an analysis on how the selling by exchanges of both trading services and price information affects price discovery and the distribution of

gains from trade in an asset market, or Juranek and Walz (2020) for an analysis of vertical versus horizontal integration of PTSP and exchanges when vertical integration has a cost advantage for exchange trades but a disadvantage for OTC trades).

9.2 Blockchain

Blockchain (i.e., a ledger of transactions maintained in a public, distributed, commonly accessible network) may reshape the post-trading industry. Blockchain has a number of features that relate to our model. First, Blockchain offer traders the choice to reveal their counterparty identity. In this way, it makes optional disclosure possible (see Malinova and Park (2016) or Khapko and Zoican (2020)). Second, the distributed ledger technology in Blockchain could lower post-trade costs by eliminating the need for multiple records of a trade and enhancing clearing and settlement efficiency.¹⁹

With uniform fees, a decrease in cost has two effects. First, they deteriorate observed liquidity but increase cum-fee liquidity. Second, when settlement costs were too high in the absence of Blockchain, a lowering of these costs may lead to the creation of a market. By allowing selective disclosure of counterparty information and in an environment of trade-specific fees, Blockchain allows to implement the socially preferred outcome assuming that post-trade fees are competitively determined. Intuitively, revealing counterparty information allows to implement the *own* equilibrium. Hiding information implements the *all* equilibrium. Taken together, this allows to implement the socially preferred outcome.

10 Conclusion

Post-trade costs are economically sizeable and depend on the type of transaction. That is, they are lower for internalized than for non-internalized trades. In practice, post-trade infrastructures allocate these costs across trades via different fee schedules, and

¹⁹For example, the Australian Stock Exchange (ASX) plans to replace the central depository with a blockchain solution at the end of 2020. According to ASX, Blockchain should provide greater market efficiencies through better record keeping, automated reconciliation, more timely transactions and enhanced data quality.

the charged fees affect trading decisions. We capture this setting by integrating two commonly-used post-trade fee schedules into a trading model with different market design choices about counterparty information. A uniform fee structure that charges all trades the same fee, and a trade-specific fee structure that charges internalized trades a lower fee. In doing so, we construct the first model to integrate the trade and post-trade phases in order to show how post-trade fees combined with market design choices affect traders' strategies and, thus, market quality and ultimately welfare.

Our model exposes the following key trade-off faced by traders providing liquidity under trade-specific fees when having counterparty information: target all potential counterparties by setting a liquid quote that likely entails a high post-trade fee, or accept reduced execution probabilities by setting an illiquid quote that attracts only own-broker counterparties (not subject to the fee). The existence of this trade-off thus depends on the market design about counterparty information and the outcome of this trade-off essentially hinges on the cost differential between internalized and non-internalized trades. With counterparty information available, traders already restrain themselves to within-broker trading at intermediate cost levels, although from a welfare perspective this switch should only occur at higher cost levels. The reasoning is that traders do not internalize the welfare-improving trading gains from across-broker trades. This inefficiency does not arise with uniform fees or anonymity, as traders are then unable to target only own-broker counterparties. At higher post-trade costs, in contrast to uniform fees, trade-specific fees combined with information about counterparties yield the socially preferred outcome. The reason is that trade-specific fees prevent costly across-broker trades from taking place. Regulators can obtain the socially preferred outcome, and thus minimize inefficiencies, by tailoring the market design and the implemented fee structure to the post-trade cost level, or by allowing traders to decide whether to reveal their broker affiliation under the trade-specific fee structure. Our analysis thus shows that optional disclosure in combination with trade-specific fees improves social welfare outcomes when it is important to induce the *own* equilibrium. Blockchain can make such a market design possible.

Our theoretical study has a number of limitations which may explain why the market designs that appear more efficient in the paper are not always observed in reality. One such limitation is that the behavior of exchanges is not explicitly incorporated. Our analysis shows that exchanges may not necessarily choose the socially optimal design and hence regulatory scrutiny is potentially warranted. For example, exchanges may deter low trading volume outcomes by preventing optimal disclosure. While our model is a first step towards integrating market design choices about trading and post-trading features, we do not consider the possible effects of competition between vertical integrated players and stand-alone trading venues or post-trade-service providers. We leave a deeper analysis of such interactions for future research.

