Liquidity: What you see is what you get?

Vincent van Kervel*

April 2012

Abstract

Competition between electronic limit order books improves the overall liquidity of equity markets in most studies. However, my model shows that liquidity offered on the limit order books combined may strongly overestimate the actual liquidity available to investors. The excess is caused by high-frequency traders operating as market makers, who may duplicate their limit order schedules on several venues to increase their execution probabilities. Then, after a trade on one venue they will quickly cancel outstanding limit orders on others. The magnitude of the cancellations depends on the fraction of investors that may access several venues simultaneously, i.e., who use Smart Order Routing Technology (SORT). The reason is that market makers incur higher adverse selection costs when the investor trades at a competing venue first. Consequently, a higher fraction of SORT investors reduces the incentives of market makers to place duplicate limit orders. The empirical results confirm the main prediction of the model, as trades on the most active venues are followed by cancellations of limit orders on competing venues of more than 53% of the trade size.

JEL Codes: G10; G14; G15;

Keywords: Market microstructure, Fragmentation, Market makers, High Frequency Trading, Smart Order Routing

---

*Department of Finance, Tilburg University, TILEC, CentER. Email: v.l.vankervel@uvt.nl., Phone: +31 13 466 4056. I am indebted to my thesis supervisors Hans Degryse and Frank de Jong. I thank Larry Glosten for invaluable comments and the invitation to stay at Columbia University, whose hospitality is gratefully acknowledged. In addition, I thank Jean-Edouard Colliard, Peter Hoffmann, Albert Menkveld and seminar participants at the University of Amsterdam and the VU University of Amsterdam for suggestions. Comments and suggestions are appreciated. All errors or mine.

1 Introduction

Two important trends have drastically changed equity markets in recent years. First, technological innovations have lead to high-frequency trading, a trading strategy whereby a computer algorithm analyzes market data and trades at extremely high speed. Second, competition between trading venues has caused a dispersion of trading volumes and liquidity across venues, i.e., the market has become fragmented. These two changes in the structure of equity markets might strongly affect the optimal behavior of investors. Indeed, as high-frequency traders operate on several trading venues simultaneously, the order flow and liquidity of these venues become strongly interrelated.[1] This paper argues that the interrelation of liquidity across trading venues causes substantial overestimation of liquidity aggregated over these trading venues.

I underpin this hypothesis in a model of competition between two centralized limit order books. The model predicts that high-frequency traders, who supply liquidity by acting as traditional market makers, have an incentive to duplicate their limit orders on both venues. They will do so because this strategy increases their execution probabilities and therefore expected profits. Then, after execution on one venue, they will quickly cancel their limit orders on competing venues. Effectively, the depth aggregated over all venues overstates true liquidity, as a single trade reduces liquidity on many venues simultaneously. The empirical results strongly support this hypothesis: a trade reduces liquidity at competing venues by 53% of the trade size within a second.

My first contribution to the literature is a theoretical model. The model is based on the framework of a pure limit order market with adverse selection of Sandås (2001), extended to a two-venue setting. High-frequency market makers supply liquidity with limit orders, whereas potentially informed traders demand liquidity with market orders. In this setting, only a fraction of the traders has the technological infrastructure to submit market orders to both venues simultaneously, i.e., smart order routing technology (SORT). Effectively, I introduce market segmentation because non-SORT traders cannot access the liquidity of both venues. In equilibrium, a limit order faces higher adverse selection costs when executed against a SORT trader. The reason is that conditional upon execution, there is a probability that she traded on the competing venue already and therefore the combined

[1] The high level of interaction between markets became apparent during the flash crash, i.e., between the E-mini S&P 500 futures and the individual stocks (SEC-CFTC "Findings regarding the market events of May 6 2010", 2010).

trades are larger and more informed on average. Consequently, a lower fraction of SORT traders reduces adverse selection costs and increases equilibrium liquidity supply. However, the additional liquidity follows from market segmentation only, and will be cancelled after a trade on the competing venue; hence the term duplicate liquidity.

The model offers the following predictions. First, due to the information content of a trade on one venue, limit orders on the same side of the order books of competing venues will be cancelled. Cancellations would not occur in a single venue setting, which implies that I must test whether the observed cancellations are significantly greater than zero. Second, after a trade new limit orders will be placed on the opposite side of the limit order books of all venues, consistent with asymmetric information models. Third, duplicate limit orders will only be placed on venues where non-SORT investors operate (mostly the traditional market), because market makers face lower adverse selection costs when trading with non-SORT investors. When all investors use SORT, competition between market makers only improves liquidity, as documented by Glosten (1998) and Foucault and Menkveld (2008). Fourth, the model explains why entrant venues are typically very liquid but execute relatively little volume. Although market makers offer substantial liquidity there, only a small fraction of investors use SORT to tap this liquidity.[2] Fifth, the decision of one investor to adopt smart order routing technology increases her accessible liquidity, but also imposes a negative externality on all other investors (the SORT and non-SORT traders as well as market makers).

My second contribution to the literature is an empirical investigation. The dataset contains the entire limit order books of all relevant trading venues with publicly displayed data, for a sample of FTSE 100 stocks in November 2009. These stocks are traded in a fairly fragmented environment, as the traditional market (the London Stock Exchange) executes 66% of lit trading volume, leaving 34% to four competing venues (Chi-X, Bats Europe, Turquoise and Nasdaq OMX). I test the models main predictions by investigating the short-term correlations between the supply and demand of liquidity across trading venues. Specifically, I regress changes in liquidity supply at the bid or ask side of one venue on contemporeneous and lagged liquidity demand of all five venues, i.e., buy and sell trading volumes. Lagged trading volumes of up to ten seconds away measure investors' responses to trades over time. I sample one observation per 100 milliseconds in order to analyze high-frequency trading strategies. To the best of my knowledge, this paper is the

[2] In the model, transaction costs and trading speeds are assumed equal across trading venues, and therefore do not explain differences in market shares.

first to analyze the impact of high-frequency trading strategies on the liquidity supply across trading venues.

The empirical results strongly support the models' main prediction that trades are followed by limit order cancellations on competing venues. That is, within 100 milliseconds, transactions on the three most active trading venues are followed by cancellations on the *same side* of competing limit order books of 38 to 85% of the transaction size. As a result, liquidity aggregated over all venues overstates liquidity available to investors, since e.g. a 100 share trade reduces liquidity by more than 138 shares. Further, these cancellations increase to 53 to 149% of the transaction size after one second (depending on the trading venue). This finding is particularly relevant to algorithms designed to split up large trades over time, as the liquidity impact of each individual small trade is indeed larger than previously thought. The fact that liquidity shocks immediately spill over to other venues is not captured by static liquidity measures, such as the quoted depth.

The analysis also confirms several other predictions of the model. First, the magnitude of cancellations is stronger on the primary market after trades on entrant venues (46 to 52%), than vice versa (14 to 30%). Second, trades reveal information on the true asset value, as a trade on the ask side is followed by new limit orders on the bid side of 30 to 70% of the tradesize.

The main policy implication of the model is that fair markets require investors to be able to split up trades *simultaneously* across several venues. When a trader leaves a millisecond delay between the split, the market effectively becomes segmented. That is, after high-frequency traders observe the first part of the trade, they will quickly cancel their duplicate limit orders on competing venue before the second part arrives. This high-frequency trading strategy is known as latency arbitrage, and relates to the trend of increasingly faster trading mechanisms.[3]

Most related to this work is literature on competition between electronic limit order books. Pagano (1989) predicts that all trading activity should divert to the trading system with the lowest transaction costs, and only unstable equilibria may exist when two venues have identical cost structures. In contrast, Glosten (1998) shows that two electronic limit order markets can coexist when tick sizes are discrete and time priority rules apply. Since time priority is absent *across* venues, competition between liquidity suppliers increases, which in turn raises aggregate liquidity. This point is further developed in Foucault and

[3]See for example "Trading at the speed of light", Sept 12, 2011 on www.ft.com.

Menkveld (2008), who coin this channel the “queue-jumping” effect. Competition between exchanges also arises through differences in the tick size, where the venue with the smallest tick size becomes most liquid (Biais, Bisière, and Spatt (2010) and Buti, Rindi, Wen, and Werner (2011)). My model confirms all these findings, and adds the presence of duplicate limit orders when some investors cannot access all trading venues.

Empirical research on competition between exchanges typically focusses on its impact on aggregate liquidity and welfare. O’Hara and Ye (2011) find that competition between exchanges reduces the effective cost of trading. Jovanovic and Menkveld (2011) find that competition between an entrant (Chi-X) and incumbant market (Euronext) has an ambiguous effect on total welfare. Degryse, de Jong, and van Kervel (2011) show that aggregate liquidity increases by competition between venues with publicly displayed limit order books, but worsens by competition of opaque markets. Instead, the current paper studies the effect of competing exchanges on the liquidity supply of each of the individual exchanges.

Rather than assuming a fraction of non-SORT traders, market segmentation may also arise when investors have different trading speeds. Biais, Foucault, and Moinas (2011) show that high-frequency trading facilitates the search for trading opportunities, but increases adverse selection costs for slow traders. As a result, the equilibrium level of investment in high-frequency technology exceeds the welfare maximizing level. Hoffmann (2011) shows that high-frequency traders have lower adverse selection costs on average, which in turn causes slow traders to post less aggressive limit orders. The result is an ambiguous effect on welfare, which depends on the stocks fundamental volatility. High-frequency traders might extract rents from liquidity motivated traders when they operate as intermediaries (Cartea and Penalva, 2011). McInish and Upson (2011) provide empirical evidence that high-frequency traders pick off slower traders in the US, due to regulation that effectively causes slow investors to trade against stale quotes. Hasbrouck and Saar (2011) argue that trading speeds affect the competition between liquidity suppliers in a single trading venue. This paper shows that different trading speeds might in fact cause duplicate limit orders in fragmented markets.

Finally, this paper relates to recent research on high-frequency traders who act as market makers (e.g., Jovanovic and Menkveld (2011), Menkveld (2011) and Guilbaud and Pham (2011)). Such market makers gain the bid-ask spread by offering liquidity at both sides of the limit order book, while simultaneously managing advere selection costs (e.g., Glosten and Milgrom (1985), Glosten and Harris (1988)) and inventory risk (e.g., Ho and

Stoll (1981)). This paper extends these works by analyzing market making when trading is fragmented across electronic limit order books.

The remainder of the paper is structured as follows. Section 2 describes the duplicate limit order hypothesis and the model. Section 3 presents the empirical work, after which I conclude.

2 The model

In this section I first describe the duplicate limit order hypothesis. Then, I place it into a model that quantifies this duplicate limit order effect. In essence, the model is a combination of Glosten (1998), Sandås (2001) and Foucault and Menkveld (2008). I contribute to Foucault and Menkveld (2008) by allowing for adverse selection. Rather than analyzing the single exchange setting in Sandås (2001), I focus on two competing centralized limit order books. Compared to Glosten (1998), I introduce market segmentation by constraining some traders to have access to one trading venue only, and show how this causes overestimation of consolidated liquidity. Glosten (1998) and Sandås (2001) are special cases of my model.

