BIS Working Papers

No 1115

Sharks in the dark: quantifying HFT dark pool latency arbitrage

by Matteo Aquilina, Sean Foley, Peter O'Neill and Thomas Ruf

Monetary and Economic Department

August 2023

JEL classification: D47, G10, G14.

Keywords: high-frequency trading, dark pools, latency arbitrage, stale quotes, reference prices.

This publication is available on the BIS website (www.bis.org).

ISSN 1020-0959 (print)
ISSN 1682-7678 (online)

Sharks in the Dark: Quantifying HFT Dark Pool Latency Arbitrage*

Matteo Aquilina,[†] Sean Foley,[‡] Peter O'Neill[§] and Thomas Ruf[¶]

Abstract

We investigate stale reference pricing and liquidity provision in dark pools using proprietary, participant-level regulatory data. We show a substantial amount of stale trading occurs, imposing large costs on passive dark pool participants. Consistent with these costs, HFTs almost never provide liquidity in the dark, instead frequently consuming liquidity, in particular in order to take advantage of stale reference prices. Finally, we show that market design interventions randomizing dark execution times are successful at countering dark pool latency arbitrage, protecting passive providers of dark liquidity. Our results have substantial implications for practitioners and policymakers aiming to improve liquidity provision in dark pools.

Keywords: high-frequency trading, dark pools, latency arbitrage, stale quotes, reference prices

*We would like to thank Michael Aitken, Peter Andrews, Patrick Augustin, Hao Ming Chen, Brian Eyles, Simon Hargreaves, Edwin Schooling Latter, Ted Macdonald, Albert Menkveld, Richard Payne, Talis Putnins, Jia Shao, Patrick Spens, Felix Suntheim, Martin Taylor, Carla Ysusi and Bart Z. Yueshen. We are also grateful to participants of the FCA/LSE conference on financial regulation and the 5th Behavioural and Financial Markets Conference for their comments. The opinions expressed are those of the authors and not of the Financial Conduct Authority or the Bank for International Settlements. Any errors and omissions are our own. This article is an extension of the FCA Occasional Paper 21 'Asymmetries in Dark Pool Reference Prices.'

†Bank for International Settlements. e-mail: matteo.aquilina@bis.org

‡Macquarie University. e-mail: sean.foley@mq.edu.au

§Financial Conduct Authority (UK) and University of New South Wales.
e-mail:*peter.oneill@unsw.edu.au* Peter O'Neill thanks the Capital Markets Cooperative Research Center for funding a period as a visiting researcher at the FCA during this study.

¶University of New South Wales.

1 Introduction

A defining feature of modern financial markets is the fierce competition on speed, which determines liquidity providers' profitability (Baron, Brogaard, Hagströmer, and Kirilenko, 2019) and investors' trading costs. In such a world, fast traders impose adverse selection costs on slower ones by acting on information signals faster than their competitors (Foucault, Hombert, and Roşu, 2016). New information can be complex, i.e. its impact on prices may not be clear, resulting in competition in both the interpretation of, and reaction to, new information (Budish, Cramton, and Shim, 2015). However, when the consequence of the signal for prices is certain, a race occurs based on reaction times alone. These situations, variously labeled 'toxic arbitrage' (Foucault, Kozhan, and Tham, 2016) or 'latency arbitrage' (Bartlett and McCrary, 2019), are harmful to liquidity in financial markets when they increase adverse selection for liquidity providers (Shkilko and Sokolov, 2020).

Dark pools are uniquely susceptible to this latency arbitrage for two reasons: they rely on outside prices from a reference market in order to determine execution prices, and dark liquidity is often pegged to this reference price. Changes in reference prices creates an 'old' current price and a 'new' price once the update is propogated to the dark venue, a situation which does not exist on lit venues. While prices may vary between lit venues, some uncertainty remains whether a quote update in market A will be followed by convergence in market B or reversal in market A. By contrast, in situations where the dark pool reference price (and its pegged dark limit orders) are stale, fast traders aware of the new price can engage in near risk-free latency arbitrage by trading at these now stale prices at the expense of other market participants.

In this study, we quantify the extent of stale trading in dark pools and the costs it imposes on investors. Furthermore, our unique data set spanning complete order book information (user-level trade and quote entry) for multiple dark pools allows us to identify the market participants most often responsible for latency arbitrage, making inferences about their liquidity provision strategies of different groups of participants in dark pools possible. Lastly, we examine the effectiveness of two market design solutions — randomized dark uncross times, and frequent batch auctions —

aimed at addressing the negative consequences of the speed race. We show that measures designed to protect passive liquidity provision in dark pools can successfully reduce the negative implications and costs of latency arbitrage. This study contributes to the literature on the design of modern markets that operate at high frequency (with persistent race conditions) and sheds light on the merits of mechanisms to 'slow down' markets, such as 'speed bumps', batch auctions, and minimum resting times.

Due to data availability, there has been limited research into stale trading in the dark. A study by the Canadian securities regulator IIROC (Anderson, Devani, and Zhang, 2016) finds that 4 percent of all dark pool trades in Canada occur at stale prices. The Tabb Group consultancy analyzed ten months of trading data for a large buy-side firm, finding that midpoint trades were priced at the far touch or worse (i.e. at or behind the best bid or ask) 11 percent of the time (Alexander, Giordano, and Brooks, 2015). In a related strand of literature, Ding, Hanna, and Hendershott (2014) find frequent deviations between the US NBBO reference feed (the SIP) and the NBBO constructed from proprietary feeds, proxying for the differing 'views' of the market that slow and fast traders experience.

Using a conservative identification, we document that at least 4 percent of dark trading in our UK-based sample occurs at stale reference prices, in line with the results of Anderson, Devani, and Zhang (2016). We find an average cost of 2.4 basis points for executions at stale prices in the dark and estimate the total annual cost to the passive side across all UK dark pools from this type of latency arbitrage to be GBP4.2 million for the year 2014. To provide a point of reference, Hasbrouck (2018) estimates the loss of slow to fast traders in lit markets at up to 2.2 basis points.

Speed differentials between fast and slow participants in lit trading venues have been a hotly debated topic for some time. Some studies disagree whether these differences benefit or harm market liquidity as there are two opposing effects: On the one hand, speed allows fast liquidity providers to quickly update their stale quotes and thus reduces their adverse selection risk, which in turn lowers transaction costs for other market participants (Brogaard, Hagströmer, Nordén, and Riordan, 2015). On the other hand, fast liquidity takers, called latency arbitrageurs or 'high-frequency bandits' (Menkveld and Zoican, 2017), will try to beat liquidity providers in this race and execute against their stale quotes before they can update them. A number of studies (Foucault, Kozhan, and Tham, 2016; Shkilko and Sokolov, 2020) have documented reduced liquidity and increased transactions costs as a consequence. Budish, Cramton, and Shim (2015) argue that arbitrage bandits in aggregate have the advantage because the (single) liquidity provider has to beat *all* bandits to avoid adverse selection. Menkveld and Zoican (2017) find that the net benefit to liquidity depends on the relative proportion of bandits vs. uninformed liquidity traders present in the market at any point. Foucault, Kozhan, and Tham (2016) add the relative speed of bandits over liquidity providers as a crucial determinant of this adverse selection cost.

How do these insights transfer to the dark? By construction, not much is known about the identity of traders in dark pools. Our results provide some insights into who is active in dark pools and who provides and consumes liquidity. We show that it is almost exclusively high frequency trading firms that are on the benefiting side of dark pool executions at stale prices (between 96 and 99 percent of the time). Furthermore, stale trading does not happen at random; rather, in the vast majority of cases (83 percent), it is driven by aggressive dark orders from HFTs. This is even more notable given another of our findings: HFTs as a group almost never provide marketable liquidity in the dark (although they do post non-marketable limit orders in some stocks). These data points, taken together, imply very strategic order submissions by HFTs that are mostly liquidity consuming, including those instances where they take advantage of stale reference prices. We are aware of one comparable empirical study by the Australian securities regulator (ASIC), which finds that HFTs were on the 'winning side' of trades in dark pools that referenced stale prices 85% of the time compared with 31–32% for other users, (ASIC, 2015, p. 54).

Regulators have also paid close attention to the negative consequences of the speed race, in particular with regard to reduced liquidity or the public perception that 'markets are rigged' (Bartlett and McCrary, 2019; Lewis, 2014). In the case of dark pools, regulators have fined dark pools operators for systematic mispricings[1]

[1] Goldman Sachs's U.S. dark pool Sigma X briefly used incorrect references prices (Hope, 2014).

and for misleading investors about the nature of the reference pricing.[2] Spurred by the highly competitive equity market landscape, and with support from regulators, some venue operators have proactively implemented market design changes to address latency arbitrage and stale trading in the dark. The two most commonly discussed proposals are speed bumps and batch auctions.

Speed bump mechanisms typically slow down incoming orders but may allow amendments to existing limit orders to pass through without delay, giving liquidity providers a head start over potential arbitrageurs. Currently, speed bumps are used by, among others, U.S. exchange and dark pool IEX (Hu, 2019), the Chicago Stock Exchange, the Alpha Exchange in Canada (Chen, Foley, Goldstein, and Ruf, 2016) and NYSE. In the case of IEX's speed-bump, for instance, inbound orders are delayed by 350 microseconds (IEX, 2015) but the repricing of certain type of pegged dark limit orders is not. Consequently, as long as the reference price feed is not stale by more than 350 microseconds, reference price latency arbitrage can be prevented.[3]

Batch auctions (Baldauf and Mollner, 2020; Budish, Cramton, and Shim, 2015) do away with the conventional design of continuous, sequential trading, which by its very design creates conditions for speed races, and replaces it with the principle of frequent auctions (often sub-second) that mitigate the speed advantages of faster traders, with slower traders granted sufficient time to observe the outcome of the last auction and submit their orders for the next, with no disadvantage to trade priority. No public exchange venue has implemented this mechanism to date, but at least one dark pool operator has: UK based Turquoise introduced a 'random uncross' feature, which it says is beneficial for 'latency sensitive flow' (LSEG, 2017). This design matches eligible resting orders in the dark pool at random points in time, preventing fast traders from exploiting their speed advantage.