A Appendix A: Proofs

Proof of Proposition 1. We first discuss how we obtain the candidate equilibrium as displayed in Proposition 1. When trading is anonymous, it is argued in the main text that results for the uniform and trade-specific fee structure coincide. Since there is no information about broker-ID, traders always address all potential counterparties and the “all” strategy will prevail. Hence, the execution probability $\gamma = 0.5$ and the fee (when uniform fees apply) and expected post-trade fee (when trade-specific fees apply) is

$$E(\text{fee}(MO)) = E(\text{fee}(LO)) = \left(\frac{N-1}{N}\right)c.$$

The equilibrium bid and ask quotes as shown in Equation (4) follow immediately from including the above mentioned values for γ and fees, and solving the system of indifference conditions (1), or filling in Equation (2). The observed and cumulative spreads are obtained from the difference in ask and bid quotes, and adding $2((N-1)/N)c$, respectively.

As a second step, we investigate whether this candidate equilibrium indeed constitutes an equilibrium by showing that all possible deviations yield lower profits. We argue that any other bid or ask quote announced by a LO trader are suboptimal. First, suppose a buy LO trader would post any $B > B_{2,all}^*$. Such a quote does not increase the likelihood of trading (which equals $1/2$) but only yields an inferior price, and thus profit. Second and similarly, $B < B_{2,all}^*$ is not optimal as it would yield zero profits: arriving traders then prefer to submit a LO themselves rather than hitting the worse priced LO. By symmetry, the same applies for a sell LO. Given anonymity, attracting only *own* broker counterparties is not possible.

We next show why a trader will not post a quote when $c > c_3$. A LO trader will only play the “all” strategy if it yields a positive expected profit. The expected payoff of a buyer linked to any broker submitting a limit order with price $B_{2,all}^*$ is:

$$\pi_{2,all} = \frac{1}{2} \left[V_h - \left(\frac{V_h + 2V_l}{3} + \left(\frac{N-1}{3N} \right) c \right) - \left(\frac{N-1}{N} \right) c \right] \quad (16)$$

Then, $\pi_{2,all} < 0$ if $c > \frac{N(V_h - V_l)}{2(N-1)} \equiv c_3$. □

Proof of Proposition 2. We first focus on the first part of the proposition, i.e., the trade-specific fee structure.

The first bullet point deals with the “*all*” equilibrium. We proceed in three steps:

1. The “*all*” equilibrium is characterized by the quotes $A_{1,all}^*$ and $B_{1,all}^*$ that solve the system of equations (1) where $E(fee(LO)) = \left(\frac{N-1}{N}\right)c$ (as the LO can hit all type of broker-ids); $E(fee(MO)) = c$ (as the MO should also attract other broker-ids), and $\gamma = 1/2$. This results immediately in the quotes $A_{1,all}^*$ and $B_{1,all}^*$ and the observed spread $S_{1,all}^*$. The cum-fee spreads $S_{1,all}^{*cum,j' \neq j}$, $S_{1,all}^{*cum,j=j}$, and $S_{1,all}^{*cum,avg}$ are obtained after adding the relevant (weighted) post-trade fees.
2. The equilibrium payoff of the submitter of a LO then becomes $\gamma[A - V_l - E(fee(LO))] = \frac{V_h - V_l}{3} - c \frac{2N-1}{3N} = \Pi^{all}$.
3. We now need to check for deviations. The submitter of a buy LO could deviate and offer a bid B' such that $B' - V_l - fee = \Pi^{all}$, with $fee = 0$. This offer gives such a deviator a payoff equal to $\Pi' = \gamma[V_h - B' - fee]$, with $\gamma = (1/(2N))$, and $fee = 0$. We have that $\Pi' = \frac{V_h - V_l}{3} + c \frac{2N-1}{6N^2}$. The profit $\Pi^{all} \geq \Pi'$ whenever $c \leq \frac{2N(N-1)(V_h - V_l)}{(2N+1)(2N-1)} \equiv c_1$.