2.1 Duplicate limit order hypothesis

In a fragmented trading environment time priority is not enforced between trading venues, whereas price priority is enforced only when the trader has access to both venues. Price priority implies that limit orders with a better price are executed before those with a worse price, while time priority entails that limit orders placed first are executed first.[4]

Because of the absence of time priority across venues, liquidity suppliers can improve their execution probabilities by placing similar limit orders on several venues simultaneously. After execution on one venue, they will cancel remaining limit orders on the other venues.[5] Therefore, I predict that a trade is immediately followed by cancellations of limit orders on the same side of competing limit order books. A tradeoff arises however, as there is a probability that both limit orders will be executed simultaneously, causing the liquidity supplier to trade too much.

[4] In the US, price priority across markets is enforced by law, Reg NMS. Time priority however, crucial for this hypothesis, is not enforced.

[5] Note that this strategy does not work in a single exchange setting due to time priority.

Who might pursue this duplicate limit order strategy? I argue that high-frequency traders who operate as market makers can strongly benefit from the increased execution probability. That is, placing duplicate limit orders increases their trading rate and expected profits. At the same time, using state of the art technology allows them to monitor several venues simultaneously and cancel limit orders quickly after observing trades and other news. Being able to cancel quickly reduces the expected cost of simultaneous execution and adverse selection.[6] In contrast, this strategy is likely not very attractive to "regular traders." For some traders, the technology required for continuous monitoring might be too expensive. Other traders might use algorithms to optimally split up large quantities over time, in which case they will not cancel limit orders since each child order is part of a large parent order.

2.2 Model setup

This section quantifies the cost of executing duplicate limit orders in an adverse selection framework. I show that duplicate limit orders face higher adverse selection costs because conditional upon executing both limit orders, the incoming trade is larger and therefore more informed on average.

Consider two venues, A and B, and two types of investors, market makers and traders. The risk neutral market makers supply liquidity by placing limit orders on one or both venues. They are profit maximizing and use high-frequency trading technology to quickly access all venues. The risk averse traders demand liquidity by placing market orders. Traders have private information or liquidity motives, and therefore want to trade quickly.[7]

The asset has a fundamental value X_t, which incorporates all information available up until period t. The next periods fundamental value is given by

$$X_{t+1} = X_t + \mu + \varepsilon_{t+1}, \tag{1}$$

where μ is a trend and ε_{t+1} a random innovation. Traders have some private information on ε_{t+1}.

[6] In this context, trades on two venues occur simultaneously when the liquidity supplier is not fast enough to adjust his outstanding limit orders after the first trade. Effectively, his quotes are stale when the second trade comes in.

[7] Since traders are risk averse and may have liquidity motives, the market does not break down like in a Kyle (1985) framework.

Trading occurs sequentially over periods indexed by t. Each period t consists of three stages. First, the market makers arrive consecutively and place limit orders on one or both venues. They do so until no market maker finds it optimal to place additional limit orders. Then the trader arrives, placing market orders with total size x that consume liquidity on one or both venues. Finally, the market makers update their expected fundamental value of the asset conditional on the incoming market order using some price impact function. Now, the game starts over and is repeated for every trade. Because of high-frequency trading technology, the last two stages might last only a few milliseconds.

Market makers place their limit orders on a discrete pricing grid, $\{p_1, p_2, ...p_k\}$ for the ask side. The current expected value (midpoint) is p_0, the best ask price p_1 and the tick size $\Delta = p_i - p_{i-1} > 0$. I focus on the ask side only, prices larger than p_0, as the bid side is analogous. Denote the number of shares offered on venue $j \in \{A, B\}$ at each price level by $\{Q_{j1}, Q_{j2}, \ldots, Q_{jk}\}$. Price and time priority exists within each venue, but not between venues.

2.3 The trader

The trader is randomly drawn from a population of traders, which consists of three types. The first type only goes to venue A (fraction α), the second type only to venue B (fraction β), and the third type uses smart order routing technology (SORT) to access both venues simultaneously (fraction $\gamma = 1 - \alpha - \beta$). Simultaneous is defined here as sending trades to both venues so fast that the market makers are unable to update their limit order schedules inbetween the trades. When both venues offer the same best price, the SORT investors are indifferent as to where to send their trades to. In this case, they simply use a tie-breaking rule, which posits that with probability π they first buy shares on venue A, and with probability $(1 - \pi)$ they first buy shares on venue B. The parameters α, β and π are constant, as I focus on a high-frequency environment.

Four reasons motivate why some investors are not able to trade on both venues simultaneously. First, human traders with access to both venues might trade too slowly, creating a delay of several milliseconds when they split up a trade across two venues. In this case, high-frequency market makers have sufficient time to update their limit order schedules inbetween trades. As a result, human traders effectively have access to one exchange only.[8] This argument seems realistic, as McInish and Upson (2011) show that slow traders are adversely selected because they often trade against stale quotes. Second, smart order routing technology might be too expensive for some traders, as it requires fixed costs for technological infrastructure, software, programmers and access to data feeds etc. Third, fixed costs of sending your trade to two venues could make it more economical to trade through one price but save on the transaction costs (such that fixed clearing and settlement costs are paid only once). Fourth, investors may deliberately decide to split up large trades over time, for example to benefit from the resiliency of liquidity. In this case, they might not need the liquidity offered on additional venues.

The trader is a buyer or seller with equal probability, and has a reservation price $p_m > p_1$ at which she is not willing to buy.[9] The expected order size is similar for all types, with mean ϕ and exponential density function

$$f(x) = \frac{1}{\phi} \exp(-\frac{x}{\phi}) \qquad if\ x > 0\ (market\ buy). \tag{2}$$

The cumulative distribution function is $F(\cdot)$. Assuming an exogenous order size simplifies the analysis, although in reality the size depends on several factors, such as the current state of the limit order book, the trader's expected fundamental value and her current holdings. However, I mainly focus on the behavior of the market makers, described next.

2.4 The market makers

Since traders are informed about the innovations in the assets true value, market makers update their expectation of the fundamental value based on the size of the incoming trade x. The market makers observe trades on both exchanges. They use a price impact function $h(x)$ which is non-decreasing, since buy trades typically contain positive information with respect to the true value (similarly, sells contain negative information). Then, larger orders

[8] In fact, when market makers can update inbetween the two trades, the slow trader will never prefer to split up his trade across two venues. The reason is that the market makers update their limit order schedules symmetrically across two venues. Therefore, if it is optimal to submit the first part of the trade to venue A, then it is also optimal to submit the second part there.

[9] This small assumption prevents the trade from walking up the limit order book too much in case of a thin order book, but does not affect the outcome of the model.

cause more adverse selection costs and greater price impacts.

$$E(X_{t+1}|x) = E(X_t) + \mu + h(x). \tag{3}$$

As in Sandås (2001), the price impact function is linear with coefficient λ,

$$h(x) = \lambda x. \tag{4}$$

For simplicity, I assume market makers face a fixed limit order *execution cost* c, while the *submission cost* is zero.[10] Given this setup, we can calculate the expected profit of a limit order placed on any location q in the queue of limit orders on each venue. The profit depends on the expected value of the asset conditional upon execution of the limit order. For a limit order on venue A, this value is $E(X|x > q)$ when the trader immediately goes to A (denoted $E_q(X)$ for brevity), and $E(X|x > q + Q_{B1})$ when the trader first buys all shares Q_{B1} on venue B and then goes to A. Denote the probability that the incoming order x is larger than q as $\overline{F}_q = 1 - F(q)$, then the profit of a limit order on price level 1 of venue A is

$$\Pi_{A,q} = (\alpha + \gamma\pi)\overline{F}_q(p_1 - c - E_q(X)) + \gamma(1 - \pi)\overline{F}_{q+Q_{B1}}(p_1 - c - E_{q+Q_{B1}}(X)). \tag{5}$$

In the first term, the limit order executes against traders going to venue A only (α) and against SORT investors who choose to trade on venue A first ($\gamma\pi$). Then, the expected profit is simply the price minus the fixed cost c and the expected value of the asset conditional on $x > q$. The second term represents SORT traders who first buy all the shares offered at price p_1 on venue B, and then buys shares on venue A ($\gamma(1 - \pi)$). Indeed, market makers only realize profits when their limit orders are executed. Therefore, conditional upon execution, the second term of the expected profit is lower since the incoming trade is larger and more informed, i.e., $E(X_{t+1}|x > q + Q_{B1}) > E(X_{t+1}|x > q)$.

Not surprisingly, we observe that the expected profit of limit orders on venue A depends on the number of shares offered on venue B. Therefore, to obtain the equilibrium liquidity supply we need to solve for the profit equations of limit orders on both venues

[10] Placing limit orders is costless on the venues we analyze empirically.

simultaneously. The profitabilty of limit orders on price level 1 of venue B is

$$\Pi_{B,q} = (\beta + \gamma(1-\pi))\overline{F}_q(p_1 - c - E_q(X)) + \gamma\pi\overline{F}_{Q_{A1}+q}(p_1 - c - E_{Q_{A1}+q}(X)). \quad (6)$$

2.5 Equilibrium

The model is in equilibrium when no market maker can profitably place an additional limit order on any price level (as in Glosten (1994), Proposition 2). Therefore, the expected profit of the *marginal limit order*, the single share offered at the end of the queue of limit orders, must equal zero for all price levels on each venue. Following Sandås (2001), I substitute $q = Q_{A1}$ (the marginal limit order) and integrate the profit equation over the distribution of the incoming order x, using equations (2), (3) and (4)

$$\Pi_{A1} = \int_{Q_{A1}}^{\infty} (\alpha + \gamma\pi)(p_1 - c - X_t - \lambda x)\frac{1}{\phi}\exp(-\frac{x}{\phi})\mathrm{d}x + \int_{Q_{A1}+Q_{B1}}^{\infty} \gamma(1-\pi)(p_1 - c - X_t - \lambda x)\frac{1}{\phi}\exp(-\frac{x}{\phi})\mathrm{d}x = 0.$$

The first integral goes to infinity, because the marginal limit order is executed for any trade larger or equal to Q_{A1}. For demonstrational purposes previous equations contain γ, which I next substitute with $(1 - \alpha - \beta)$ to calculate the solutions. Solving the integral gives

$$\Pi_{A1} = (\alpha + \pi(1-\alpha-\beta))(p_1 - c - X_t - \lambda(\phi + Q_{A1}))\exp(-\frac{Q_{A1}}{\phi}) + (1-\pi)(1-\alpha-\beta)(p_1 - c - X_t - \lambda(\phi + Q_{A1} + Q_{B1}))\exp(-\frac{Q_{A1}+Q_{B1}}{\phi}) = 0.$$