A third category of design changes focus on order characteristics. For example,

[2]Barclay's was fined by the SEC for claiming it was using faster direct feeds to price dark pool trades despite actually using slower 'SIP' feeds (SEC, 2016).

[3]This is acknowledged in competing exchange BATS' submission to the SEC regarding IEX's application to become an exchange: 'this 350 microsecond delay provides IEX the ability to update the prices of resting orders that are pegged ... before market participants with faster access to market data can access those now stale prices on IEX' (Swanson, 2015, p. 1).

Deutsche Bank's Super X dark pool stops matching any orders that are stale by more than a second (Deutsche Bank, 2016). In addition, many dark pool operators disclose information on latency and how it is managed.[4]

In an additional contribution to the literature, the final part of this study investigates two market design interventions aimed at curbing latency arbitrage in general, and stale trading in particular. We find that both batch auctions as well as random uncrossings are effective tools in reducing adverse selection costs and negative effects on market liquidity. Our results provide empirical support for the predictions of theoretical models of high frequency arms races (Foucault, Kozhan, and Tham, 2016) as well as those proposing design changes to counter speed races and latency arbitrage (Baldauf and Mollner, 2020).

The remainder of the paper is organized as follows. Institutional details of dark pools and the data sources are discussed in Sections 2 and 3, respectively. Section 4 examines the prevalence of dark pool trades at stale reference prices, the cost to investors, what causes stale reference prices and HFT liquidity provision in dark pools. The effect of stale reference prices on price impact are covered in Section 5. Section 6 investigates two market design interventions aimed at combating costs associated with stale reference prices and Section 7 concludes.

2 Institutional Details

A dark pool is a trading venue which lacks both pre-trade transparency and a traditional limit-order-book structure. While in lit venues market participants can observe the orders submitted by other participants, in dark pools all orders are hidden. The main advantage of submitting an order to a dark pool is that the trader's intention is not revealed to the entire market. Dark pools also offer better prices than are available on the lit market (so called 'price improvement'), typically

[4]For example, Goldman Sach's compares the timestamps in market data feeds 'to the time that a quote is received by SIGMA X (based on GSEC's internal clock). If this process identifies a latency greater than a defined threshold, SIGMA X will automatically suspend crossing functionality in the relevant security' (GSEC, 2016, p. 3).

trading at prices between the best bid and offer (BBO). The main disadvantage of dark pools is execution uncertainty;[5] as it is impossible to know whether there is a willing counterparty, one cannot know beforehand whether a trade will actually take place. Indeed, (Zhu, 2014) argues that information events result in *correlated* liquidity demands, reducing dark pool execution probabilities at times of heightened information asymmetry.

Orders sent to dark pools usually include a price limit – the maximum price at which a participant is willing to buy (or the minimum price at which a participant is willing to sell).[6] However, within the boundaries of these price constraints, the dark pool operator is responsible for determining the price at which trades take place as a direct consequence of the absence of pre-trade transparency. To determine trade prices, dark pools rely on reference prices drawn from lit markets.[7]

Dark pool operators typically have two options to determine which venues to use to calculate a reference price. The first option is to rely on a single venue, usually the 'primary' market (which for the purposes of this analysis is the London Stock Exchange (LSE)). The second option is to consider multiple (lit) venues. In the first case, dark pools use the Best Bid/Offer (BBO) prices available on the LSE. In the second case, operators construct what is known as the European BBO (EBBO), which includes orders from all other venues.[8]

Having chosen a source of reference prices (the BBO or the EBBO), dark pools then choose to match prices at either the quoted midpoint, the BBO, or both.[9] Many

[5] This is because there may not be liquidity available at the desired time to trade, or the liquidity could be 'one-sided', i.e. (Zhu, 2014). For example, at the midpoint there may be a resting sell order rather than buy orders to facilitate sells. In this paper, we examine whether there is also 'price uncertainty' for dark pool executions arising from latency in reference prices.

[6] Similar to lit markets, participants in dark pools may submit a 'market order' without a price that executes at the prevailing BBO (or midpoint). In practice, these are rarely used.

[7] For more evidence of the manipulation of markets to alter reference prices, see Chau, Aspris, Foley, and Malloch (2021), Aspris, Foley, and O'Neill (2020), Foley, Hu, Huang, and Li (2023), and Frino, Ibikunle, Mollica, and Steffen (2018).

[8] The EBBO differs from the National Best Bid or Offer used in the US as it often excludes smaller lit markets, such as Equiduct and Aquis.

[9] The dark pools currently operated by BATS/Chi-X and Turquoise use only the midpoint price. Other dark MTFs, such as ITG Posit, UBS MTF and Goldman Sachs Sigma X, also use the best bid or ask price (depending on the direction of the trade).

jurisdictions have recently prohibited these non-price improving trades,[10] including Europe with the enactment of MiFID II.[11]

Dark pools in the UK and Europe are, from a regulatory perspective, classified as either multilateral trading facilities (MTFs) or broker crossing networks (BCNs). MTFs fall within the scope of MiFID, require regulatory approval, and are independent venues. This makes them subject to requirements of non-discriminatory access, with transparent rulebooks, and usually central clearing (CCPs). BCNs are not regulated under MiFID and are non-independent crossing networks under a broker's direct control. They are thus allowed to discriminate by participant types, with rules disclosed in bilateral contracts. MiFID II prohibits BCN venues forcing operators to register as MTFs or Systematic Internalizers (SIs).[12] In the US, the 'Alternative Trading System' (ATS) framework would apply to MTFs, but also BCNs.

For our analysis, we separate dark pools into three categories. First, MTFs that are operated by lit exchanges, including Turquoise and BATS Europe,[13] which are integrated with their respective lit exchanges infrastructure, and match trades exclusively at the midpoint. Data for these venues contain orders and trades with participant identifiers. Secondly, MTFs (such as Posit, SigmaX and UBS) that are operated by investment banks and other brokers, who match trades at the BBO as well as the midpoint. Available data for these contain less information (only trade level data, not order data). Third, BCNs (such as 'Crossfinder' and 'SuperX') for which specific venues cannot be determined are excluded from the sample used in this study.[14]

[10] See Foley and Putniņš (2016) for an examination of this in Canada and Australia.

[11] MiFID II Article 4(1)(a) and (2).

[12] Neumeier, Gozluklu, Hoffmann, O'Neill, and Suntheim (2021) provide further details on the regulatory changes applicable to dark pools in Europe.

[13] Turquoise is operated by an investment firm, Turquoise Global Holdings, but is is majority owned by London Stock Exchange Group (LSEG).

[14] MiFID allows reporting of these as simply OTC (Over the Counter) trades.

3 Data and Methodology

3.1 Data

We rely on two data sources in this study. First, we use the FCA's proprietary order book and transaction-level data, which are being collected for market monitoring and research purposes. The transaction data contain both lit and dark trades for the four UK trading venues LSE, BATS, Chi-X and Turquoise, timestamped at the millisecond to the clock of each exchange's matching engine.[15]

These venues account for 99.6% of all FTSE 350 on-exchange lit traded volume in the UK, and all exchange-operated dark MTFs, thus providing us with a representative sample of lit and dark trading activity for UK-listed stocks.[16]

Each transaction record includes buyer or seller initiated flags, price, quantity and information on the order type along with the identity of the market participants. The corresponding order book data contain all order messages (submissions, amendments, and cancellations) and allows us to compute reference prices at the millisecond granularity.

Second, we complement the above exchange-based transactions with millisecond time-stamped dark pool trade data from Thomson Reuters Tick History, which includes the broker-operated dark MTFs by UBS, Sigma X (Goldman Sachs), ITG Posit and Instinet Blockmatch. We exclude Liquidnet as their reference prices are determined through bilateral negotiation rather than the MiFID reference price waiver, and Smartpool and Blink MTF as they have negligible volumes. The remaining sample covers 94%[17] of dark MTF (exchange and broker operated) trading.[18]

[15] BATS and Chi-X are part of the same legal entity, having merged in 2012, but they maintain separate order books.

[16] Estimates were calculated for the period 1/1/14 to 30/06/15 for the FTSE 100 and FTSE 250 Index using information from Fidessa (2017).

[17] Liquidnet accounts for 5.42% of dark trading in the FTSE 100 and 8.65% in the FTSE 250. Estimates from Fidessa. Note that Fidessa does not include Goldman Sach's MTF, Sigma X in its estimates (Rosenblatt, 2014).

[18] As BCNs are unregulated venues under MiFID I, post-trade reporting does not require these venues to be named so they are reported as 'OTC' trades. These venues are excluded from the analysis due to the inability to separate OTC trades organized on a BCN from other OTC trades.

Our analysis uses a randomly selected sample of 57 stocks from the FTSE 100 and 57 from the FTSE 250. Together, these two indices make up the FTSE 350, and provide us with a diversified and representative sample of stocks with both relatively high and low levels of liquidity.[19] The opening and closing auction periods in both samples are excluded, as they specifically prohibit dark trading. Although the full order book data spans 2014 and 2015, the analysis is restricted to five separate weeks, spaced approximately two-and-a-half months apart, for computational reasons.[20]

3.2 HFT and Participant Identification

The de-anonymized nature of the data allows for the identification of participants at the firm level, but not at the trading desk or client level. We group participants into HFTs and non-HFTs, and further divide non-HFTs based on whether they use co-location at the primary exchange to investigate the role that the resulting differences in speed have on trading outcomes.[21] The vast majority of HFTs in the sample are co-located, and 99.84% of all dark trades by HFTs emanate from co-located servers. Many participants that are not HFTs are also co-located.