The second bullet point in Proposition 2 focuses on the “*own*” strategy. We again make the following reasoning in 3 steps:

1. The “*own*” equilibrium is characterized by the quotes $A_{3,own}^*$ and $B_{3,own}^*$ that solve the system of equations (1) where $E(fee(LO)) = 0$ (as the broker-id is own), $E(fee(MO)) = 0$ (as the broker-id=own), and $\gamma = (1/(2N))$. This results immediately in the quotes $A_{3,own}^*$ and $B_{3,own}^*$ and the observed spread $S_{3,own}^*$ and cum-fee spread $S_{3,own}^{*cum}$.
2. The equilibrium payoff of the submitter of a LO then becomes $\gamma[A - V_l - E(fee(LO))] = \frac{V_h - V_l}{2N+1} = \Pi^{own}$.
3. We now need to check for deviations. The submitter of a buy LO could deviate and offer a bid $B'' - V_l - fee(other) = \Pi^{own}$ (where $fee(other) = c$). This offer

would give such a deviator a payoff equal to $\Pi'' = \gamma[V_h - B'' - fee(all)]$ (with $\gamma = 1/2$ and $fee(all) = ((N - 1)/N)c$). We have that $\Pi'' = \frac{N(V_h - V_l)}{2N+1} - c \frac{2N-1}{2N}$. The profit $\Pi^{own} \geq \Pi''$ whenever $c \geq \frac{2N(N-1)(V_h - V_l)}{(2N+1)(2N-1)} \equiv c_1$.

This shows that the “*all*” strategy prevails whenever $c \leq \frac{2N(N-1)(V_h - V_l)}{(2N+1)(2N-1)} \equiv c_1$. Otherwise the “*own*” strategy prevails.

The last part of the proposition (on uniform fees) has been proven already in Proposition 1 as results for uniform fees coincide with anonymity. This completes the proof. $\square$

Proof of Proposition 3. We first deal with the case where the PTSP implements a trade-specific structure. The equilibrium quotes when the LO trader opts for anonymity have been derived in Proposition 1. Similarly, the equilibrium quotes when the LO trader reveals her identity have been shown in Proposition 2, where LO traders optimally select the “*all*” strategy for $c \leq c_1$, and the “*own*” strategy otherwise.

We now need to show that the equilibrium is the one as displayed in Proposition 3 by showing that deviations are not profitable. The additional complication here is that a LO submitter not only decides on whom to target (*all*, or *own*) but also on whether to *conceal* or *reveal* its broker-id.

With the “*all*” strategy, a LO trader always prefers “*all, conceal*” over “*all, reveal*” since the profits of concealing while playing the “*all*” strategy dominates the one with revealing. This result is immediate as the execution probability γ is the same but when revealing the LO submitter gets a worse price. Therefore, limit order traders always deliberately opt not to reveal their identity when they choose the “*all*” strategy.

We now proceed to show the first bullet point related to when the “*all*” strategy is chosen. We proceed in three steps:

1. The equilibrium payoff of a LO trader when concealing its broker-id becomes $\gamma[A - V_l - E(fee(LO))] = (1/2)([A - V_l - ((N - 1)/N)c] - \frac{V_h - V_l}{3} - 2c \frac{N-1}{3N}) = \Pi^{all, conceal}$.
2. We now need to check for deviations. The submitter of a buy LO could deviate and offer a bid B''' such that $B''' - V_l - fee = \Pi^{all, conceal}$, with $fee = 0$. This offer gives

a deviator a payoff equal to $\Pi''' = \gamma[V_h - B''' - \text{fee}]$, with $\gamma = (1/(2N))$, and $\text{fee} = 0$.

We have that $\Pi''' = \frac{V_h - V_l}{3N} + c \frac{N-1}{3N^2}$.

3. The profit $\Pi^{all,conceal} \geq \Pi'''$ whenever $c \leq c_2 = \frac{N(V_h - V_l)}{2N+1}$.

We now turn to the second bullet point of the proposition related to when the “own” strategy is chosen. When the LO trader opts for “own, reveal”, we need to study whether the LO trader does not want to deviate towards “all, conceal” as that deviation gives larger profits than “all, reveal”, and “own, conceal”. We proceed in three steps:

1. The equilibrium payoff of the submitter of a LO with “own, reveal” becomes

$$\frac{1}{2N} [A - V_l] = \frac{V_h - V_l}{2N+1} = \Pi^{own, reveal}.$$

2. The submitter of a buy LO could deviate and offer a bid B'''' satisfying: $B'''' - V_l - \text{fee}(all) = \Pi^{own, reveal}$ (with $\text{fee}(all) = ((N-1)/N)c$. Notice that the deviator hides her identity such that the belief of the arriving investor could be anything; among those, we select the one with symmetric beliefs which given the symmetry of our model seems a reasonable choice. This offer B'''' would give the deviator a payoff equal to $\Pi'''' = (1/2)[V_h - B'''' - ((N-1)/N)c] = \frac{N(V_h - V_l)}{2N+1} - \frac{N-1}{N}c$.

3. We have that $\Pi^{own, reveal} \geq \Pi''''$ when $c \geq c_2$.

In other words, when the PTSP applies a trade-specific fee structure, for $c \leq c_2$, LO traders choose the “all” strategy and conceal. Otherwise, the “own” strategy and they reveal their broker-ID.

When the PTSP applies a uniform fee structure, concealing or revealing of broker-id is irrelevant as it does not influence the fee charged. Results are therefore those as under anonymity. □

Proof of Corollary 1. The proof is immediate since $c_1 \leq c_2$. □

Proof of Corollary 2. The proof is an immediate result from Proposition 1, 2, and 3. In particular, taking the partial derivative of the observed and cum-fee spreads w.r.t. c and N yields the statement in this corollary. □

Proof of Proposition 4. 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 market design choice $k \in \{AnonymousTrading, OptionalDisclosure, Transparency\}$ and fee structure of the PTSP $f \in \{Uniform, Trade - Specific\}$, denote the stationary probabilities of each of these three possible states under a given strategy $s \in \{all, own\}$ as $\varphi_{k,s}^f = \{\varphi_{k,s}^{f,1}, \varphi_{k,s}^{f,2}, \varphi_{k,s}^{f,3}\}$. In Appendix B, we derive the exact expressions for the various $\varphi_{k,s}^f$. Based on these stationary probabilities, we can now easily develop the measures for trading volume (i.e. the trading rate) and the degree of internalization (i.e., the internalization rate) for each strategy under a given market design choice and fee structure.

The trading rate TR in a period under strategy s , market design choice k and fee structure f 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:

$$TR_{k,s}^f = \varphi_{k,s}^{f,2} + \varphi_{k,s}^{f,3} \quad (17)$$

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

$$IR_{k,s}^f = \frac{\varphi_{k,s}^{f,2}}{\varphi_{k,s}^{f,2} + \varphi_{k,s}^{f,3}} \quad (18)$$

and can be seen as the percentage of trades that is internalized. The non-internalization rate is then equal to $1 - IR_{k,s}^f$. The proof is then immediate by filling in the stationary probability distribution results of Appendix B in the definition of trading rate and internalization rate. $\square$

Proof of Proposition 5. Welfare for each combination of market design choice and fee structure is obtained by plugging in the results from Proposition 4 into Equation (11) (proof of which strategy produces highest welfare is obtained by simple comparison of contained in the discussion above the proposition. $\square$

Proof of Proposition 6. Comparing observed spreads as well as trading volumes for the different combinations of market design choices and fee structures shows that welfare may be higher even though observed spreads are low, or trading volumes high. In contrast, a comparison of average cum-fee spreads shows that the lowest one always coincides with highest welfare. $\square$

Proof of Corollary 3. Recall that $c_\alpha = \frac{N(V_h - V_l)(N-1)}{(N+\alpha-1)(2N+1)}$. When $\alpha = 0$, $c_\alpha = \frac{N(V_h - V_l)}{(2N+1)} = c_2$. This implies that the combination of transparency and trade-specific fees when $\alpha = 0$ coincides with the social planner's preferred outcome. $\square$

B 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.²⁰ These three states form a finite state space. For each possible strategy s corresponding to transparency regime k and fee structure f , a Markov chain (with the property that the next state depends only on the current state) could be constructed with transition matrix $\widehat{M}_{k,s}^f$, which is a 3×3 matrix capturing all transitions from one state to another (see Colliard and Foucault (2012) for a similar approach). These matrices reflect the transition probabilities corresponding to the equilibrium decisions under the considered strategy, and could be written as follows:

$$\widehat{M}_{k,all}^f = \begin{bmatrix} \frac{1}{2} & \frac{1}{2N} & \frac{N-1}{2N} \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}; \quad (19)$$

and

$$\widehat{M}_{k,own}^f = \begin{bmatrix} \frac{1}{2} + \frac{N-1}{2N} & \frac{1}{2N} & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}; \quad (20)$$