The zero expected profit condition implies that the first line of the equation is positive while the second line is negative. In equilibrium, the market makers expect to lose to traders that go to venue B first, and profit from traders that go to venue A first. Similarly, for venue B we have

$$\Pi_{B1} = \int_{Q_{B1}}^{\infty} (\beta + \gamma(1-\pi))(p_1 - c - X_t - \lambda x)\frac{1}{\phi}\exp(-\frac{x}{\phi})\mathrm{d}x + \int_{Q_{A1}+Q_{B1}}^{\infty} \gamma\pi(p_1 - c - X_t - \lambda x)\frac{1}{\phi}\exp(-\frac{x}{\phi})\mathrm{d}x = 0,$$

which I solve to obtain

$$\Pi_{B1} = \pi(\beta + (1-\pi)(1-\alpha-\beta))(p_1 - c - X_t - \lambda(\phi + Q_{B1}))\exp(-\frac{Q_{B1}}{\phi}) + \pi(1-\alpha-\beta)(p_1 - c - X_t - \lambda(\phi + Q_{A1} + Q_{B1}))\exp(-\frac{Q_{A1}+Q_{B1}}{\phi}) = 0.$$

The two equations with two unknowns can be solved implicitly[11]

$$\begin{aligned} Q_{A1} &= \frac{p_1 - c - X_t - \lambda\phi}{\lambda} - \frac{\gamma(1-\pi)Q_{B1}}{\gamma(1-\pi) + (\alpha + \gamma\pi)\exp(\frac{Q_{B1}}{\phi})}, \\ Q_{B1} &= \frac{p_1 - c - X_t - \lambda\phi}{\lambda} - \frac{\gamma\pi Q_{A1}}{\gamma\pi + (1-(\alpha+\gamma\pi))\exp(\frac{Q_{A1}}{\phi})}. \end{aligned} \tag{7}$$

The zero expected profit condition holds for prices deeper in the order book too. Now, the expected profit consists of three terms, as a limit order on price level 2 on venue A might get executed by traders of type α (who first buy Q_{A1} and then Q_{A2}), by type $\gamma\pi$ (who first buy Q_{A1}, Q_{B1} and then Q_{A2}), or by type $\gamma(1-\pi)$ (who first buy Q_{B1}, Q_{A1}, Q_{B2} and then Q_{A2}). For brevity, denote $Z(x) = \alpha(p_2 - c - X_t - \lambda x)\frac{1}{\phi}\exp(-\frac{x}{\phi})$, then

$$\Pi_{A2} = \int_{Q_{A1}+Q_{B1}}^{\infty} Z(x)\mathrm{d}x + \int_{Q_{A1}+Q_{B1}+Q_{A2}}^{\infty} Z(x)\mathrm{d}x + \int_{Q_{B1}+Q_{A1}+Q_{B2}+Q_{A2}}^{\infty} Z(x)\mathrm{d}x. \tag{8}$$

I solve the system for prices deeper in the order book in similar fashion to equation (7), which gives an implicit solution of the form $Q_{A2} = f(Q_{A1}, Q_{B1}, Q_{B2}; parameters)$, and similarly for other price levels.

[11] This solution is unique. In Equation (7), $\partial Q_{A1}/\partial Q_{B1} < 0$ and $\partial Q_{B1}/\partial Q_{A1} < 0$, since the first term does not depend on Q_{A1} or Q_{B1}, while we subtract a second term which consists of non-negative parameters only.

2.6 Testable implications

In this section I discuss two static predictions of the model, which hold at each point in time, and two dynamic predictions, which hold before and after a trade.

The static predictions follow from the solutions for Q_{A1} and Q_{B1} in equation (7). The first terms are identical, and equal the optimal quantity offered in a single venue setting (the solution obtained by Sandås (2001)). However, we subtract a non-negative second term, implying that the offered quantities on the individual venues are weakly lower in a fragmented market. The second term reflects the adverse selection costs incurred by the market makers when a SORT trader buys at the competing venue first. Based on the solution, I derive the following static predictions (the proofs are in section A.1 of the Appendix).

Proposition 1 *When the liquidity at both venues has reached equilibrium:*

1. *An increasing fraction of SORT traders γ strictly reduces consolidated depth.*
2. *For $\gamma > 0$ and SORT traders do not always go to venue A first or always to venue B first, i.e., $0 < \pi < 1$, the consolidated liquidity is strictly higher than the liquidity offered in a single exchange setting.*

A larger fraction of SORT traders implies less segmented markets, as many traders have access to the liquidity of both venues simultaneously. Effectively, market makers face higher adverse selection costs when trading with a SORT trader, because there is a probability that she traded on the competing venue already. Therefore, a higher fraction of SORT traders increases expected adverse selection costs and reduces equilbrium liquidity supply. The second part of the proposition relates to the benefit of competition between exchanges as documented by Glosten (1998) and Foucault and Menkveld (2008), which stems from market makers' ability to jump the queue of limit orders on one venue by placing limit orders on the other. Such "queue jumping" erodes the profitability of existing limit orders on the first venue and effectively increases competition between liquidity suppliers.

To demonstrate how the equilibrium liquidity depends on the individual parameters, I take the first derivative from the solution of Q_{A1} with respect to α, β, γ and π. The results can be summarized as follows, and hold symmetrically for Q_{B1} (see the Appendix).

Proposition 2 *Other things equal, the number of shares offered at price level 1 of venue A, Q_{A1}:*

1. *Increases by π, the probability that SORT traders go to venue A first; by α, the fraction of investors that only have access to venue A; and by β, those with access only to venue B when holding α constant.*

2. *Decreases by Q_{B1}; by γ, the fraction of SORT investors; and by β when holding γ constant.*

As expected, the liquidity offered on venue A strictly increases in α, and decreases in γ. These effects are solely driven by adverse selection costs, which reduce when more investor go to venue A first (α and π) and increase when more investors go to venue B first and then to venue A ($\gamma(1-\pi)$).[12] Since $\gamma = 1 - \alpha - \beta$, liquidity on A increases with β when holding α constant (such that the fraction of smart order routers decreases), and decreases with β when holding γ constant (such that the fraction of α decreases). An increase in Q_{B1} implies larger adverse selection costs when a SORT trader purchases Q_{B1} first and then Q_{A1}.

Notice the following extreme case. When the fraction of SORT investors $\gamma = 0$, the second term of equation (7) becomes zero and the quantity offered on each venue is identical to that of a single exchange setting. However, while consolidated liquidity is twice that of the single exchange setting, investors are not better off since they can trade on one venue only. Effectively, duplicates of limit orders are placed on both venues, such that consolidated liquidity overstates liquidity available to investors by a factor of two.

The market segmentation that arises when $\gamma < 1$ leads to important dynamic changes in market makers' liquidity supply, i.e., before and after a trade. This is the main prediction of the model.

Proposition 3 *Define the "consolidated liquidity impact" of a trade as the difference in equilibrium consolidated liquidity before and after a one unit trade.*

1. *When $\gamma = 1$ or in the single exchange setting, the consolidated liquidity impact equals one.*

[12] The signs of the derivatives with respect to π and Q_{B1} hold on the relevant domain of Q_{A1}, i.e., for Q_{A1} smaller or equal to the equilibrium solution.

2. *The consolidated liquidity impact decreases by γ. For $\gamma < 1$, the consolidated liquidity impact is strictly larger than one, such that the impact of a trade on liquidity is larger than the trade size.*

In the first case, a one unit trade reduces consolidated liquidity with one unit because the price impact of the trade equals the slope of the limit order book. However, a decrease in γ reduces the market makers expected adverse selection costs, who in turn increase their liquidity supply. But given that the information content of a trade is held fixed, it must be that the increased liquidity is cancelled after the trade.[13] Effectively, the private information of a trade is incorporated on the competing venue via cancellations of limit orders. The numerical example of the next subsection explains this in more detail.

Market segmentation does not arise when $\gamma = 1$ or in the single exchange setting, so market makers have no incentive to place duplicate limit orders. The channel of duplicate limit orders occurs in addition to the effect that for $\gamma > 0$, consolidated liquidity increases because of competition between liquidity suppliers (the second part of Proposition (1)). In what follows, I define duplicate liquidity as the liquidity impact minus one, i.e., the impact of a trade on consolidated liquidity due to additional cancellations of limit orders.

Next we analyze the impact of a trade on the liquidity supply at the other venue.

Proposition 4 *Define the "cross-venue liquidity impact" as the difference in equilibrium liquidity at the competing exchange before and after a one unit trade. The cross-venue liquidity impact as always negative, and increases in magnitude by the liquidity of the competing exchange. Further, a similar effect takes place on the bid side after a trade on the ask side, and vice versa.*

The cancellations are simply a consequence of the private information revealed by the trade. The price impact changes the market makers' expected fundamental value, and accordingly they cancel and resubmit their entire limit order schedules around this value. If market makers offer more liquidity on a venue, the cancellations following trades on the competing venue will be larger too. Proposition (2) describes the impact of the models' parameters on Q_{A1}, which directly translate into the cross-venue liquidity impact. Also, since a trade on the ask side increases the expected fundamental value, the liquidity supply

[13] The proposition that γ reduces consolidated liquidity is proved in the Appendix, which makes this proof redundant.

on the bid side of both venues will increase accordingly. This holds symmetrically for trades on the bid side.

To summarize this section, a *reduction* in the fraction of SORT investors lowers liquidity via less competition between liquidity suppliers, but this negative effect is more than offset by the increase in duplicate limit orders. However, the duplicate limit orders will be cancelled after a trade, as shown in the numerical example of the next section.

2.7 Numerical example

In this section I substitute the models parameters with realistic values and analyze the equilibrium. In particular, we are interested in the impact of the fraction of traders that might go to one venue only (α and β) on outstanding liquidity of both venues.

I choose the following parameter values. The average trade size is 1 unit ($\phi = 1$), the best ask price is £10.00 and the tick size is 0.5 cent, which is the relevant case for the sample stocks with a price of £10.00. The fixed order execution cost c is 0.1 cent (one fifth of the tick size). The price impact of a one unit trade $\lambda = 20$ basis points, and the tie-breaking rule $\pi = 0.5$. The models' results come out clearest when I set the fundamental value X_t just above £9.99, such that the depth in the order book is constant at each price level for the case that all investors use SORT ($\gamma = 1$).[14]

The numerical outcomes for the four best price levels are shown in Table (1), where α and β vary. The first row shows the single exchange setting, $\alpha = 1$, which is the Sandås (2001) solution. This is the benchmark case, and shows that 1.53 units are offered on the best price level, and 2.5 units on all subsequent price levels. In this case, the quantities offered beyond the best price are constant because of the tradeoff between improved prices and higher adverse selection costs, which equals the tick size divided by the price impact.

The second column shows the situation where all investors use smart order routing technology ($\gamma = 1, \alpha = \beta = 0$). Compared to the benchmark case, consolidated liquidity is 63% higher on price p_1 (2.5 versus 1.53), and identical on all subsequent levels. This corresponds to part 2 of Proposition (1), and shows that liquidity summed over the first price level and beyond indeed dominates the benchmark case. Consistent with competition

[14] Specifically, I set the fundamental value $X_t = 9.993943$. Small changes in X_t relative to the fixed pricing grid cause changes in offered liquidity at the best price, which interact with liquidity on the competing exchange and in turn with liquidity deeper in the order book.

between market makers, the SORT traders reduce the expected profits to market makers with 33% (from 2.24 to 1.50 basis points per trade, in the bottom part of the Table).