We follow the approach of Aquilina and Ysusi (2016) for identifying HFT participants, resulting in an almost identical list, differing only because of the presence of additional participants in a more recent sample. The criterion used for defining HFTs is that they are a subset of algorithmic trading participants that use proprietary capital to generate returns using computer algorithms and low-latency infrastructure.

Objective measures of HFTs have been proposed by Hagströmer and Nordén (2013) and Kirilenko, Kyle, Samadi, and Tuzun (2017), such as high order-to-trade ratios and inventory mean reversion. These measures aim to proxy for characteris-

[19] These stocks were selected by the FCA using a stratified sample. In unreported results, we verify that our sample of securities do not significantly differ from the population means on variables including trading activity, relative share of LSE and Dark venues, price and volatility.

[20] Specifically, these weeks are the continuous five-day trading weeks starting: 13th of January, 31st of March, 16th of June, and 1st of September in 2014, and the 22nd of June in 2015.

[21] Co-location refers to the placement of a market participant's servers in close physical proximity to an exchange to reduce transmission latency. Information on who uses co-location is obtained from FCA supervisors.

tics that latency sensitive participants may demonstrate, but do not guarantee these participants are truly latency sensitive, nor that others do not exhibit these characteristics. For example, an HFT engaged in predominantly liquidity consuming (aggressive) trading strategies, will have a lower order-to-trade ratio than an HFT engaged in liquidity providing (passive) market-making strategies.

Therefore, the use of the FCA's internal supervisory knowledge, and the knowledge of the platforms from which the original list was obtained, is recognized as the most accurate means of identification. Many of these firms now identify publicly as HFTs and established their lobby group, 'The Modern Markets Initiative', in 2013. Baron, Brogaard, Hagströmer, and Kirilenko (2019) list a subset of those that can be identified by name, as Sweden provides non-anonymized exchange trading. In our sample, 30 participants are classified as HFTs. The overlap with the publicly disclosed names from the Modern Markets Initiative is very high.

4 Extent and Nature of Stale Trading

In this section, we discuss the prevalence of dark pool trades at stale reference prices in exchange operated dark pools, investigate possible causes, quantify their costs to market participants and discuss the provision of liquidity in dark pools.

4.1 Extent of Stale Trades

Our approach to identifying stale reference prices is relatively simple - we look for situations in which dark midpoint trades occur at prices that are no longer the midpoint, but were until very recently (similar to the methodologies of Anderson, Devani, and Zhang (2016) and ASIC (2015)). First, we identify all trades that take place in dark pools; then we look for quotes on the primary market whose midpoint matches the price at which each dark pool trade took place. We look for a matching quote up to one millisecond *after* the trade time and as far back as necessary to observe a match.[22] If this process returns multiple matches, we use the most recent quote match from the EBBO.[23] For a trade to happen at a stale price, there has to be at least one quote update in the time interval between the matching quote and the trade itself. In other words, if the matching quote does not change before the trade takes place then the trade is not stale, if the matching quote does change then the trade could be stale.[24] In situations with stale trade prices, the matching quote prevails until the new (non-matching) quote arrives. We use the ending timestamp of the matching quote (or equivalently the starting time of the non-matching quote) to measure staleness. Latency for trade i, our measure of staleness, is therefore measured as follows:

$$Latency_i = (TradeTime_i - EndofMatchTime_i) \tag{1}$$

It is easy to see that for non-stale trades the matching quote will still be active at the time of the trade, and that therefore *Latency* would be zero. Recognizing that a one millisecond threshold would not allow for clock synchronicity and timestamp rounding effects, in our analysis, only trades that occur with a *Latency* of at least two milliseconds are classified as stale.[25]

Our methodology eliminates the risk of mis-classifying stale trades, but is likely to significantly understate the extent of stale reference prices if some market participants can react to a quote update in well under 2 milliseconds. If e.g. the shortest possible time for a market participant to receive a quote update from the LSE, process it and send an order to the dark pool is 350 microseconds, participants may be able to successfully race the quote update to the dark venue if prices are stale by more than this amount. Indeed, estimates (Standard and Poors Capital IQ, 2015) put the

[22]This is because one millisecond reflects the upper bound of clock synchronization accuracy provided by the exchanges, and the minimum granularity of timestamps provided to us.

[23]This is conservative as it reduces the number of stale trades we identify.

[24]We are only interested in quotes that change the midquote price - updates to depth will not change the midpoint, and so are ignored.

[25]Compared to Anderson, Devani, and Zhang (2016), we require latency to be at least two milliseconds rather than one and restrict our analysis to dark pools referencing only the primary exchange, making identifying matches more certain and diminishing clock synchronization issues inherent in timestamps from different markets.

minimum transmission time for the dark venues in our sample at the time of our sample period at less than 500 microseconds.

Figure 1: Histogram of Stale Reference Price Age

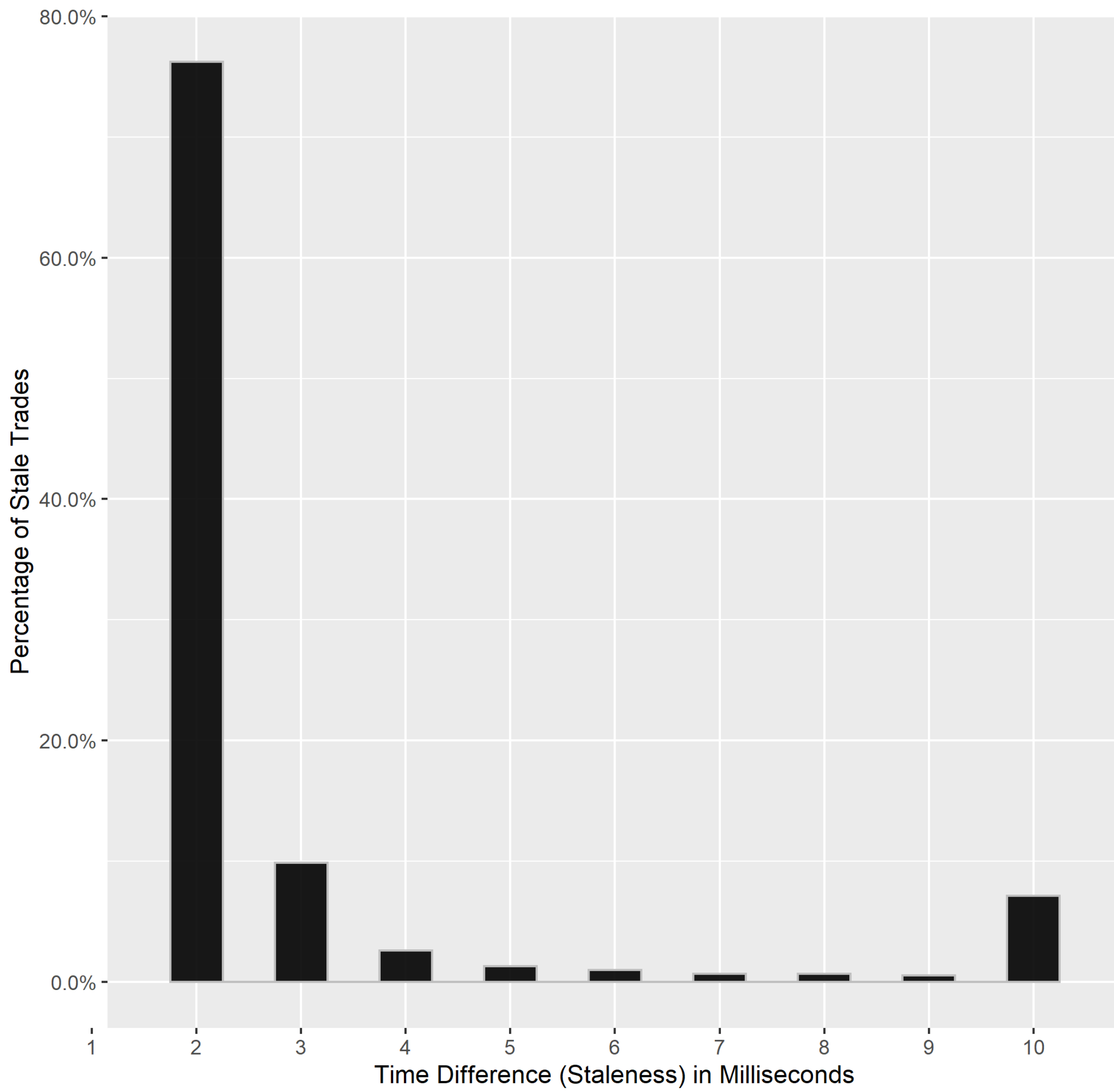


Figure 1 supports the notion that our identification is likely conservative. It displays the distribution of *Latency*, our measure of staleness, which required at least a 2 millisecond lag between reference quote and time of trade in order to classify a trade as stale. We find that over 70% of all trades classified as stale occur 2 milliseconds after the reference quote was superseded by an update. A lag of 3 millisecond only occurs in 10% of cases and larger lags are rarer still. This would indicate that we are likely missing a significant amount of stale trades below the 2 millisecond lag due to our timestamp granularity.

The upper panel of Figure 2 reports the proportion of stale dark trades against the market capitalization (in logs) of individual securities averaged across the entire sample period; the lower panel reports this proportion against the average nominal share price (in logs). We make two salient observations: first, the average proportion of stale dark trades varies significantly by security, between 0% and 7.8%. Second, we observe a strong positive correlation of stale trading with market capitalization and a relatively weak correlation with nominal stock price.

Given that, based on our analysis reported later on, almost all stale trading involves HFTs, this observation is potentially explained by that fact that HFTs have a known preference for highly liquid, high volume stocks which allows them to quickly unwind in the lit market any position they assume in a dark pool.

Figure 2: Proportion of Stale Dark Trades by Stock (%) Observations are calculated as sample means per stock and presented against the natural log of the sample mean market capitalization in pound sterling and the natural log of the stock's sample mean share price on the x-axis in Panel A and B respectively.