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 $\varphi_{k,s}^f$ is a row vector satisfying $\varphi_{k,s}^f = \varphi_{k,s}^f \cdot \widehat{M}_{k,s}^f$, i.e. $\varphi_{k,s}^f$ is a normalized left eigenvector of $\widehat{M}_{k,s}^f$ associated with the eigenvalue 1. Do note that as this Markov chain is irreducible and aperiodic, the stationary distribution $\varphi_{k,s}^f$ is unique. Let $\varphi_{k,s}^{f,1}$, $\varphi_{k,s}^{f,2}$ and $\varphi_{k,s}^{f,3}$ be the stationary probability of occurrence of states 1, 2 and 3 under the considered strategy. Then the stationary probability distribution could be denoted as $\varphi_{k,s}^f = (\varphi_{k,s}^{f,1}, \varphi_{k,s}^{f,2}, \varphi_{k,s}^{f,3})$.

²⁰We do not need to make a distinction between buyers and sellers because both sides of the market are symmetric in our model.

This distribution $\varphi_{k,s}^f$ could be derived for each of the strategies as:

$$\begin{aligned}\varphi_{k,all}^f &= (\frac{2}{3}, \frac{1}{3N}, \frac{N-1}{3N}) \\ \varphi_{k,own}^f &= (\frac{2N}{2N+1}, \frac{1}{2N+1}, 0)\end{aligned}\tag{21}$$

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

C Appendix C: Asymmetric Fee Allocation between LO and MO

In this Appendix, we derive the results as presented in Table 2. Recall that our general system of indifference equations (1) applies. With asymmetric fee allocation, $E(\text{fee}(\text{MO}))$ and $E(\text{fee}(\text{LO}))$ are influenced by how the total trading cost of a particular match is split between a MO and a LO, i.e., they depend upon α . We now explain the values taken by $E(\text{fee}(\text{MO}))$ and $E(\text{fee}(\text{LO}))$ for the different settings. Filling those values in the system of indifferent equations (1) and solving yields the corresponding bid and ask quotes, and the resulting observed and cum-fee spreads. The values for $E(\text{fee}(\text{MO}))$ and $E(\text{fee}(\text{LO}))$ are summarized in Table C.1:

Table C.1: Expected Fees for MO and LO in case of uneven split

Design \ Fee structure		Uniform		Trade-specific	
		<i>all</i>	<i>own</i>	<i>all</i>	<i>own</i>
Anonymous	MO	$2\alpha \frac{N-1}{N}c$	-	$2\alpha \frac{N-1}{N}c$	-
	LO	$2(1-\alpha) \frac{N-1}{N}c$	-	$2(1-\alpha) \frac{N-1}{N}c$	-
Transparent	MO	$2\alpha \frac{N-1}{N}c$	-	$2\alpha c$	0
	LO	$2(1-\alpha) \frac{N-1}{N}c$	-	$2(1-\alpha)c$	0
Optional Discl.	MO	$2\alpha \frac{N-1}{N}c$	-	$2\alpha \frac{N-1}{N}c$	0
	LO	$2(1-\alpha) \frac{N-1}{N}c$	-	$2(1-\alpha) \frac{N-1}{N}c$	0

Note: the second row indicates the *all* or *own* strategy. The second column within Uniform and Trade-specific indicates whether a LO or MO is under consideration. The symbol ‘-’ indicates this combination does not apply.