From columns three to six the fraction of SORT traders gradually decreases, which increases consolidated liquidity (part 1 of Proposition (1)). When we move towards the full duplicate limit orders case ($\gamma = 0$, column (6)), consolidated liquidity is twice that of the single exchange setting. In each case, the Table reports the market shares and the per trade expected profits to markets makers of both venues.

The main prediction of the model is that a higher fraction of SORT traders reduces the amount of duplicate limit orders (Proposition (3)). To illustrate this point, we analyze the impact of a 2.5 share trade on the liquidity of both limit order books. The price impact is $2.5\lambda = 1$ tick, meaning that all prices shift up exactly one level after market makers have revised their limit orders. When all investors use SORT (column (2)), the trade consumes the 1.25 units of Q_{A1} and 1.25 of Q_{B1}, and then the limit order books are immediately in equilibrium (market makers will not need to revise limit orders on the ask side). In effect, there are no duplicate limit orders and the consolidated liquidity impact of a trade is exactly one. In contrast, when no investors use SORT (column (6)), the trade will consume half of the offered liquidity on both venues (1.25 of Q_{A1} and Q_{B1}), but then the remaining shares will be cancelled because of the price impact of these trades. In effect, after the market makers revision, the 2.5 share trades reduce liquidity on both venues with 5 shares, meaning that 100% of the order size is cancelled.

The second prediction of the model is that the cross-venue liquidity impact of a trade depends on the liquidity of the competing exchange (Proposition (4)). We confirm this in column (7), where 50% of the investors have access to venue A only and 50% use SORT. Thus, venue A is very liquid compared to B. As above, when we trade 1.25 on venue A and B simultaneously, the entire liquidity schedules shift up one price level. Therefore, cancellations on Q_{A1} are 0.17 (from 1.42-1.25 to 0), and Q_{A2} are 1.12 (from 2.37 to 1.42), whereas Q_{B1} are 0 and Q_{B2} is -0.10, a replenishment (from 1.15 to 1.25). In general, the replenishment implies that after a trade on one exchange, limit orders may get cancelled on the competing venue and replaced on the current venue to restore equilibrium.

Indeed, column (7) shows the realistic setting in Europe, where a large fraction of investors is able to trade only on the traditional venue A (50%), and the remaining investors use SORT to access a new entrant venue B. The model correctly describes the following stylized facts. Despite that only SORT investors trade on venue B, liquidity is still fairly high here: 88% relative to venue A at p_1, which reduces to 29% when we sum liquidity on levels p_1 to p_4. This is consistent with the stocks analyzed empirically. In addition, the liquidity at p_1 of venue A is about 93% of that in the single venue setting in column (1), which closely matches the finding of Degryse et al. (2011) that competition between exchanges reduces liquidity of the traditional market with about 10%. Lastly, while the entrant offers a substantial fraction of total liquidity, it has a relatively low market share of trading volume (22%, highly consistent with the empirical results).

3 Empirical results

This section first presents a brief overview of the sample stocks trading environment, followed by a data description and an explanation of the liquidity measures, the DepthAsk(X) and DepthBid(X). Then I test the duplicate limit order hypothesis and discuss the results.

3.1 Market structure FTSE100 stocks

The FTSE100 stocks are primarily listed on the London Stock Exchange (LSE), which is the fourth largest stock exchange in the world. In the sample period, November 2009, the LSE executes approximately 61% of trading volume (excluding dark pool and OTC volumes).[15] These stocks are traded on the trading system SETS, an electronic limit order market organised by the LSE which integrates market makers liquidity provision. Note that the market makers in the model are regular investors, who use high-frequency technology to operate like traditional market makers. Continuous trading occurs between 08:00 and 16:30, local time.

Once stocks are listed on the LSE, alternative trading venues may decide to organize trading in them as well.[16] Four important entrants have emerged which also employ publicly displayed limit order books: Chi-X, Bats, Turquoise and Nasdaq OMX Europe. These venues are regulated as Multi-lateral Trading Facilities (MTFs), the European equivalent to ECNs. While these entrants in effect have the same market model as the LSE, they differ with respect to trading technology (speed in particular), fixed and variable trading fees, and some of the types of orders that may be placed (e.g., pegging a limit order price

[15] As reported by Fidessa, see http://fragmentation.fidessa.com.

[16] This feature makes the current study inherently different from literature on cross-listings, where firms may choose to list on several exchanges to improve access to global capital.

to the midpoint, such that it always equals the midpoint $+n$ ticks). Investors can demand ("take") liquidity by issuing a market order or supply ("make") liquidity by issuing limit orders at any moment in time. All markets allow for visible, partially hidden (iceberg) and fully hidden limit orders. The hidden portion of iceberg orders becomes visible after (partial) execution of the visible part. Accordingly, limit orders have a price, transparency, time priority within a trading venue, but not between trading venues.

Chi-X started trading in April 2007 and is the most succesful entrant in terms of market share with 24% of trading volume in November 2009. Turquoise and Nasdaq OMX started trading FTSE 100 firms as of September 2008, and Bats two months later. Their market shares are substantially lower, with 5.5%, 1.8% and 7.6%, respectively. In May 2010 Nasdaq OMX closed down, as they did not meet their targeted market shares.[17] As of July 2009, the five trading venues use identical tick sizes, which depend on the stock price. All the new competitors employ a maker - taker pricing schedule, where executed limit orders receive a rebate of 0.18 to 0.20 basis points, while market orders are charged 0.28 to 0.30 basis points of traded value. These make-take fees are relatively small compared to a ticksize of 5 basis points (0.5 pennies) for a £10.00 stock.

The trading venues with publicly displayed limit order books execute approximately 60% of total volume, while the remaining 40% is executed on dark pools, Broker-Dealer Crossing Networks, internalized and Over-The-Counter.

3.2 Data

The current analysis is based on a subsample of ten FTSE100 stocks, each randomly selected from one market cap decile of the 100 constituents (i.e., a size stratified sample).[18] The sample period consists of 10 trading days (November 2 - 13, 2009), and high-frequency data are taken from the Thomson Reuters Tick History database. For each stock, the data contain separate limit order books for the five trading venues.

For each transaction, I observe the price, traded quantity and execution time to the millisecond,[19] while for each limit order placement, modification or cancellation, the dataset

[17] See "Nasdaq OMX to close pan-European equity MTF", www.thetradenews.com.

[18] I choose ten stocks during ten trading days as computational limitations prevents me from using the full sample of stocks or more trading days.

[19] If a single market order is executed against several outstanding limit orders, separate messages are generated for each limit order.

reports the timestamp and the ten best prevailing bid and offer prices and their associated quantities.[20]

While the time stamp is per millisecond, I take snapshots of the limit order books at the end of every 100^{th} millisecond, resulting in approximately 30 million observations. Higher frequencies are not useful when comparing multiple trading venues, as it may lead to inaccuracies because of latency issues, i.e. millisecond reporting delays. Per snapshot, I observe the available liquidity and that periods trading volumes on the buy and sell sides of every venue. The advantage of taking snapshots is that the data become evenly spaced, such that every observation receives the same weight in the regressions. Accordingly, lagged variables in the regressions become easily interpretable.

I do not directly observe hidden and iceberg limit orders. However, from trades I can construct the executed hidden quantity, based on the state of the order book directly before and after the trade and the traded quantity. As such, I do observe hidden liquidity that gets 'hit' by a market order. These data are identical to those offered by several information vendors, meaning I use the information set available to the market.

Panel A of Table 2 presents summary statistics for the sample stocks. As I select stocks from within each size decile, there is a large variation in market cap: the mean is £21 billion with a £37 billion standard deviation. Accordingly, also trading volume (in shares and pounds) and realized volatility vary substantially. A large part of this variation stems from Itv PLC, the smallest stock in the sample. In contrast, the market shares of the five trading venues are fairly stable between firms, and highly representative for the entire FTSE100 index.

Panel B of Table 2 presents summary statistics on the average number of limit orders and transactions per minute. While the LSE's market share is largest by far, the number of transactions lie much closer together (i.e., Chi-X trades are smaller on average). On top of that, the number of limit orders on Chi-X greatly exceeds the LSEs, on average 218 versus 160 per minute.

Worth mentioning is the ratio of limit orders to trades, which is 31:1 for the LSE, 51:1 for Chi-X and increases as a venues market share goes down to 123:1 for Nasdaq. This is mostly due to high-frequency traders placing many limit orders, and shows that new entrants are highly active despite a small overall market share. Chi-X and Bats are the

[20] This feature of the data makes it difficult to follow a limit order over time, because its location in the order book can changes and may even fall outside the observable range of ten price levels.

most succesfull new competing venues in terms of market share, number of transactions and limit order book activity.

Finally, in Table 3 reports summary statistics on trading volume per minute, denominated in GBPs. For each venue I show the total buy and sell volume (left panel), and the volume executed against hidden limit orders (right panel). Overall, the buy and sell sides are fairly symmetric.

3.3 The DepthAsk(X) and DepthBid(X) liquidity measures

This subsection explains the DepthAsk(X) and DepthBid(X) measures, also used in Degryse et al. (2011).

The DepthAsk(X) aggregates all shares offered at prices between the midpoint and the midpoint plus X basis points. Similarly, the DepthBid(X) sums the shares offered within the midpoint minus X basis points and the midpoint. The midpoint is the average of the best bid and ask price available in the market, and I choose $X = 10$ basis points relative to the midpoint. The price constraint X garantuees that I sum liquidity at prices close to the midpoint, i.e. only at good price levels. This is important, as liquidity offered deeper in the order book is less likely to be executed, and therefore less relevant to investors. The number of shares in the interval are then converted to the value in GBPs.

Formally, define price level $j = 1, 2, ..., J$ on the pricing grid and the midpoint M, the average of the best ask and bid price available in the market, then for venue v,

$$DepthAsk(X)_v = \sum_{j=1}^{J} P_{j,v}^{Ask} Q_{j,v}^{Ask} \mathbf{1}\left(P_{j,v}^{Ask} < M(1+X)\right), \tag{9a}$$

$$DepthBid(X)_v = \sum_{j=1}^{J} P_{j,v}^{Bid} Q_{j,v}^{Bid} \mathbf{1}\left(P_{j,v}^{Bid} > M(1-X)\right). \tag{9b}$$

The measures are calculated at the end of every 100 millisecond interval and represent liquidity offered at the bid and ask side, per trading venue. When taking higher values for X, liquidity deeper in the order book is also incorporated. Then, comparing different price levels X reveals the shape of the order book. For example, if the depth measure increases rapidly in X, the order book is deep while if it increases only slowly, the order book is relatively thin. The order book is asymetric when the absolute difference between

DepthAsk(X) and DepthBid(X) is high.

The measure has several features that make it highly suitable for the empirical approach. First, the measure is calculated per venue, for the bid and ask side, which allows us to analyze correlations between the provision and consumption of liquidity across venues, sides and over time. Second, the measure can directly be related to trading volumes, as both are denoted in pounds. Third, the Depth measure incorporates limit orders beyond the best price levels, making it robust to small, price improving limit orders. Such orders are often placed by high-frequency traders, who mostly drive the dynamics in the model. Fourth, by choosing a fixed interval the measure is independent of the tick size, which varies across stocks. For a detailed discussion and comparison of this measure with related liquidity measures such as the Cost of Roundtrip ($CRT(D)$) and Exchange Liquidity Measure ($XLM(V)$), I refer the interested reader to Degryse et al. (2011).