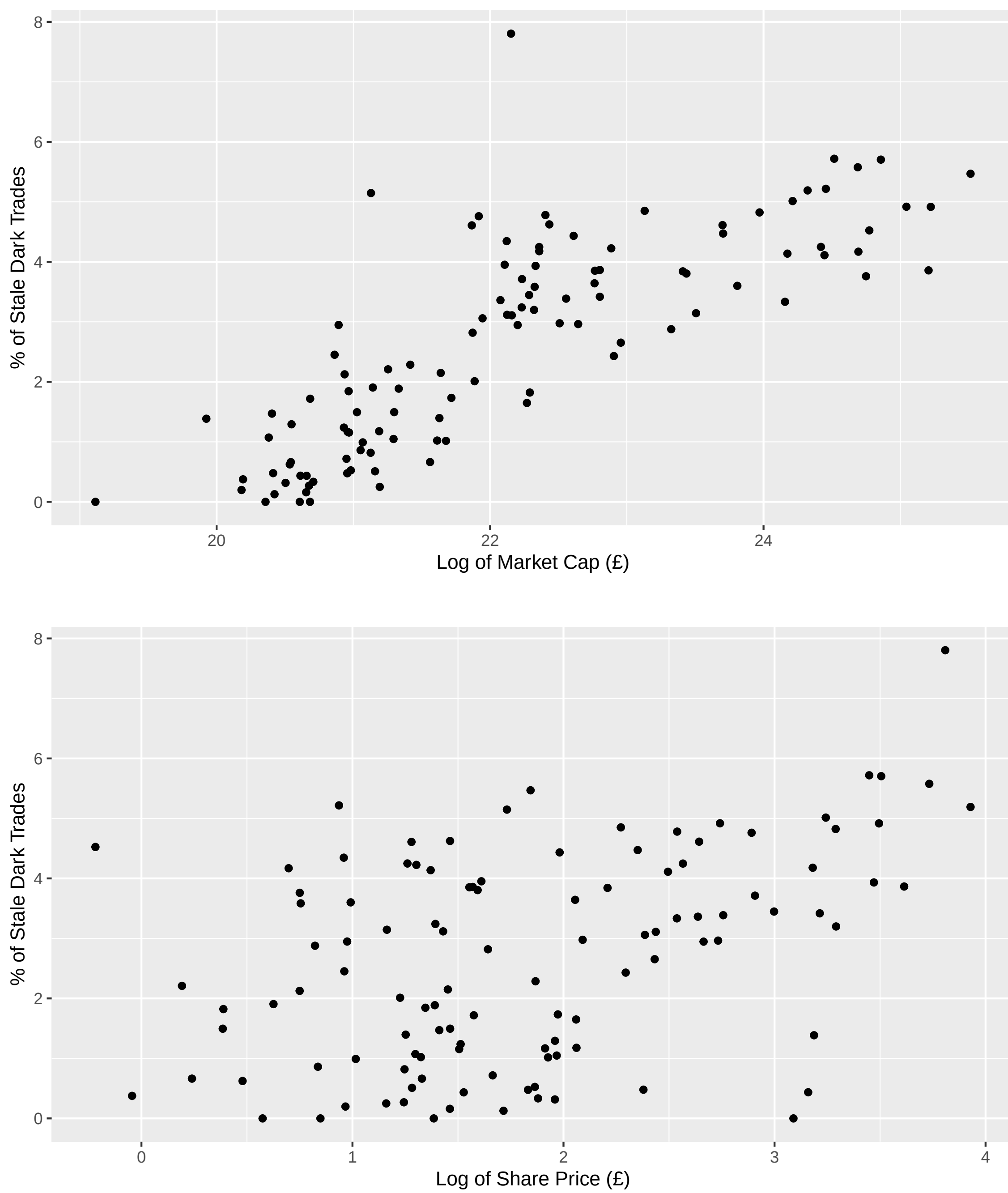


Table 1 contains a number of sample characteristics of our trade-by-trade data. Overall, we identify 3.94% of all trades as stale over the sample period, almost identical to IIROC's figure of 4% for Canada (Anderson, Devani, and Zhang, 2016). This value appears to be trending upwards over time, averaging 3.36% in 2014 and 4.05% in 2015. This upward trend in latency might seem at odds with the overall trend in the marketplace of lower exchange latency and increasing participant speed. However, as we will demonstrate in Figure 3 and Table 2, there is a positive relationship between message traffic and latency. Therefore, if exchange upgrades do not keep pace with message traffic, an increase in latency will be observed.

Table 1: Descriptive Statistics:
This table presents descriptive statistics of the trade-by-trade data used in price impact regressions. The data originates from the matching engines four UK trading venues from five one week periods over January 2014 to June 2015. The venues include LSE, BATS, Chi-X and Turquoise. The total number of observations is 723,979 for each variable. $StaleTrade$ is a dummy variable taking the value of 1 if a trade is deemed to be stale, we multiply by 100 to express as a percentage in this table. $PriceImpactBps$ is the price impact of a trade, calculated using orderbook aggressor flags as at the midpoint 100 milliseconds after the trade. $Quantity$ is the trade quantity in 100s of shares and $Consideration$ is the trade value in GBP. $Staleness$ is the age in milliseconds of the stale quote being referenced. $QuotedSpread$ is calculated from the EBBO at the time of the trade in basis points. $VFTSE$ is the volatility index of the FTSE100 which is taken from options on FTSE100 stocks.

Variable	Mean	Std. Dev	p5	p95
Stale Trade (%)	3.94	19.50	0.00	0.00
PriceImpactBps (100 ms)	0.91	2.69	-1.59	4.87
Quantity (Shares 100s)	1.49	5.40	0.04	6.06
Consideration (£1000s)	6.90	14.15	0.43	19.93
Staleness (ms)	5.98	106	1.00	3.00
Quoted Spread	6.89	8.07	1.50	18.20
VFTSE	12.39	0.99	11.08	14.31

4.2 Cause of Stale Trades

Having established a significant proportion of dark trades at stale reference prices, we next seek to identify the factors responsible for the occurrence of stale trades. Issues of message delays and latency have been shown to contribute to market fragility and as such have become a point of interest among regulators and academics in the wake of several major market malfunctions such as the 2010 Flash Crash. For example, Menkveld and Yueshen (2019) demonstrate that delays in market data feeds during the 2010 Flash Crash may have contributed to the massive deterioration in liquidity at the heart of the episode via a rare breakdown in the crucial arbitrage relationship between the E-mini S&P 500 Futures and the SPY S&P 500 ETF. The importance of processing latency in markets is also established by Kirilenko and Lamacie (2015) who find that processing latency predicts the volatility of asset prices. Exchanges have also been increasingly forthcoming about this relationship. The Eurex exchange states that large bursts of incoming messages can delay the dissemination of quote updates by many milliseconds, for periods lasting over 3 milliseconds (Eurex, 2016, p. 28). The CME has introduced 'First in First Out' priority at the exchange gateway to ensure equitable outcomes amongst participants due to variances in processing time (CME, 2014).

Motivated by these concerns among academics and exchange operators, we investigate whether message volume plays a role in the prevalence of stale dark trades. Because all sample stocks share the same infrastructure, it is not the message traffic in a single stock that matters, but message traffic across all stocks. We begin by establishing an association between overall message traffic and the percentage of stale trades we identify. We estimate several specifications of the following equation:

$$Percstale_{s,d,v} = Messages_d + Controls_{s,d,v} + \epsilon_{s,d,v} \tag{2}$$

where $Percstale_{s,d,v}$ is the percentage of stale trades in stock s on day d and venue v and $Messages_d$ is the average number of messages across two millisecond time intervals on date d across the top 400 stocks on all venues. Stock-day-venue tuples with less than 40 trades are dropped from the analysis to reduce noise from unrepresentative observations.

Table 2 shows those results. We find that the general level of messages is positively associated with the proportion of stale trading. This points to large volumes of quotes slowing down the matching engine, consistent with the examination of 'quote stuffing' by Egginton, Van Ness, and Van Ness (2016). We also find evidence that increased volatility within the FTSE 100 increases stale dark trades, consistent with the notion that rapidly moving prices create more opportunities for predatory dark trading. While returns are insignificantly related to stale trades, this is consistent with pronounced directional movements generating correlated dark midpoint trading, which would render dark latency arbitrage unprofitable (Zhu, 2014). Finally, wider quoted spreads are associated with fewer stale dark trades. This reflects the slim margins associated with latency arbitrage - a movement in the midpoint of a tick-constrained stock necessarily represents an arbitrage opportunity of at least half the spread. However, a quote update in a multi-tick spread may not move the midpoint sufficiently to result in an arbitrage opportunity.[26]

The evidence presented in Column 1 of Table 2 indicates that a higher incidence of message traffic results in a greater proportion of stale trades. Whilst this evidence is not *causal*, we attempt to better identify this relationship by partitioning our sample. In columns 2 and 3 we partition our sample into days with above (below) median levels of message traffic. If message traffic truly has a causal impact on stale trades, we would expect to see significant results only on days with many trades (column 2) and insignificant results on days in which there are few trades (and hence unconstrained capacity on the matching engine).[27] Our results exactly mirror this conjecture: stale trades are heavily correlated with message traffic in the above-median days, but exhibit no correlation on low-message traffic days. Extending this analysis, columns 4-8 further break days down into quintiles of message traffic activity — from low activity in column 4 to high activity in column 8. Consistent with the extreme nature of activity required to burden a messaging engine documented

[26]Dyhrberg, Foley, and Svec (2022) demonstrate how wide tick ranges impact trader behaviour.

[27]Foley, Krekel, Mollica, and Svec (2023) provide direct evidence that message traffic can delay cryptocurrency matching engines by up to 15 milliseconds.

by Egginton, Van Ness, and Van Ness (2016), we observe a large and significant relationship between message traffic and stale quotes *only* on those days within the highest quintile of message traffic activity. This evidence supports the causal relationship between message traffic and stale trades.