Notice that when $\alpha = 0.5$, the values in Table C.1 coincide with our main analysis. Solving the system of indifference equations while using the respective fee values shown in Table C.1, and taking into account that $\gamma = 0.5(1/(2N))$ when the *all* (*own*) strategy applies yields the outcomes as shown in Table 2. The value $c_\alpha = \frac{N(V_h - V_l)(N-1)}{(N+\alpha-1)(2N+1)}$ results

from the analysis of optional disclosure with trade-specific fees. It follows from comparing the profit of a LO under the *all* strategy $(\frac{1}{2} [V_h - (\frac{V_h+2V_l}{3} + (\frac{N+1}{3N})c) - (\frac{N-1}{N})c])$ and the profits of a LO under the *own* strategy $(\frac{1}{2N} [V_h - (\frac{V_h+2V_lN}{2N+1})])$, and showing that profitable deviations are not possible. The proofs of this proceed similarly as the proofs of Propositions 1, 3, and 2, and are therefore omitted for brevity.

References

- AFME (2012): “Post Trade Settlement Committee Task Force on CSD Account Structure,” *AFME Policy Paper*.
- (2015): “Post Trade Explained: The Role of Post-Trade Services in the Financial Sector,” *AFME Policy Paper*.
- ANAND, A., J. HUA, AND T. MCCORMICK (2016): “Make-Take Structure and Market Quality: Evidence from the U.S. Options Markets,” *Management Science*, 62, 3271–3290.
- BARUCH, S. (2005): “Who Benefits from an Open Limit Order Book?” *Journal of Business*, 78, 1267–1306.
- BLOOMFIELD, R. AND M. O’HARA (1999): “Market Transparency: Who Wins and Who Loses?” *Review of Financial Studies*, 12, 5–35.
- BROLLEY, M. AND K. MALINOVA (2017): “Informed Trading in a Low-Latency Limit Order Market,” *Working Paper*.
- CESPA, G. AND T. FOUCAULT (2014): “Sale of price information by exchanges: does it promote price discovery?” *Management Science*, 60, 148–165.
- COLLIARD, J.-E. AND T. FOUCAULT (2012): “Trading Fees and Efficiency in Limit Order Markets,” *Review of Financial Studies*, 25, 3389–3421.
- COLLIARD, J.-E. AND P. HOFFMANN (2017): “Financial Transaction Taxes, Market Composition, and Liquidity,” *Journal of Finance*, 72, 2685–2716.
- DEGRYSE, H. AND N. KARAGIANNIS (2019): “Once Upon a Broker Time? Order Preferencing and Market Quality,” *Working Paper*.
- DEGRYSE, H., M. VAN ACHTER, AND G. WUYTS (2009): “Dynamic Order Submission Strategies with Competition between a Dealer Market and a Crossing Network,” *Journal of Financial Economics*, 91, 319–338.

- DUFFIE, D. AND H. ZHU (2011): “Does a Central Clearing Counterparty Reduce Counterparty Risk?” *Review of Asset Pricing Studies*, 1, 74–95.
- ESMA (2018): “Final Report - Guidelines on Internalised Settlement Reporting under Article 9 of CSDR, ESMA70-151-1258,” .
- EUROPEAN COMMISSION, (2016): “Capital Markets Union and Post-Trade,” *Working Paper*.
- FOTAK, V., V. RAMAN, AND P. K. YADAV (2014): “Fails-to-Deliver, Short Selling, and Market Quality,” *Journal of Financial Economics*, 114, 493–516.
- FOUCAULT, T. (1999): “Order Flow Composition and Trading Costs in a Dynamic Limit Order Market,” *Journal of Financial Markets*, 2, 99–134.
- FOUCAULT, T., O. KADAN, AND E. KANDEL (2005): “Limit Order Book as a Market for Liquidity,” *Review of Financial Studies*, 18, 1171–1217.
- (2013a): “Liquidity Cycles and Make/Take Fees in Electronic Markets,” *Journal of Finance*, 68, 299–341.
- FOUCAULT, T., M. PAGANO, AND A. ROELL (2013b): *Market Liquidity - Theory, Evidence, and Policy*, Oxford University Press.
- FOUCAULT, T., D. SRAER, AND D. J. THESMAR (2011): “Individual Investors and Volatility,” *Journal of Finance*, 66, 1369–1406.
- GLOSTEN, L. R. (1998): “Competition, Design of Exchanges and Welfare,” *Working Paper*.
- GOETTLER, R. L., C. A. PARLOUR, AND U. RAJAN (2005): “Equilibrium in a Dynamic Limit Order Market,” *Journal of Finance*, 60, 2149–2192.
- HOFFMANN, P. (2014): “A Dynamic Limit Order Market with Fast and Slow Traders,” *Journal of Financial Economics*, 113, 156–169.