Table 4 contains summary statistics on the Depth(10) and Depth(50) measures for the bid and ask side, reported in GBPs and calculated per exchange. The statistics are based on single observations per tenth of a second per stock, equal weighted over all stocks. Depending on the tick sizes, the Depth(10) aggregates liquidity of two to five price levels on the bid and ask side. First and most strikingly, we observe that the liquidity offered on Chi-X is 86% of the LSE, while they execute only a third of the LSE volume. As predicted by the model, this implies that indeed a substantial fraction of investors only have access to the LSE. The liquidity available at Bats is roughly 40% of the liquidity at the LSE, while Turquoise and Bats have approximately 20% each. The ask side contains on average 3% more liquidity than the bid side, meaning that the order books are very symmetrical.

The regression analysis works with *changes* in DepthAsk(X) and DepthBid(X), i.e., the value of the current minus the previous observation. As Eq. (9) shows, these changes depend on the activity in the limit order book and on the level of the midpoint. The model relates changes in the depth measures due to limit order book activity only, i.e., the placement, cancellation, modification and execution of limit orders. Therefore, I define $Chg\_DepthAsk(X)$ as the difference in DepthAsk between each period, holding the midpoint constant

$$Chg\_DepthAsk(X)_{i,t} = DepthAsk(X, M_{t-1})_{i,t} - DepthAsk(X, M_{t-1})_{i,t-1}. \tag{10}$$

The measure simply shows how much liquidity in GBPs is added or removed from one period to the next.

3.4 Methodology

The model predicts that high-frequency traders operating as market makers place duplicate limit orders on several exchanges. Therefore, we expect that a trade is followed by cancellations of limit orders on the same side of the limit order books of competing venues. This prediction holds equally for buy and sell trades. The second prediction of the model is that the cancellations should occur to a greater extent on venues with a high share of non-SORT traders, i.e., the traditional exchange. The reason is that the market makers place more duplicate limit orders on the LSE than on Chi-X, which in turn will be cancelled after a Chi-X trade. The third prediction, consistent with theories of asymmetric information, is that new limit orders will be placed on the bid side after a buy trade (and on the ask side after a sell), including some duplicate limit orders.

Cancellations of limit orders occur when the $Chg\_DepthAsk(10)$ is negative (after controlling for trading volume). Thus, I regress the $Chg\_DepthAsk(10)$ of venue v on contemporeneous and lagged buy and sell volumes of all venues. I add lags of ten seconds, which are 100 periods as the data are sampled at a 100 millisecond frequency, because the model's predictions apply to a high-frequency trading environment and should be incorporated very quickly. Instead of estimating 100 individual lagged coefficients, I add five variables that average trading volume of 1, 2-4, 5-10, 11-20 and 21-100 periods away, per venue for buy and sell volumes. Section A.2 in the appendix explains in more detail how I obtain the cumulative impact of a transaction over time.

A trade is classified as Buy or $Sell$, and define trading venue $v = 1, ..., 5$, for the current and five lagged groups l, stock i and time t. I test the model's predictions with the following regressions:

$$\begin{aligned} Chg\_DepthAsk(10)_{it}^{V} &= c_i + \sum_{v=1}^{5}\sum_{l=0}^{5}\left(\beta_{l,v}^{Buy} \times Buy_{it-l}^{v} + \beta_{l,v}^{Sell} \times Sell_{it-l}^{v}\right) + \\ & \sum_{v=1}^{5}\left(\beta_{v}^{Buy} \times BuyHid_{it}^{v} + \beta_{v}^{Sell} \times SellHid_{it}^{v}\right) + \varepsilon_{it}. \qquad (11) \end{aligned}$$

The term after the firm fixed effects represents the buy and sell volumes (in GBPs) for the five venues covering the six lagged groups. The term on the second line controls for contemporeneous hidden buy and sell liquidity (observed when executed), which is added for the following reason. The effect of a buy trade on $Chg\_DepthAsk(10)$ of that venue should mechanically be -1 : a one pound trade reduces the depth with exactly one pound. However, since a trade executed against a hidden limit order does not reduce DepthAsk(10), I control for executed hidden liquidity.

This regression is executed ten times: for the bid and ask sides of five trading venues. It shows how many pounds close to the midpoint are submitted or cancelled after a one pound buy or sell trade on some venue. Note that the effects do not die out over time, as for example a buy trade might contain positive price information, such that some limit orders will permanently be cancelled on the ask side.

3.5 Results

The regression results are reported in Table (5), with the change in DepthAsk and DepthBid of all venues as dependent variables. Each column represents one regression, showing separate coefficients for buy and sell volumes, per trading venue. The dependent and independent variables are all measured in GBPs. Within each venue, the displayed coefficients represent the cumulative effect over time (the running sum). I only show the effect within one tenth of a second, after 1 second and after 10 seconds. Intermediate lagged values are estimated to improve the model fit, but for brevity not reported. Next I discuss the findings for the DepthAsk only, as the results for the DepthBid are symmetric.

In line with the Proposition (4), the first column shows that a one pound buy trade at Chi-X is immediately followed by cancellations on the LSE of -0.21 pounds (-21%). After ten seconds, the effect is -0.61, meaning that more than half of the Chi-X trade size is cancelled on the LSE. The coefficients for Bats are similar, -0.27 immediately and -0.54 after ten seconds (all significant at the 1% level). This effect is economically very large, and implies that trades on entrant venues are immediately followed by cancellations on the traditional market. Note that the effect cannot be explained by investors who simultaneously place trades on several venues, since the regression controls for trades on other venues. Also, the regression controls for the execution of hidden limit orders. The effect of Nasdaq and Turquoise trades on LSE liquidity are negative, but surprisingly small.[21]

The immediate effect of a one pound LSE buy trade on LSE DepthAsk is -0.83 pounds. This implies that while the trade removed 1 pound, either 17 cents is immediately replen-

[21] I need to investigate the possibility that Nasdaq and Turquoise mainly attract trades when only they offer the best price in the market. In this case, we would not expect cancellations.

ished, or, first a new limit order is placed which immediately provokes the trade. The latter explanation is consistent with the findings of Hasbrouck and Saar (2009). A Chi-X buy trade reduces Chi-X DepthAsk with -1.31 (column (2)), implying that beyond the reduction of 1 pound, an additional and significant 31 cents is cancelled. Coefficients of the other venues lie between -0.70 and -1.26.

Consistent with the second part of Proposition (4), the LSE indeed responds more strongly to Chi-X and Bats trades than vice versa. In column (2) and (3), LSE trades reduce liquidity on Chi-X and Bats with -0.18 and -0.05 after ten seconds (compared to -0.61 and -0.54 above).

The cancellation effect also holds particularly strong between Bats and Chi-X, in both directions. Bats and Chi-X have immediate cross-coefficients of -0.58 and -0.18 (column (2-3)), and are the most succesfull entrants in terms of market share. These findings suggest that a fraction of high-frequency market makers operate on the LSE, Chi-X and Bats simultaneously, generating the strong correlations of liquidity between these venues. In contrast, Turquoise and Nasdaq seem more independent, as their liquidity does not respond much to trades on the LSE, Chi-X and Bats and vice versa.[22]

The immediate effect of any venues sell trades on LSE DepthAsk(10) is economically large and positive, with coeffients ranging from 0.10 to 0.30 (column (1), bottom panel). This result is consistent with an information effect: the sell trade conveys negative information about the stock, such that market makers improve prices (and quantities) of their ask limit orders. In addition, the finding is consistent with selling investors who use algorithms that optimally place market orders on the bid sides and new limit orders on the ask sides (as predicted by theory of e.g., Parlour (1998)).

Concluding, the results confirm that consolidated liquidity is overstated in a fragmented market, because a single transaction is followed by substantial cancellations of limit orders on the other venues. In the model, these cancellations would not occur in a single venue setting or when all traders use smart order routing technology.

[22] A possible explanation is that at times, Nasdaq and Turquoise offer zero DepthAsk(10). Obviously, in such periods their liquidity does not respond to competitors trades, pushing the estimated coefficients toward zero.

3.6 Drivers of the cancellation effect

In this section I analyze the driving forces behind the cancellation effect. In particular, I am interested in the variation in cancellation rates of the venues across stocks and over time, and how they are influenced by observable market characteristics.

I estimate the previous model for hour t of stock i, but use as dependent variable the change in ask (or bid) liquidity summed over all venues, i.e., the change in consolidated Depth(10). These regressions represent the impact of a trade at venue v on market wide depth, consistent with Proposition (3). In a new dataset I store the estimated coefficients $Coef_{v,it}$ accumulated over one second after the trade, for buy trades on the ask side liquidity and sell trades on the bid side liquidity. The sample size is extended from 10 to all 21 trading days in November 2009 for the same 10 stocks used in the previous analysis, which results in 1890 observations. Next, I filter the data by dropping observations where the trading venue v of $Coef_{v,it}$ has executed less than £10.000 or 50 transactions. Without this restriction, the estimated coefficients on the cancellation rates have too high standard errors and are not economically meaningful. As a consequence, we disregard the regressions of Turquoise and Nasdaq OMX altogether because these venues are insufficiently active.[23] In addition, I winsorize $Coef_{v,it}$ at the 1% and 99% level to reduce the impact of outliers.

Based on the new dataset, I run the following two regressions for firm i, venue $V = \{LSE, Chi\text{-}X, Bats\}$, and hour h

$$Coef_{V,ih} = c_i + \delta_h + Frag_{ih} + LS\_LSE_{ih} + \varepsilon_{ih}, \tag{12}$$

$$\begin{aligned} Coef_{V,ih} = {} & c_i + \delta_{(h)} + Frag_{ih} + LS\_LSE_{ih} + Ln(entrant\ trades)_{ih} + Order\ Imb_{ih} + \\ & Volat_{ih} + Ln(turnover)_{ih} + Ln(Depth(10)\ cons)_{ih} + \varepsilon_{ih}. \end{aligned} \tag{13}$$

Using estimated coefficients as dependent variable in a second step regression does not bias the coefficients of the second step.[24] $Frag$ is the degree of fragmentation of stock-hour ih, defined as (1-HHI). HHI is the sum of squared market shares of the five venues based on trading volume (also used in Degryse et al. (2011)). LS_LSE is the liquidity share of the LSE, defined as the ratio of LSE Depth(10) over consolidated Depth(10). $Ln(entranttrades)$ is the logarithm of the number of transactions at entrant venues. Order

[23] We do incorporate these venues when calculating consolidated liquidity, although the results are qualitatively unaffected when leaving them out.

[24] The measurement error from the first step only increases the standard errors of the coefficients in the second step.

imbalance ($Order\ Imb$) is the difference between the logarithm of buy volume and sell volume. $Volat$ is the realized volatility based on 5-minute underlying stock returns, calculated as the sum of 12 squared 5-minute returns, hour by hour. $Ln(turnover)$ and $Ln(Depth(10)\ cons)$ are the logarithms of trading volume (in GBP) and consolidated Depth(10). We add firm fixed effects and daily dummy variables $\delta_{(h)}$ to absorb fixed stock and day characteristics.