Table 2: Regression of Stale Trades over Time
This table reports coefficient estimates for the percentage of stale trades across all venues for each stock date in the sample using the following specification:

$$Percstale_{s,d} = Messages_d + VFTSE_d + Spread_{s,d} + PriceVol_{s,d} + TimeTrend + FE_{s,v} + \epsilon_{s,d}$$

where $Percstale_{s,d,v}$ is the percentage of stale trades in stock s on day d and venue v. Stock-day-venue tuples with less than 40 trades are dropped from the analysis to reduce noise from unrepresentative observations. $Messages_d$ is the average number of messages across two millisecond time intervals on date d across the top 400 stocks on all venues. $VFTSE_d$ is the natural log of the average FTSE Volatility Index. $Spread_{s,d}$ is the time weighted quoted spread for stock s and day d in pence. $Return_{s,d}$ is the % change in a stock's opening and closing price. $Trend_d$ is a time trend variable which increments one for each calendar date in the sample. FE_s and FE_v are stock and venue fixed effects, respectively. Columns 2-8 report regressions on different subsets of the sample. Columns 2 and 3 report regressions on dates where $Messages_d$ is above and below the median. Columns 3-8 report quintile subsets by daily message traffic. Standard errors are clustered at the stock, date and venue level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	% Stale	% Stale	% Stale	% Stale	% Stale	% stale	% Stale	% Stale
VARIABLES	Full Sample	Above Median	Below Median	Low	Quintile 2	Quintile 3	Quintile 4	High
Messages	0.002**	0.005***	-0.002	-0.007	-0.018	-0.028	-0.006	0.009***
	(2.521)	(3.024)	(-0.869)	(-1.486)	(-0.933)	(-0.640)	(-0.715)	(4.619)
VFTSE	0.030***	0.041***	0.038***	0.061	0.052**	0.143	0.095***	0.037
	(3.223)	(3.103)	(2.755)	(1.486)	(2.124)	(1.488)	(3.467)	(1.621)
Quoted Spread	-0.581***	-0.473	-0.613***	-0.698	-0.557**	-0.495	-0.363	0.556
	(-3.258)	(-1.487)	(-3.142)	(-1.361)	(-2.301)	(-1.542)	(-1.532)	(0.525)
Return	0.000	0.000	0.000	0.001	0.001	0.001	-0.001	0.003*
	(0.362)	(0.384)	(0.220)	(0.883)	(1.115)	(0.786)	(-1.327)	(1.934)
Trend	-0.000	-0.000**	0.000	0.000	0.000	-0.001*	0.000	0.001
	(-0.372)	(-2.319)	(0.020)	(1.086)	(0.925)	(-1.729)	(1.072)	(0.629)
Constant	-0.049**	-0.073**	-0.054	-0.080	0.017	-0.154	-0.162*	-0.101**
	(-2.254)	(-2.223)	(-1.608)	(-0.773)	(0.148)	(-0.599)	(-1.765)	(-2.033)
Venue Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Stock Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Observations	4,144	2,025	2,119	765	853	818	861	847
R-squared	0.445	0.504	0.444	0.499	0.520	0.480	0.553	0.676

With the previous test we have established a strong association between the overall amount of message traffic and the percentage of trades we identify as stale. That specification however fails to recognize that potential delays in message transmission play out at a much more granular time frame (i.e. milliseconds) and it also suffer from potential endogeneity concerns: for instance it is plausible that days with considerable amount of news result in many stale trades (as investors digest the news) and in a high number of messages. We therefore exploit the detailed data at our disposal and attempt to show that an increased amount of messages reaching the exchange causally results in a higher likelihood of trades being stale. In order to do so, we run a probit regression at the individual transaction level using the following specification:

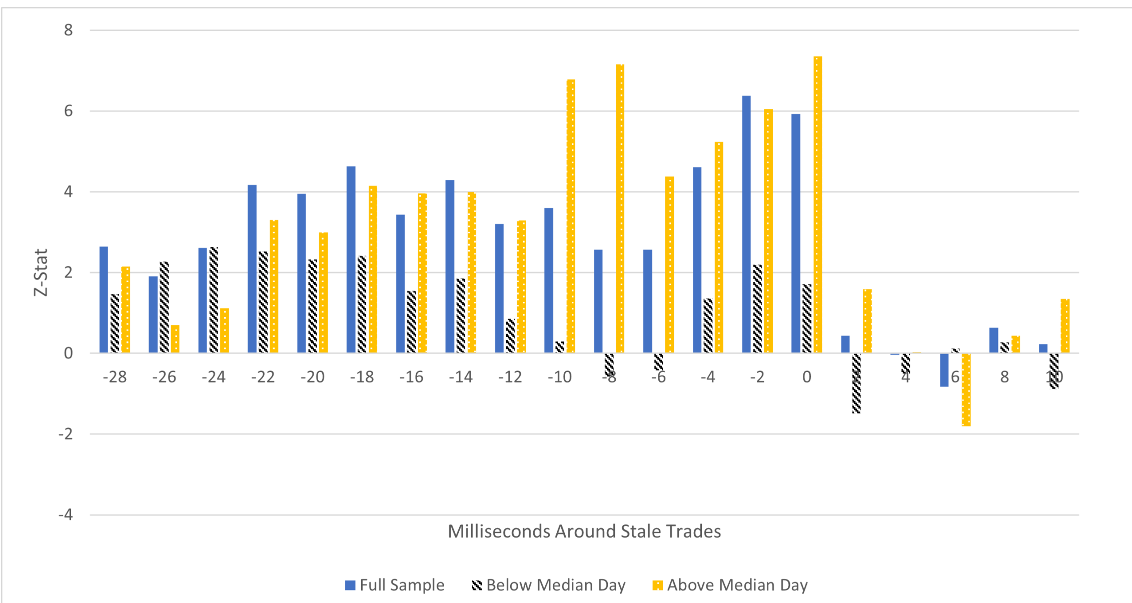
$$StaleTrade_{i,s,d,v} = \alpha + \beta \sum_{\tau=-10}^{-2} TotMktMessages_{d,s,t(i)+2*\tau} + FE_d + FE_v + \epsilon_{i,s,d,v} \quad (3)$$

$StaleTrade_{i,s,d,v}$ is a dummy variable which takes a value of one for a dark trade which is stale, and zero otherwise, for trade i in stock s on date d and venue v. $TotMktMessages_{d,s,t(i)+2*\tau}$ is total number of order messages between 10 millisecond and 2 milliseconds prior to the trade $t(i)$. Venue and date fixed effects are also included and standard errors are clustered at the date level. We use lagged values of the number of messages on the market excluding messages pertaining to the stock for which the trade we are analyzing occurs to avoid endogeneity concerns. We find strongly statistically significant results. As a further check, we estimate the same regression but with the total number of messages in the 10 milliseconds after the trade occurs and do not find statistically significant results. This provides evidence that stale trades are caused by exchange or participant infrastructure not having sufficient capacity to process large influxes in messages without inducing latency.

Figure 3 examines this relationship by reporting the Z-statistics of $TotMktMessages$ in each interval in graphical form.

The evidence in Table 2 leads us to believe that any observed relationship between message traffic and stale trades should only exist during periods of extreme message traffic. To test this, we re-estimate our regressions on two partitions of the data: days with above(below) median message traffic. The resulting Z-statistics are also presented in Figure 3. Consistent with our findings in Table 2, the amount of message traffic in the 30 milliseconds prior to a trade shows no significant relationship for the below median partition. The above median partition, however, exhibits a strong, statistically significant relationship with stale trades from 22 to 0 milliseconds before the trade.

Figure 3: Z-Statistics of Probit Regression of Stale Trades Against Market Messages
This chart reports robust z-statistics of a probit regressions of dark trades in the sample. The dependent variable $StaleTrade$ takes the value one stale dark trades, and zero otherwise. This is regressed against $TotMktMessages$, which is the change in the total number of messages in the 2 millisecond period around the trade time (14 intervals prior to 5 intervals after), excluding messages in the stock in which the trade occurs. The 95% critical value is annotated in a dashed horizontal line. Robust z-statistics in parentheses. The pseudo Rˆ2 is 0.1012 and the number of observations in the full sample is 1,041,340. The regression is also performed for days in which the mean message traffic is above the median day for the sample, and below.



4.3 Measuring the Cost of Stale Trades

For every stale dark trade, one party loses while another gains on the transaction. To measure the cost of stale reference prices to the disadvantaged party, the absolute value of the difference between the trade price and the LSE midpoint price at the time of trade i is multiplied by the volume of the dark trade:

$$Cost_i = |(TradePrice_i - LSEMidTradeTime_i)| * Volume_i \tag{4}$$

This reflects the cost relative to the counter factual of the reference price not being stale, assuming that the dark trade would have otherwise occurred.

Table 3: Costs of Stale Reference Prices
The following table presents estimates of the cost of dark pool latency arbitrage to users of dark pools in the UK. This is presented separately for two subsets of stale trades. Stale reference prices that are inside (or outside) the Primary BBO, which sum to the total column. Outside BBO costs represents real opportunity costs, whilst inside BBO trades represent assumed opportunity costs. The first row expresses the average cost of a dark pool trade at a stale price in basis points (bps). This is calculated as, for trade i, $Cost_i/Consideration_i * 10000$. The following assumptions are made: 253 trading days in the year when scaling the 25 days in the sample; constant proportionality of rates across stocks when scaling 114 stocks to the FTSE 350; constant proportionality when scaling the subset of dark venues to the full population. This subset is estimated to be 32.41% of total dark trading using Rosenblatt's Dark Liquidity report for 2014.