- HOLLIFIELD, B., R. A. MILLER, P. SANDÅS, AND J. SLIVE (2006): “Estimating the Gains from Trade in Limit-Order Markets,” *Journal of Finance*, 61, 2753–2804.
- HOLTHAUSEN, C. AND J. TAPKING (2007): “Raising Rival’s Costs in the Securities Settlement Industry,” *Journal of Financial Intermediation*, 16, 91–116.
- JURANEK, S. AND U. WALZ (2020): “Organizational Design, Competition, and Financial Exchanges,” *Scandinavian Journal of Economics*, 122, 132–163.
- KADAPAKKAM, P.-R. (2000): “Reduction of Constraints on Arbitrage Trading and Market Efficiency: An Examination of Ex-Day Returns in Hong Kong after Introduction of Electronic Settlement,” *Journal of Finance*, 55, 2841–2861.
- KHAPKO, M. AND M. ZOICAN (2020): “How Fast Should Trades Settle?” *Management Science*, *forthcoming*.
- KOEPL, T. V., C. MONNET, AND T. TEMZELIDES (2012): “Optimal Clearing Arrangements for Financial Trades,” *Journal of Financial Economics*, 103, 189–203.
- LOON, Y. C. AND Z. ZHONG (2012): “The Impact of Central Clearing on Counterparty Risk, Liquidity, and Trading: Evidence from the Credit Default Swap Market,” *Working Paper*.
- MADHAVAN, A. (1995): “Consolidation, Fragmentation, and the Disclosure of Trading Information,” *Review of Financial Studies*, 8, 579–603.
- MALINOVA, K. AND A. PARK (2015): “Subsidizing Liquidity: The Impact of Make/Take Fees on Market Quality,” *Journal of Finance*, 70, 509–536.
- (2016): “Market Design for Trading with Blockchain Technology,” *Working Paper*.
- MERRICK, J. J., N. Y. NAIK, AND P. K. YADAV (2005): “Strategic Trading Behavior and Price Distortion in a Manipulated Market: Anatomy of a Squeeze,” *Journal of Financial Economics*, 77, 171–218.

- NAIK, N. Y., A. NEUBERGER, AND S. VISWANATHAN (1999): “Trade Disclosure Regulation in Markets with Negotiated Trades,” *Review of Financial Studies*, 12, 873–900.
- NYSE EURONEXT (2008): “Form 10-K Annual Report for the Securities and Exchange Commission,” [http :
://www.sec.gov/Archives/edgar/data/1368007/000119312508064701/d10k.htm](http://www.sec.gov/Archives/edgar/data/1368007/000119312508064701/d10k.htm).
- OXERA (2014): “Global Cost Benchmarking of Cash Equity Clearing and Settlement Services,” *ASX Report*.
- PAGANO, M. AND A. RÖELL (1996): “Transparency and Liquidity: A Comparison of Auction and Dealer Markets with Informed Trading,” *Journal of Finance*, 51, 579–611.
- RINDI, B. (2008): “Informed Traders as Liquidity Providers: Anonymity, Liquidity and Price Formation,” *Review of Finance*, 12, 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, 4601–4641.
- SEC (2006): “Securities and Exchange Commission Release No. 34-53742; File No. SR-NSCC-2006-04,” .
- SKJELTORP, J. A., E. SOJLI, AND W. W. THAM (2018): “Identifying Cross-Sided Liquidity Externalities,” *Management Science*, 62, 3085–3391.
- TAPKING, J. (2007): “Pricing of Settlement Link Services and Mergers of Central Securities Depositories,” *ECB Working Paper No. 710*.
- TAPKING, J. AND J. YANG (2006): “Horizontal and Vertical Integration in Securities Trading and Settlement,” *Journal of Money, Credit, and Banking*, 38, 1765–1795.

VAN ACHTER, M. (2010): “A Dynamic Limit Order Market with Diversity in Trading Horizons,” *Working Paper*.

WEINSTEIN, S. AND S. YEKINI (2015): “Transparency in Securities Transactions and Custody Chains,” *Working Paper*.