Summary statistics of the regression variables are presented in Table (6). The top panel shows that the degree of fragmentation is fairly constant across stocks and over time, with a mean of 0.59 and standard deviation of 0.09. This also holds for the LSE liquidity share, with a standard deviation of 0.11. In contrast, the number of trades on the entrants vary substantially, with a mean of 435 and a standard deviation of 485. The order imbalance is highly volatile too, since a one standard deviation increase implies that buy volume is 43% larger than sell volume. The bottom panel shows the cancellation rates of trades at the individual venues on consolidated liquidity. The impact on consolidated liquidity is approximately equal to the sum of the impact on the individual venues, as reported in Table (5). The standard deviations of Turquoise and Nasdaq are very large, which justifies excluding them from further analysis.

The regression results for the three trading venues are presented in Table (7). $Frag$ should be a good proxy for the number of traders with SORT technology, which corresponds to Proposition (3). In the model, market makers face higher adverse selection costs when trading with SORT traders, because there is a risk that the SORT trader already bought shares on the competing venue, which in turn implies a larger and more informed trade on average. Thus, we expect that $Frag$ correlates negatively with the amount of cancellations.

The coefficient on $Frag$ is indeed negative for Chi-X and Bats (columns (2-3)), but not for the LSE (column (1)). While this result seems contrasting, a likely explanation is that $Frag$ also proxies for the activity of market makers who operate on several venues. Then, a higher level of fragmentation implies more cross-market activity and *more* cancellations. This explanation mainly holds for trades on the LSE, which is always active, whereas Chi-X and Bats are only active when the market is fragmented.[25] When I control for other variables that proxy for the activity of cross-market market makers (columns (4-6)), $Frag$ becomes strongly negative (and significant). Then, a one standard deviation increase in

[25] An alternative explanation is that some liquidity providers on the LSE do not operate on the entrant venues, i.e., do not cancel their limit orders after they trade on the LSE.

$Frag$ (0.09) reduces cancellations of LSE trades by 5% and Chi-X and Bats trades by 30% and 43%.

The liquidity share of the LSE is negatively correlated with the amount of duplicate limit orders, and significant in all regressions. A large LSE liquidity share means that market makers are less active on the entrants which results in fewer cross-market activity and cancellations. These coefficients are economically large, as a one standard deviation change reduces the cancellation rates with 20-40% (depending on the trading venue).

The remaining control variables in columns (4-6) have the expected signs. An increasing number of entrant trades increases the cancellation rates, which is consistent with market makers operating more actively on entrant venues. Similarly, more overall liquidity (consolidated depth) is consistent with a higher amount of duplicate limit orders. Turnover has a negative sign, which means that in periods of high trading activity the market makers place less duplicate limit orders.

The bottom panel of Table (7) shows the results using the cancellation effects of sell trades on bid liquidity as dependent variable, and are highly symmetric.

4 Conclusion

Equity markets have evolved rapidly in recent years due to the increasing number of trading venues and heavy investments in high-frequency trading technology. I show that a specific type of high-frequency traders, those who operate like modern day market makers, might in fact cause a strong overestimation of liquidity aggregated across trading venues. The reason is that these market makers place duplicate limit orders on several venues, and after execution of one limit order they quickly cancel their outstanding limit orders on competing venues. As a result, a single trade on one venue is followed by reductions in liquidity on all other venues.

The empirical analysis confirms that trades are followed by substantial cancellations on competitors. That is, within 100 milliseconds after trades on some venues $39-85\%$ of the order size is cancelled on competitors. After one second this number increases to $98-125\%$, which shows that the impact of a trade on liquidity is in fact twice the trade size. Note that the reduction in liquidity is due to cancellations of limit orders, since the analysis controls for transactions on all individual trading venues.

The analysis is executed on a sample of FTSE100 stocks, which are fairly fragmented in the sample period. The main advantage of using these stocks is the rich data: I observe the entire limit order books of the five trading venues that compete for these stocks. The data represent all publicly displayed liquidity, which is the same information set available to the general public. By sampling the data at one observation per tenth of a second, I study high-frequency trading behavior. The paper contributes to the literature by analyzing the impact of high-frequency trading strategies on the demand and supply of liquidity across exchanges.

The results relate to the benefits and drawbacks of equity market fragmentation. While previous research typically shows a beneficial effect of fragmentation on liquidity, the analysis argues that the benefits, while still positive, are mitigated because of duplicate limit orders. Note that duplicate limit orders will not arise when all investors use smart order routing technology (SORT).

An additional result of the model is that liquidity shocks are highly correlated across trading venues. Therefore, static measures of consolidated liquidity (and quoted depth) overstate the liquidity available to investors, since a single trade reduces liquidity on all venues. However, liquidity measures such as the quoted spread and effective spread are unaffected by the duplicate limit orders. My results are particularly important to algorithms designed to minimize execution costs by splitting up trades over time. Indeed, the strong cancellations mitigate the benefits of order splitting—a result that cannot be observed by static liquidity measures.

My model focusses on two exchanges only, but can already predict a substantial fraction of duplicate limit orders. Therefore, the relevance of the model is only strenghtened by the fact that most European stocks trade on more than four exchanges (with publicly displayed limit order books) and some US stocks on so much as twelve exchanges. A larger number of trading venues encourages market makers to duplicate their limit orders.

The predictions of the model are very relevant to US markets too. Indeed, it seems that all US traders use SORT because the regulator prohibits the execution of trades at prices inferior to the best available price. However, duplicate limit orders will still arise when some traders are unable to split up trades *simultaneously* across venues. When a trader leaves a millisecond delay between the split, high-frequency traders can observe the first part of the trade and quickly cancel duplicate limit orders on other venues before the second part of the trade arrives. Therefore, the results of the model also hold when the trading speeds varies across investors.

The main policy implication of the model is that fair markets require traders to have simultaneous access to all trading venues. In any other case non-SORT or slow traders pay a higher price for liquidity, which is transfered to the SORT traders. This result is consistent with McInish and Upson (2011), who find that high-frequency traders are able to exploit slower traders in other ways too (see also Biais et al. (2011) and Hoffmann (2011)). A second implication is that the decision of a trader to acquire SORT increases the liquidity she has access to, but also imposes a negative externality on other traders. The overall welfare effects depend on the cost of acquiring smart routing technology, and are therefore ambiguous.

A Appendix

A.1 Theory

This section summarizes the proofs of the propositions in the model.

- The consolidated liquidity in a fragmented market is strictly larger than liquidity in a single exchange setting (Proposition (1)).

Proof. Denote the liquidity in a single exchange setting as Q_1, which is the solution from Sandås (2001). I repeat equation (7), and rewrite as

$$Q_{A1} = \frac{p_1 - c - X_t - \lambda\phi}{\lambda} - \frac{\gamma(1-\pi)Q_{B1}}{\gamma(1-\pi) + (\alpha + \gamma\pi)\exp(\frac{Q_{B1}}{\phi})} \equiv Q_1 - c_{B1}Q_{B1} \qquad \text{A1(a)}$$

$$Q_{B1} = \frac{p_1 - c - X_t - \lambda\phi}{\lambda} - \frac{\gamma\pi Q_{A1}}{\gamma\pi + (1-(\alpha+\gamma\pi))\exp(\frac{Q_{A1}}{\phi})} \equiv Q_1 - c_{A1}Q_{A1}. \qquad \text{A1(b)}$$

I need to show that $Q_{A1} + Q_{B1} > Q_1$, which is equivalent to $c_{B1}Q_{B1} + c_{A1}Q_{A1} < Q_1$.

$$\begin{aligned} c_{B1}Q_{B1} + c_{A1}Q_{A1} &= c_{B1}Q_{B1} + c_{A1}Q_1 - c_{A1}c_{B1}Q_{B1} \\ &= c_{A1}Q_1 + (1-c_{A1})c_{B1}Q_{B1} \\ &< c_{A1}Q_1 + (1-c_{A1})Q_1 = Q_1. \end{aligned}$$

In the first equality I simply replace Q_{A1} with the first line of equation A1(a); which I then rewrite in the second line. The inequality holds because $c_{B1}Q_{B1} < Q_1$, since $c_{B1} < 1$ and $Q_{B1} \leq Q_1$. ■

- Both venues co-exist when $\alpha < 1$ and $\beta < 1$. Co-existence means that both venues attract a positive amount of liquidity and market share.

Proof. It is sufficient to show that liquidity on both venues is strictly positive, because SORT traders will always use liquidity of both venues when the incoming order size is sufficiently large (i.e., when the size goes to infinity). This is proved in the last equality above, that $Q_{A1} = Q_1 - c_{B1}Q_{B1} > 0$, and symmetrically for venue B. ■

- The first derivates of the solution of Q_{A1} with respect to $\pi, \gamma, \alpha, \beta$ and Q_{B1} are straightforward and not reported, but available upon request (Proposition (2)).

- The market makers' profit of limit orders on price level 1 is given by

$$\int_0^{Q_{A1}} (\alpha + \gamma\pi)\int_q^{\infty} \Pi_i(x)\mathrm{d}x + \gamma(1-\pi)\int_{q+Q_{B1}}^{\infty} \Pi_i(x)\mathrm{d}x\mathrm{d}q. \tag{14}$$

Calculating the profit of market makers involves integrating over the profitability of all marginal shares at each price level. I calculate the profit of each marginal share by integrating over the incoming trade x, which are traders who either go immediately to venue A, $(\alpha + \gamma\pi)$, or traders that go to venue B first and then to A, $\gamma(1-\pi)$. For brevity, denote $\Pi_i(x) = (p_i - X_t - c - \lambda x)f(x)$, then the market makers profit of all limit orders on price level 1 of venue A is

A.2 Empirical

This section shows that the methodology in section 3.4 measures cumulative effects over time. That is, in regression (11) I add contemporeneous terms and lagged values of trading volumes 100 periods ago (i.e., ten seconds). Instead of estimating 100 coefficients, I create six variables representing the averaged lagged volumes of the current, 1, 2-4, 5-10, 11-20 and 21-100 periods away, per venue for buy and sell volumes. Define t as the current period,

i as the start and j as the end of the intervals (e.g., $i = 2$ and $j = 4$). Then

$$Vol_{i-j} = \frac{1}{j-i+1}\sum_{n=i}^{j} Vol_{t-n}.$$

The periods of lagged values are chosen such that they maximize the model fit. An example of how the data looks like is shown in the table below, where a £1.00 trade occurs on some venue at time $t = 1$. The first four columns show the values of the contemporeneous and three lagged groups. The fifth column shows the cumulative effect of the regression coefficients over time, calculated as a running sum of the individually estimated coefficients. By constructing the variables as averages, the long-term effect of a trade is simply the sum of the estimated coefficients. The standard errors are also calculated based on this sum.

Data example of a trade at time $t = 1$.
Each β_t represents the estimated coefficient of lagged average volumes t periods away, as described in regression (11).