Calculation	**Total**	**Inside BBO**	**Outside BBO**
Average bps per Trade	2.36	1.97	4.31
Total Measured Cost	£44,793	£30,915	£13,878
Scaled Per Year	£453,000	£313,000	£140,000
Scaled to FTSE 350	£1.4m	£928,000	£417,000
Scaled to all UK Dark Venues	£4.2m	£2.9m	£1.3m

Table 3 presents the costs of stale reference prices for the exchange operated dark pools in the sample, finding a total measured cost of £44,793 for the 5-week sample period for our 114 stocks, which scales to approximately £1.4m for the entire FTSE350 for the year 2014. Assuming the prevalence of stale prices was similar in broker operated dark pools, the total yearly cost would be £4.2m across all dark venues for 2014. For comparison, consider the trading revenues of some of the largest

HFTs operating in the UK. Knight Capital Group Europe's trading revenue for 2014 was $83.18m and Jump Trading's gross revenue for 2014 was $97.1m.[28]

This, however, relies on the assumption of a constant level of reference price latency across all of the exchanges in the calculations. This is unlikely to be the case across all dark pools as there is considerable variation even within the sample used. In particular, some broker dark pools - which are not included in our sample as we cannot directly identify the trades that take place there - allow stale reference prices of up to a second in duration.[29] Therefore, our estimate represents a lower bound on the total cost of reference price latency, as it is likely that such costs will be higher on dark pools we do not observe. It should be noted however, that these costs include both trades inside and outside the BBO. Only trades outside the BBO reflect 'pure' arbitrage opportunities - where the losing counterparty could almost certainly have obtained a better price on the lit market due to the resting liquidity. This is not the case with stale trades inside the BBO, for which we assume that the dark trade would still have occurred had the reference price not been stale.

If measured in basis points per trade, costs are also relatively modest at 2.36 bps. A similar figure to Hasbrouck (2018)'s calculation of 1.83 bps lost to fast traders by slower traders in the broader lit market setting of high-frequency quote volatility.

Aside from the direct transaction cost, there may indirect costs as well. First, if activities like latency arbitrage contribute to the perception of a deterioration in 'fairness' in modern markets, this could cause investors to reduce their participation (Guiso, Sapienza, and Zingales, 2008). Second, the results are only based on UK stocks and UK venues. Latency arbitrage for stocks traded in UK-based dark venues and other European lit markets could be considerably higher, given the physical distance between the venues. Costs may also be more economically significant if liquidity providers are dissuaded from providing liquidity in dark venues in response to the costs imposed by stale reference prices. This is empirically examined in Section 5.

[28]International (2015) and Europe (2015) — two of the largest HFTs operating in the UK.

[29]Deutsche Bank Europe's Super X broker crossing network (Deutsche Bank, 2016).

4.4 Are Costs Borne Equally?

A buyer (seller) to the trade benefits from the reference price latency if the trade price is less (greater) than the prevailing mid price. For each of our three categories of participants (HFT, Colocated, Non-Colocated), we compute the proportion of stale dark trades in which they benefit separately (based on number of trades). If all market participants were subject to similar amounts of latency and in the absence of strategic order submission in the dark, we would expect all market participants to benefit and lose from staleness in the dark to a similar degree. Given the large differences in 'proximity to the market, trader sophistication (retail vs. institutional, or subscriber/member vs. public customer), and technology access (Hasbrouck, 2018), it seems unlikely that this is the case.

Table 4 reports how often each of our three trader categories a) participate in stale vs. non-stale dark trading (tab 'Any Side'); b) how often they benefit from staleness and how often each group is on the c) aggressive side or d) passive side.

The first set of results provide evidence against the idea that stale trades randomly 'happen' to market participants. On the contrary, HFT participation in stale trades seems to increase substantially from their participation in non-stale trades from about a quarter to almost half. This implies that they are able to identify latency-affected periods and act on them to their advantage, remaining unaffected by the latency associated with the reference price calculation. This could be because they are using a different (faster) feed from the primary market than the dark pool utilizes.

Can HFTs influence which side they are on? The second tab of Table 4 shows that HFTs benefit in 96% of trades with stale prices inside the BBO (losing only 3.6% of the time due to negligible HFT-on-HFT trading) increasing to 99% for stale trades outside the BBO. These findings are consistent with Baron, Brogaard, Hagströmer, and Kirilenko (2019) who find that HFTs profit in lit markets by using aggressive market orders at the expense of other participants. This result may also be explained by HFT willingness to subscribe to faster market data feeds in addition to faster processing and order submission capabilities.

Table 4: Table: Participation in Dark Trades by Participant - % of All Dark Trades in Respective Category

For the categories, 'any side', 'aggressive side', 'passive side', the proportion of trading by each participant is calculated for 'stale' and 'not-stale' such that they sum to 100%. For the category 'benefit side', for the subset of stale trades, the proportion of the time the participant class is on the 'benefit side' is calculated, such that these do not add to 100%.

	Stale		Not Stale
	Inside BBO	**Outside BBO**	
Any Side			
Non-Colocated	0.19	0.18	0.23
Colocated	0.36	0.34	0.54
HFT	0.45	0.48	0.23
Benefit Side			
Non-Colocated	0.09	0.01	
Colocated	0.12	0.08	
HFT	0.96	0.99	
Aggressive Side			
Non-Colocated	0.05		0.11
Colocated	0.13		0.45
HFT	0.83		0.44
Passive Side			
Non-Colocated	0.33		0.34
Colocated	0.58		0.63
HFT	0.09		0.03

It is worth considering why non-HFTs would purchase colocation to become *faster*, without a realistic prospect of becoming faster than HFTs? Table 4 shows that non-colocated non-HFTs are often on the passive side of trades. While their co-location allows them to reduce their interaction with stale trades relative to their overall liquidity provision (58% of stale vs 63% of non-stale trades) — they cannot really compete with HFTs: HFTs are on the winning side of stale trades 99% of the time, versus 8% of the time (for outside BBO trades) for colocated non-HFTs. However, a costly improvement in speed is still rational if the marginal benefit of the upgrade exceeds its marginal cost (Brogaard, Hagströmer, Nordén, and Riordan, 2015, p. 3413). And indeed, there is still a net benefit – colocated non-HFTs perform better than non-colocated ones (8% benefit versus 1% in Table 4). Given that, as we show in Section 4.5 colocated non-HFTs provide a substantial amount of marketable liquidity in dark pools it is still rational for them to buy such services.

Additional results in Table 4 round out this picture. For non-stale trades, HFTs and colocated non-HFTs are about equally likely to be on the aggressive side (44% vs. 45%). However, for stale trades, HFTs aggressively trade in 83% of the cases, while colocated non-HFTs only manage to do so in 13% of cases, suggesting that HFTs displace out other participant classes due to their superior speed (Aquilina, Budish, and O'Neill, 2022).

Finally, the figures for the passive side show that most dark pool trades execute with non-HFT participants on the passive side. When HFTs do execute passively, they predominantly do so when the reference price is stale. Passive stale trade execution strategies are further discussed in Section 5.

Our findings align well with those in Bartlett and McCrary (2019), which examine how HFTs exploit stale prices in a U.S. setting. They estimate the cost to non-exchange trades(which are analogous to dark MTFs in the UK) that reference these stale feeds as $3.2million USD, in a sample of Dow 30 stocks from August 2015 to June 2016. On $22.17bn traded, this amounts to 1.33 basis points, the same order of magnitude as this study's estimate of 2.36 basis points. They also estimate that the gains to liquidity takers is $3.19m versus $0.24m for gains to liquidity providers, which means 93.1% of all profits can be attributed to aggressive trading participants. These estimates only capture the subset of total latency arbitrage activity in dark pools, and in one country. Aquilina, Budish, and O'Neill (2022) estimate the total scale of latency arbitrage activity on a lit exchange, the London Stock Exchange, in

2015. They extrapolate their estimates to global equities trading, estimating around \$4.8bn total profits in 2018 and that the latency arbitrage activity has a meaningful impact on liquidity, accounting for around a third of the effective spread on UK equities markets.

4.5 Liquidity Provision in Dark Pools

The results presented in Section 4.4 indicate that HFTs are almost always on the side that benefits from stale trades. To provide evidence on the way in which they achieve this, we examine their liquidity provision in dark venues with full orderbook data. We find that HFTs almost never provide resting marketable liquidity in dark pools and therefore that there quotes are almost never stale. This is quite remarkable given that they are significant providers of liquidity on lit venues (Menkveld, 2016). Given that they do not provide significant liquidity in dark pools, HFTs benefit from stale trades by trading against the liquidity provided by other participants.

This study is amongst the first to characterize liquidity provision in modern dark pools. In lit limit order book markets, measures of liquidity are widely accepted — these include quoted spreads, effective spreads, and market depth. As the dark pools in our sample only execute trades at the midpoint of the primary market BBO only depth related measures of liquidity are relevant. Unlike lit limit order markets, dark pools appear to have significant periods of time where liquidity is 'one sided' (liquidity is only available at the bid or the ask, but not both) as well as 'no sided' (liquidity unavailable in the dark). 'No sided' liquidity can occur when there are no dark orders in the order book, but more often occurs when dark orders exist, but only with non-marketable prices. That is, there are buy orders with limit prices below the prevailing midpoint or sell orders with limit prices higher than the prevailing midpoint. In such cases, standard market depth measures are not particularly informative. We therefore propose a new measure of liquidity, defined as the percentage of time there is a bid or ask order available at a marketable price. In practice, this means that the measure of liquidity indicates periods in which people could trade in a dark pool if they were aware of the order's presence.

Our measure of liquidity is calculated in two scenarios: dark orders with quantities greater than those available at the lit market bid or ask, and dark orders with quantities below those in the lit market. This definition of liquidity supply is generous in the context of midpoint dark pools since a participant with a marketable midpoint order may also be interpreted as a consumer of liquidity. Nonetheless, they are a provider of passive dark liquidity: by resting passively on the order book, they facilitate dark executions. Without resting liquidity, participants with aggressive orders merely 'ping' dark pools without executing, like 'ships passing in the night.'[30] Figures 4-5 present cumulative frequency histograms of the dark liquidity metrics calculated for a given stock-day and venue with each chart illustrating a distribution of 8,550 observations (114 stocks x 25 days x 3 dark pools) measuring the percentage of time for a given a stock-venue-day that a class of trader is providing resting liquidity. The figures — presented separately by marketable and non-marketable orders, show that there are many stocks and/or dates, for which it is not possible for participants to access any dark liquidity.