Time	Vol_0	Vol_1	Vol_{2-4}	Vol_{5-10}	Cumulative effect
0	0	0	0	0	0
1	1.00	0	0	0	β_0
2	0	1.00	0	0	$\beta_0 + \beta_1$
3	0	0	0.33	0	$\beta_0 + \beta_1 + 0.33\beta_{2-4}$
4	0	0	0.33	0	$\beta_0 + \beta_1 + 0.66\beta_{2-4}$
5	0	0	0.33	0	$\beta_0 + \beta_1 + \beta_{2-4}$
6	0	0	0	0.20	$\beta_0 + \beta_1 + \beta_{2-4} + 0.2\beta_{5-10}$
7	0	0	0	0.20	$\beta_0 + \beta_1 + \beta_{2-4} + 0.4\beta_{5-10}$

References

B. Biais, C. Bisière, and C. Spatt. Imperfect competition in financial markets: An empirical study of island and nasdaq. *Management Science*, 56(12):2237–2250, 2010.

B. Biais, T. Foucault, and S. Moinas. Equilibrium high frequency trading. *Working Paper Toulouse School of Economics*, 2011.

S. Buti, B. Rindi, Y. Wen, and I. Werner. Tick size regulation, intermarket competition and sub-penny trading. *Working Paper Fisher College of Business, Ohio State University*, 2011.

Á. Cartea and J. Penalva. Where is the value in high frequency trading? *Working Paper Universidad Carlos III de Madrid*, 2011.

H. Degryse, F. de Jong, and V. van Kervel. The impact of dark trading and visible fragmentation on market quality. *Working paper Tilburg University*, 2011.

T. Foucault and A. Menkveld. Competition for order flow and smart order routing systems. *Journal of Finance*, 63(1):119–158, feb 2008. URL http://ideas.repec.org/p/ebg/heccah/0831.html.

L. Glosten. Is the electronic open limit order book inevitable? *Journal of Finance*, 49(4): 1127–1161, Sep. 1994.

L. Glosten. Competition, design of exchanges and welfare. *Working paper Columbia University*, 1998.

L. Glosten and L. Harris. Estimating the components of the bid/ask spread. *Journal of Financial Economics*, 21(1):123–142, may 1988.

L. Glosten and P. Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, (14):71–100, 1985.

F. Guilbaud and H. Pham. Optimal high frequency trading with limit and market orders. *Working paper Université Paris*, 2011.

J. Hasbrouck and G. Saar. Technology and liquidity provision: The blurring of traditional definitions. *Journal of Financial Markets*, 12:143–172, 2009.

J. Hasbrouck and G. Saar. Low-latency trading. *Working paper Stern school of business*, 2011.

T. Ho and H. R. Stoll. Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, 9(1):47–73, March 1981. URL http://ideas.repec.org/a/eee/jfinec/v9y1981i1p47-73.html.

P. Hoffmann. A dynamic limit order market with fast and slow traders. *Working Paper European Central Bank*, 2011.

B. Jovanovic and A. Menkveld. Middlemen in limit order markets. *Working paper New York University*, 2011.

A. Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335, 1985.

T. McInish and J. Upson. Strategic liquidity supply in a market with fast and slow traders. *Working Paper University of Texas, El Paso*, 2011.

A. Menkveld. High frequency trading and the new-market makers. *Working paper Free University Amsterdam*, 2011.

M. O'Hara and M. Ye. Is market fragmentation harming market quality? *Journal of Financial Economics*, 100(3):459–474, June 2011.

M. Pagano. Trading volume and asset liquidity. *Quarterly Journal of Economics*, 104: 25–74, 1989.

C. Parlour. Price dynamics in limit order markets. *Review of Financial Studies*, 11(4): 789–816, 1998. doi: doi:10.1093/rfs/11.4.789.

P. Sandås. Adverse selection and competitive market making: Empirical evidence from a limit order market. *The Review of Financial Studies*, 14(3):705–734, 2001.

Table (1) Numerical Example

We solve the model for the quantities offered on the four best ask price levels of venue A (upper panel) and B (second panel). We vary the parameter values of the fraction of investors with access to venue A only (α) and venue B only (β). The fraction of investors with smart order routing technology (γ) varies accordingly, defined as $1-\alpha-\beta$. The lower panels show the market shares and the per trade total expected profits to market makers of each venue. The expected profits are expressed in basis points relative to the midpoint. The remaining model parameters are held fixed. The average trade size is 1, with a per unit price impact of 20 basis points. The best ask price is £10.00, the tick size is 0.5 pennies and the fixed trading cost is one tenth of a cent. I specifically set the fundamental value to £9.993943, such that offered liquidity is constant at each price level when all investors use SORT (column 2). When both venues offer the best price, SORT traders are equally likely to go to venue A or B first ($\pi = 0.5$).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
α	1	0	0.1	0.1	0.4	0.5	0.5
β	0	0	0	0.1	0.4	0.5	0
γ	0	1	0.9	0.8	0.2	0	0.5
Venue A							
Q_{A1}	1.53	1.25	1.29	1.30	1.46	1.53	1.42
Q_{A2}	2.50	1.25	1.78	1.87	2.47	2.50	2.37
Q_{A3}	2.50	1.25	2.49	2.86	2.58	2.50	2.55
Q_{A4}	2.50	1.25	2.64	2.94	2.52	2.50	2.50
Venue B							
Q_{B1}	0	1.25	1.25	1.30	1.46	1.53	1.25
Q_{B2}	0	1.25	1.23	1.87	2.47	2.50	1.15
Q_{B3}	0	1.25	0.78	2.86	2.58	2.50	0.15
Q_{B4}	0	1.25	0.03	2.94	2.52	2.50	0.00
Market shares							
A	1.00	0.50	0.58	0.50	0.50	0.50	0.78
B	0.00	0.50	0.42	0.50	0.50	0.50	0.22
Market maker profits in basis points							
A	2.24	0.75	0.93	0.84	1.06	1.12	1.55
B	0.00	0.75	0.66	0.84	1.06	1.12	0.35

Table (2) Summary statistics of sample firms.
Summary statistics are presented for the 10 FTSE100 sample stocks, calculated using data of November 2009. Panel A shows market cap (in millions), price, average daily traded volume (millions of shares), turnover (millions of pounds), realized volatility and the market shares of the five trading venues. The trading venues are the London Stock Exchange (LSE), Chi-X, Bats Trading, Turquoise and Nasdaq OMX Europe. Panel B shows limit order book data, where statistics are equally weighted based on one observation per tenth of a second, per stock. For each venue, statistics on the average number of transactions and limit order updates per minute are shown, per venue. The stocks in our sample are Aviva, Hsbc Holdings Plc, Itv Plc, Kingfisher, Lonmin, National Grid, Pearson, Sage group, Vedanta Resources and Xstrata.

	Mean	Stdev	Max
	Panel A: Stock characteristics		
Market Cap	21,062	37,827	125,930
Price	8.06	7.13	23.3
Cumvolume	527	554	1,883
Turnover	3,104	4,188	13,650
Volatility	3.92	1.38	6.10
Share LSE	66.2	5.24	73.8
Share Chi-X	20.5	3.52	25.8
Share Bats	6.44	1.68	9.49
Share Turq.	5.17	1.07	6.29
Share Nasdaq	1.77	0.65	2.73
	Panel B: Limit order book		
Trades LSE	5.16	9.70	347
Trades Chi-X	4.32	7.46	283
Trades Bats	1.80	3.59	84
Trades Turq.	1.19	2.41	51
Trades Nasdaq	0.61	1.54	56
Limits LSE	159.84	216.31	6,999
Limits Chi-X	218.46	373.08	11,934
Limits Bats	123.92	231.60	7,308
Limits Turq.	98.25	146.95	3,296
Limits Nasdaq	75.63	154.66	7,081

Table (3) Summary statistics of trading volume per minute.
Buy and sell volumes in GBPs are reported, showing averages of volume per minute, equal weighted over 10 stocks and 10 trading days. The left panel shows all volume and the right panel volume executed against hidden limit orders.

	Volume			Hidden		
	Mean	Stdev	Max	Mean	Stdev	Max
Sell LSE	39,960	131,673	12,910,331	2,692	35,785	4,001,096
Sell Chi-X	17,492	43,216	990,636	578	3,501	155,871
Sell Bats	5,909	18,996	1,149,016	798	4,344	150,729
Sell Turq	4,082	12,535	349,188	425	2,719	86,744
Sell Nasdaq	1,429	6,667	482,222	8	358	30,040
Buy LSE	41,989	149,532	12,280,297	3,122	37,028	1,644,776
Buy Chi-X	17,674	44,352	1,186,893	625	3,944	318,701
Buy Bats	5,404	18,715	1,536,919	760	4,217	209,244
Buy Turq	4,088	13,377	396,196	532	3,317	126,089
Buy Nasdaq	1,553	7,034	342,203	26	845	65,950

Table (4) Summary statistics Depth(X) measure.
Summary statistics are presented using limit order book data, containing one observation per tenth of second per stock, for 10 trading days in November 2009. The statistics are equally weighted over observations. The mean, standard deviation and maximum of Depth(10) and Depth(50) on the ask and bid side are shown. The Depth(10) on the ask side reflects the available liquidity, in pounds, offered with prices in the interval of Midpoint and Midpoint + 10bps. Similarly, the Depth(10) on the bid side reflects the liquidity offered with prices between Midpoint - 10bps and Midpoint. Depth(50) sums liquidity within 50 basis points from the midpoint. The trading venues are LSE, Chi-X, Bats Trading, Turquoise and Nasdaq OMX Europe.

Ask side	Mean	Stdev	Max
Depth(10) LSE	66,620	117,961	8,940,000
Depth(10) Chi-X	57,693	78,948	1,100,000
Depth(10) Bats	26,311	39,358	527,947
Depth(10) Turq.	14,915	18,817	411,974
Depth(10) Nasdaq	12,824	21,345	213,733
Depth(50) LSE	446,687	371,316	9,040,000
Depth(50) Chi-X	268,467	210,066	1,650,000
Depth(50) Bats	88,363	94,112	800,103
Depth(50) Turq.	71,495	47,170	570,524
Depth(50) Nasdaq	58,096	44,992	332,529
Bid side			
Depth(10) LSE	63,244	90,327	4,420,000
Depth(10) Chi-X	55,907	77,345	1,020,000
Depth(10) Bats	25,404	38,187	553,743
Depth(10) Turq.	14,273	19,277	525,517
Depth(10) Nasdaq	13,241	22,891	538,609
Depth(50) LSE	431,658	339,853	5,100,000
Depth(50) Chi-X	270,657	213,624	1,790,000
Depth(50) Bats	85,748	92,792	1,170,000
Depth(50) Turq.	74,360	51,850	638,623
Depth(50) Nasdaq	61,156	51,443	664,181

Table (5) The cumulative impact of turnover on Depth(10).