[30]This is a common analogy used in midpoint dark pools, crossing networks and dark aggregators. (Banks, 2014) To mitigate this effect, BATS Europe's dark pools both provide lower execution fees for orders which rest on the order book (Non-IOC orders). See BATS Europe (2017).

Figure 4: Histograms of Marketable Orders in Midpoint Dark Pools

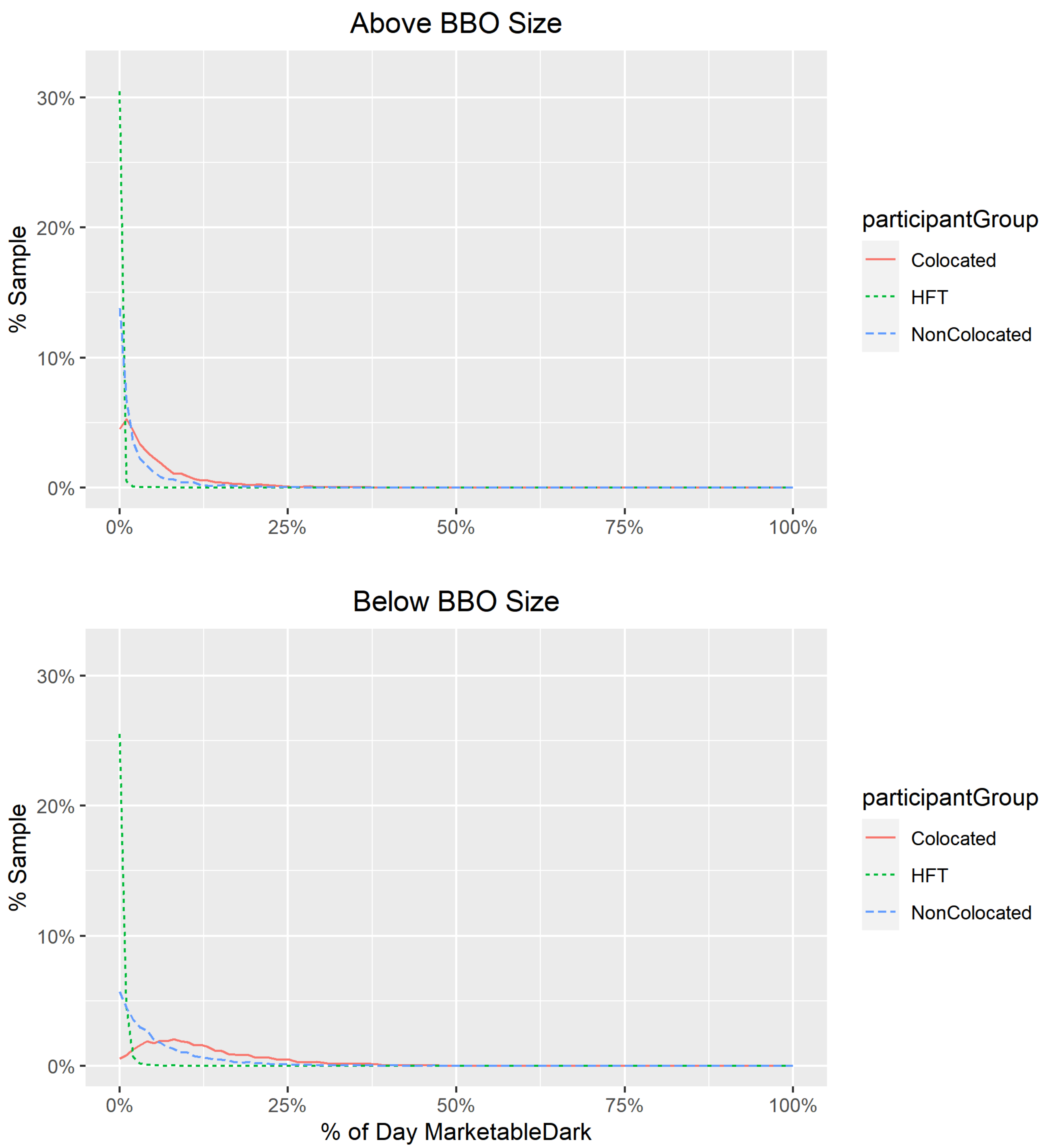
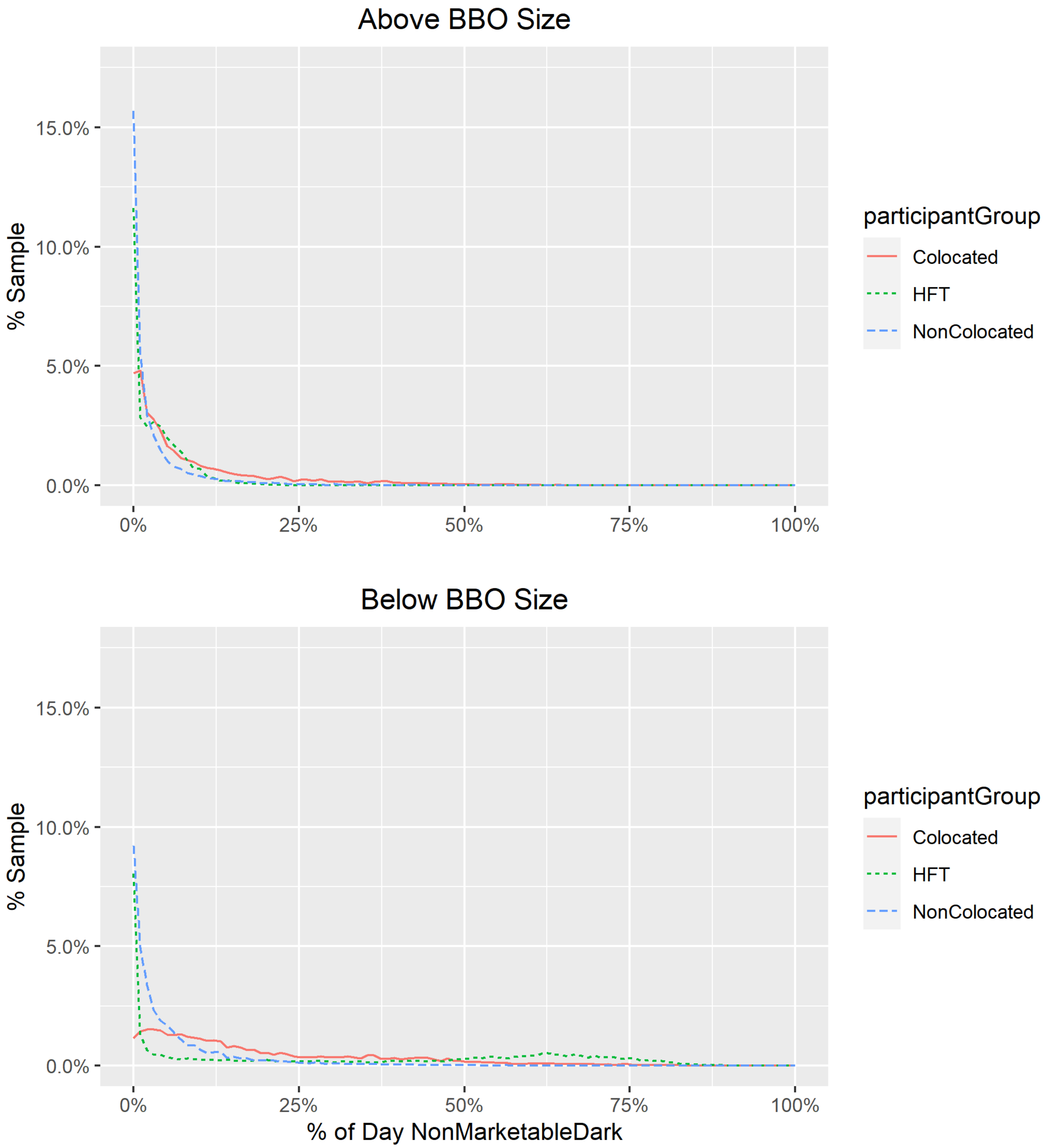


Figure 5: Histograms of Non-Marketable Orders in Midpoint Dark Pools



The largest providers of marketable orders in midpoint dark pools are co-located participants that are not HFTs. Figure 4 shows that Co-located participants (the unbroken line) are the largest suppliers of dark liquidity which is larger than the respective lit Bid or Ask ('Above BBO Size'). This is followed by Non-Co-located participants, (the dashed line). But liquidity is still relatively infrequent, with only a few percent of stock-venue-dates ever having marketable orders above BBO at 10% of the trading day. Co-located participants provide even more liquidity at sizes below the respective BBO size, but the Non-Colocated participant proportion is mostly unchanged. Confirming their role in dark pools is mainly liquidity consumption, HFTs provide very few marketable orders for very small proportions of the day.

Figure 5 presents the same analysis for non-marketable liquidity, orders in the dark pool which are not priced to execute at the current primary midpoint. These orders can be interpreted as 'stop' orders, wherein they only become executable at a given price. At Above BBO sizes, the distribution looks similar to Figure 4, except that HFTs (the broken line) now provide a similar amount of liquidity as co-located participants. And in the below BBO size panel, HFTs provide significantly more liquidity than others at sizes smaller than the BBO relatively frequently, mostly between 50-75% of the day. The difference in liquidity provided by HFT at marketable prices versus non-marketable prices is dramatic. This disparity could result from a strategy of persistently repricing resting orders to be non-marketable, in response to primary market movements. A potential rationale for this behavior is to take advantage of stale reference prices through passive executions, which we describe in Section 5.