Each column represents one regression, showing the cumulative effect over time of buy and sell turnover on changes in DepthAsk(10) and DepthBid(10) of one venue. Changes in the DepthAsk(10) reflect changes in liquidity offered with prices in the interval of Midpoint and Midpoint + 10bps. These changes stem from limit order book activity (placement, cancellations and execution of limit orders). The data consist of one observation per tenth of a second, for each stock. The independent variables are contemporeneous and lagged buy and sell trading volumes for each of the five venues, in GBP. We show the cumulative effect over time (i.e., the running sum) of current trades, and trades one and ten seconds ago. Accordingly, each panel shows the immediate, short and long-term effects of one venues transactions on another venues liquidity. The regressions also contain executed hidden volume as control variables (not reported for brevity). Standard errors are clustered per firm - halfhour, a single asterix indicates significance at the 1% level.

		Ask Side					Bid Side				
£ *Buys*	Sec	LSE	Chi-X	Bats	Turq.	Nasdaq	LSE	Chi-X	Bats	Turq.	Nasdaq
LSE	0	-0.83*	-0.25*	-0.09*	-0.02*	-0.02*	0.28*	0.24*	0.09*	0.02*	0.01*
LSE	1	-0.80*	-0.30*	-0.14*	-0.04*	-0.05*	0.35*	0.31*	0.15*	0.05*	0.05*
LSE	10	-0.67*	-0.18*	-0.05*	-0.03*	-0.04*	0.33*	0.23*	0.09*	0.04*	0.04*
Chi-X	0	-0.21*	-1.31*	-0.18*	-0.02*	-0.03*	0.26*	0.68*	0.18*	0.03*	0.03*
Chi-X	1	-0.52*	-1.47*	-0.46*	-0.09*	-0.13*	0.50*	1.00*	0.45*	0.13*	0.12*
Chi-X	10	-0.61*	-1.29*	-0.37*	-0.09*	-0.13*	0.67*	1.11*	0.46*	0.15*	0.14*
Bats	0	-0.27*	-0.58*	-1.26*	-0.04*	-0.10*	0.37*	0.60*	0.51*	0.05*	0.08*
Bats	1	-0.46*	-0.79*	-1.21*	-0.07*	-0.17*	0.52*	0.88*	0.69*	0.09*	0.16*
Bats	10	-0.54*	-0.83*	-1.01*	-0.08*	-0.15*	0.87*	1.16*	0.81*	0.14*	0.16*
Turq	0	-0.04	-0.04*	-0.05*	-0.70*	-0.03*	0.13*	0.11*	0.08*	0.14*	0.04*
Turq	1	-0.11*	-0.08*	-0.02	-0.69*	-0.04*	0.22*	0.15*	0.06*	0.19*	0.08*
Turq	10	-0.13	-0.06	0.04	-0.68*	-0.02	0.17	0.13	0.01	0.18*	0.07*
Nasdaq	0	-0.03	0.03	-0.01	0.01	-0.75*	-0.05	-0.08	0.04	0.07*	0.14*
Nasdaq	1	-0.08	0.08	0.04	0.04	-0.63*	0.00	-0.18	-0.08	0.02	0.12*
Nasdaq	10	-0.24	0.02	0.07	0.06	-0.62*	0.43	-0.19	-0.11	0.06	0.19*
£ *Sells*											
LSE	0	0.30*	0.27*	0.10*	0.02*	0.02*	-0.78*	-0.29*	-0.10*	-0.02*	-0.02*
LSE	1	0.38*	0.35*	0.17*	0.06*	0.06*	-0.75*	-0.35*	-0.16*	-0.05*	-0.06*
LSE	10	0.39*	0.29*	0.12*	0.05*	0.05*	-0.65*	-0.23*	-0.06*	-0.04*	-0.04*
Chi-X	0	0.27*	0.70*	0.19*	0.02*	0.02*	-0.24*	-1.28*	-0.20*	-0.02*	-0.03*
Chi-X	1	0.46*	0.99*	0.43*	0.12*	0.11*	-0.51*	-1.40*	-0.43*	-0.08*	-0.13*
Chi-X	10	0.57*	1.08*	0.44*	0.13*	0.12*	-0.53*	-1.15*	-0.31*	-0.08*	-0.13*
Bats	0	0.29*	0.56*	0.43*	0.04*	0.08*	-0.21*	-0.58*	-1.15*	-0.05*	-0.10*
Bats	1	0.37*	0.81*	0.62*	0.07*	0.15*	-0.38*	-0.77*	-1.11*	-0.09*	-0.16*
Bats	10	0.53*	1.00*	0.70*	0.09*	0.16*	-0.41*	-0.69*	-0.88*	-0.08*	-0.15*
Turq	0	0.10*	0.08*	0.08*	0.13*	0.04*	-0.06*	-0.03	-0.07*	-0.63*	-0.05*
Turq	1	0.17*	0.12*	0.06*	0.18*	0.08*	-0.08	-0.03	-0.06*	-0.63*	-0.03*
Turq	10	0.33*	0.11	-0.02	0.20*	0.09*	-0.17	-0.02	0.00	-0.59*	-0.04
Nasdaq	0	0.21*	0.04	0.12*	0.11*	0.16*	-0.24*	-0.04	-0.05	-0.03	-0.82*
Nasdaq	1	0.22	-0.01	0.07	0.11*	0.20*	-0.43*	-0.04	0.00	-0.03	-0.76*
Nasdaq	10	0.35	-0.14	-0.05	0.09*	0.23*	-0.84*	-0.18	-0.00	-0.03	-0.68*
R2		.123	.115	.069	.025	.026	.104	.113	.071	.022	.027

Table (6) **Descriptive statistics.**

Descriptive statistics of the independent variables of regression (11) are shown. The variables are constructed per stock-hour. The top panel shows Frag, defined as 1 - HHI based on trading volume, and the LSE liquidity share (the LSE Depth(10) divided by the consolidated Depth(10)). Next is the logarithm of the number of trades on entrants, and the Order Imbalance, defined as the logarithm of buy volume minus the logarithm of sell volume. Ln(Turnover) the logarithm of turnover summed over all venues, and Ln(Depth(10) Cons) is the logarithm of the consolidated Depth(10). Last is the realized volatility, defined as the sum of squared 5 minute returns measured hour-by-hour. The bottom panel shows the cancellation effect of a buy (sell) trade of one venue on the consolidated ask (bid) liquidity. The cancellation effect is the impact (summed over one second) of a 1 share buy-trade at one venue on market wide liquidity offered on the ask sides. These variables are the coefficients of regression (11), estimed per stock-hour using as dependent variable the consolidated Depth(10) on the ask or bid side. We only use estimates when the venue in question executes $>$50 trades and $>$£100.000, and winsorize at the 99% level. The reported variables are winsorized at the 1% level, and .

	Mean	Stdev	Max
Frag	0.59	0.09	0.78
Liq Share LSE	0.40	0.11	0.95
Ln(entrant trades)	5.63	0.98	8.38
Order Imbalance	-0.01	0.43	4.28
Ln(Turnover)	14.99	1.38	19.04
Ln(Depth(10) Cons)	12.43	1.01	16.09
Realized Vol	0.02	0.12	0.96
Cancellation effect			
Buy-Ask LSE	1.44	0.71	3.70
Buy-Ask Chi-X	2.07	1.11	5.73
Buy-Ask Bats	1.96	1.74	7.19
Buy-Ask Turq	1.14	1.65	7.57
Buy-Ask Nasdaq	1.44	3.31	14.69
Sell-Bid LSE	1.53	0.77	4.18
Sell-Bid Chi-X	1.99	1.06	5.48
Sell-Bid Bats	1.91	1.68	7.20
Sell-Bid Turq	0.95	1.50	5.41
Sell-Bid Nasdaq	1.51	2.99	14.62
Observations	1890		

Table (7) Drivers of the cancellation effect.

We show the variables that determine the cancellation effect of trades on consolidated liquidity, per trading venue. The cancellation effect is the impact (summed over one second) of a venues buy-trades on market wide liquidity offered on the ask side (top panel). The bottom panel shows the impact of sell-trades on liquidity on the bid side. This effect is estimated per stock-hour in regression (11), using as dependent variable the consolidated Depth(10) on the ask (bid) side. These coefficients are the dependent variables in the regressions below. We add the following independent variables. Frag is defined as 1 - HHI based on trading volume, and Liq Share LSE is the LSE liquidity share (the LSE Depth(10) divided by the consolidated Depth(10)). Ln(entrant trades) is the logarithm of the number of trades on entrant venues and Ln(Turnover) the logarithm of market wide turnover. Ln(Depth(10) Cons) is the logarithm of the consolidated Depth(10), and Realized Vol is the realized volatility defined as the sum of squared 5 minute returns measured per hour-by-hour. We add firm and day fixed effects. We apply robust Newey-West standard errors (HAC) with 9 lags, t-statistics are shown in parantheses.

	LSE	Chi-X	Bats	LSE	Chi-X	Bats
	Buy trades on consolidated Ask liquidity					
	(1)	(2)	(3)	(4)	(5)	(6)
Frag	1.147***	-1.087**	-0.788	-0.520*	-3.358***	-4.742***
	(4.09)	(-2.39)	(-0.71)	(-1.74)	(-4.31)	(-3.59)
Liq Share LSE	-2.261***	-3.659***	-3.130***	-1.775***	-3.110***	-2.275**
	(-9.05)	(-8.84)	(-3.36)	(-7.79)	(-7.21)	(-2.49)
Ln(entrant trades)				0.585***	0.927***	1.750***
				(7.10)	(4.71)	(4.97)
Order Imbalance				-0.291***	-0.0314	-0.341*
				(-7.00)	(-0.43)	(-1.85)
Ln(Turnover)				-0.640***	-0.774***	-1.535***
				(-9.10)	(-4.21)	(-5.05)
Ln(Depth(10) cons)				0.423***	0.506***	0.408**
				(10.93)	(6.62)	(2.20)
Realized Vol				-0.288**	0.0833	-1.107
				(-2.33)	(0.42)	(-1.10)
Observations	1,718	1,689	1,058	1,718	1,689	1,058
R2	0.143	0.078	0.018	0.261	0.143	0.047
	Sell trades on consolidated Bid liquidity					
	(7)	(8)	(9)	(10)	(11)	(12)
Frag	1.175***	-1.296***	-0.981	-0.891**	-3.028***	-4.790***
	(3.55)	(-2.86)	(-0.81)	(-2.37)	(-4.42)	(-3.03)
Liq Share LSE	-1.413***	-3.911***	-3.001***	-0.602*	-3.437***	-2.336**
	(-3.98)	(-10.33)	(-3.32)	(-1.77)	(-8.93)	(-2.46)
Ln(entrant trades)				0.572***	0.625***	1.654***
				(6.02)	(3.78)	(4.47)
Order Imbalance				0.256***	0.0803	-0.170
				(4.04)	(1.01)	(-0.91)
Ln(Turnover)				-0.651***	-0.597***	-1.461***
				(-7.73)	(-3.98)	(-4.72)
Ln(Depth(10) cons)				0.607***	0.467***	0.277
				(11.12)	(6.89)	(1.54)
Realized Vol				-0.185	-0.374**	0.287
				(-0.92)	(-2.12)	(0.17)
Observations	1,718	1,689	1,057	1,718	1,689	1,057
R2	0.067	0.090	0.017	0.211	0.139	0.043