5 Stale Trades and the Cost of Liquidity

5.1 The Role of Latency

In Section 4.3 we provided evidence of the overall cost of stale trades by comparing the prices at which stale transactions occur with the current prices available on exchanges. We have also shown that these costs are mostly paid by non-HFT firms.

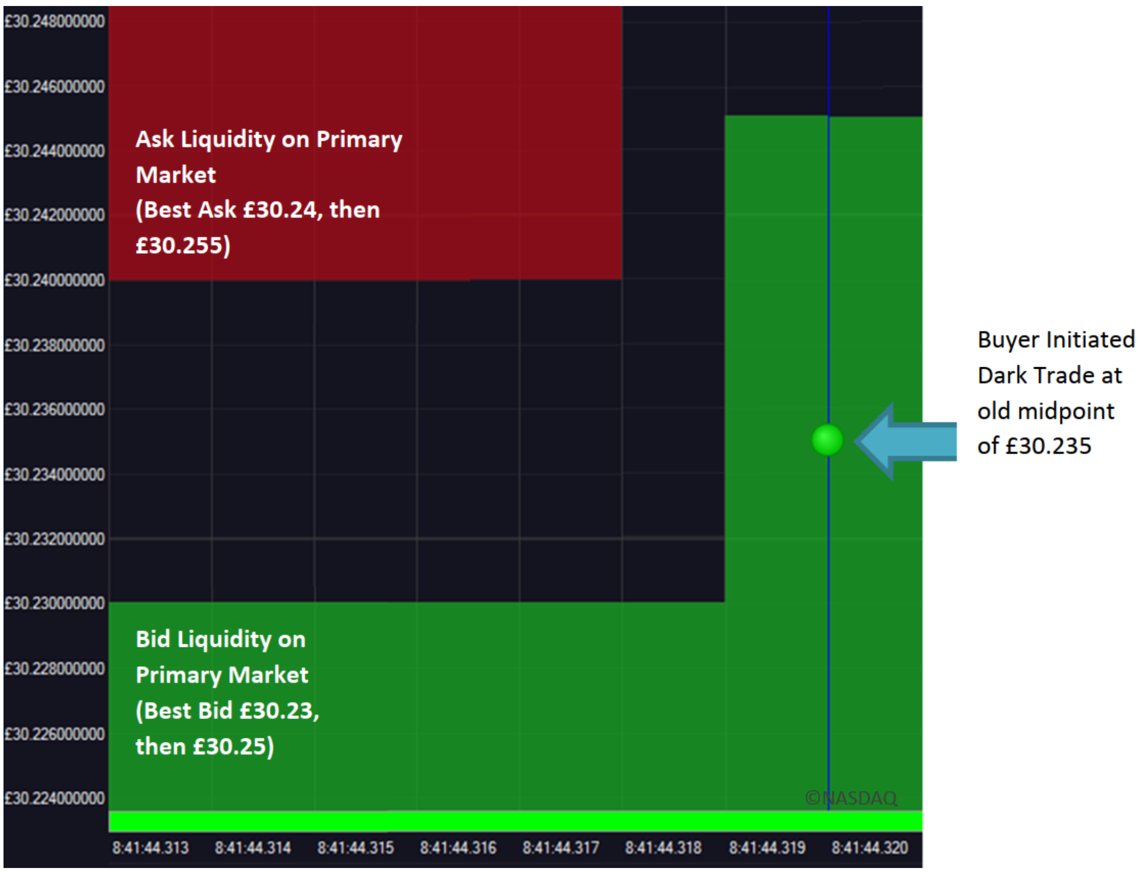
A different way of measuring these costs is to assess to what extent stale trades contribute to the cost of liquidity. In the context of lit markets Aquilina, Budish, and O'Neill (2022) show that latency arbitrage due to symmetric public information deserves a place alongside traditional adverse selection (generated by information asymmetries) as one of the primary components of the cost of supplying liquidity. An analogous argument applies in the context of liquidity provision in dark pools, where the role of symmetric public information is played by the observable price on the lit exchanges. In order to estimate the liquidity cost of stale trades we rely on a measure of price impact.

Traditionally, price impact represents the extent to which prices permanently respond to buying or selling pressure. In the context of stale reference prices, rather than responding to buying or selling pressure, price impact can be measured as the immediate price response after a trade (potentially at a stale price). If the primary midpoint moves up by £1 and a participant has a resting sell order in the dark pool that is still referencing the old price, a stale reference price arbitrageur can submit an aggressive buy order at the old price, resulting in a virtually instantaneous £1 price impact, milliseconds later. Price impact is calculated as:

$$PriceImpact_{it} = q_{it}(m_{i,t+m} - m_{i,t})/m_{i,t} \tag{5}$$

where for each trade t in stock i we measure the change in the midpoint from the time of the trade $m_{i,t}$ to a time m periods after the trade $m_{i,t+m}$, directionalized by q_{it}, the trade direction identifier. Results are reported for the time period of 100 milliseconds, which is viewed as a time period which captures 'instantaneous adverse selection' effects resulting from latency arbitrage activity. When a trade referencing a stale price benefits the aggressor, it will have a positive effect on price impact. If, however, the passive counterparty benefits from the stale price, the price impact is expected to be negative. This is illustrated in Figure 6, where an aggressive buyer trades in the dark at a stale price. As the stale midpoint is much lower than the new midpoint, the buyer (the aggressive side) benefits. When calculating price impact relative to the midpoint at the time of the trade, it is immediate and positive.

Figure 6: Stale Dark Trade Example (Aggressive Benefit)
This figure illustrates a lit market order book over time. The upper edge of the green shaded section represents the Best Bid over time, the lower edge of the red shaded section represents the Best Ask, and the unshaded section in the middle represents 'the spread', in which market participants are unwilling to place resting limit orders to buy or sell, or unable to place prices due to minimum spread requirements. The circle represents an example stale trade on a dark pool. On the lit market, trades would generally execute at the best bid or best ask. Because midpoint dark pools reference the midpoint of the lit market, they should execute in the middle of the unshaded area. Because the trade is referencing the old midpoint, the transaction price falls into the newly green shaded area below the new best bid at the time of the trade.



5.2 Understanding the Determinants of Price Impact

Having defined a measure of the cost of latency arbitrage, we now seek to understand the determinants of these costs. We are specifically interested in quantifying the affect of stale dark trades on the execution quality of associated trades. We further leverage our participant-level data to understand which types of participants exacerbate these costs. As the focus of our study is on the exploitation of short-lived latency arbitrage opportunities in dark pools, we assess price impact within 100 milliseconds of the trade happening — a virtually instantaneous effect. Recently there has been appreciation of the immediacy at which adverse selection can occur (and be observed) in microstructure research. For example, Chen, Foley, Goldstein, and Ruf (2016) examines adverse selection around speed bumps over a 20 millisecond horizon; Conrad and Wahal (2020) examines the decreasing time horizons of price impact, down to 100 milliseconds; and Menkveld (2018) examines trading interactions *within* a millisecond.

To assess the effects of stale trades on the cost of liquidity we follow Malinova and Park (2016) and estimate trade-by-trade regressions, using similar controls. The following OLS regression models are used with standard errors clustered by security, venue and date:

$$\begin{aligned} PriceImpact_{istd+m} = \alpha &+ \beta_1 Stale_{istd} + \beta_2 consideration_{istd} + \beta_3 momentum_{istd} \\ &+\beta_4 aggrHFT_{istd} + \beta_5 aggrColo_{istd} + \beta_6 passHFT_{istd} + \beta_7 passColo_{istd} \\ &+\beta_8 takebook_{istd} + \beta_9 VFTSE_{istd} + \beta_{10} spread_{istd} + \beta_{11} FE_{sd} + \epsilon_{istd} \end{aligned} \tag{6}$$

Where $PriceImpact_{isdt+m}$ is measured for trade i at for stock s at time t on day d, as specified in Equation 3, for $m = 100$ milliseconds after the trade, in basis points.[31] $Stale_{istd}$ refers to a dummy variable with the value of one if the dark trade is stale. We include various controls to proxy for information, liquidity shocks, participant and stock specific factors: $consideration_{istd}$ is the natural log of the value of the trade in British Pound Sterling; $momentum_i$ is the midpoint return in the second prior to the

[31] Results using the EBBO mid-price are both qualitatively and quantitatively unchanged.

trade, multiplied by the trade direction; $aggrHFT_{istd}$ and $aggrColo_{istd}$ are dummy variables capturing whether the aggressive side of the trade is initiated by an HFT or a co-located participant, respectively; $passHFT_{istd}$ and $passColo_{istd}$ are dummy variables representing if an HFT or a co-located participant, respectively, is on the passive side of the trade;[32] $takebook_{istd}$ is a dummy variable with a value of one for cases in which the aggressor counterparty of the dark trade also aggressively executes more than the available liquidity on the LSE within a 2 millisecond period before and after the dark trade — this aims to capture price impact relating to liquidity shocks from participants accessing multiple markets at the same time. $VFTSE_i$ is the natural log of the value of the FTSE 100 volatility index in the 15 seconds prior to the trade;[33] and $spread_{istd}$ is the quoted spread of the EBBO at the time of the trade in basis points. Stock and date fixed effects are also included.[34]

Stale trades are expected to have opposite affects on price impact depending on whether it is the aggressor or the passive provider that benefits from the stale price. To isolate and quantify these opposing effects on price impact, separate regressions are performed which include only stale trades where the benefit side is aggressive, and stale trades where the benefit side is passive (both subsample regressions include the same set of non-stale trades). Table 5 reports the results of the estimated model on three different samples for 100 millisecond price impact.[35] The first column reports estimates from the full sample of trades, both stale and non-stale. The second reports all non-stale trades, as well as only those stale trades in which the aggressive counterparty benefits, while the third reports all non-stale trades together with only those stale trades in which the passive counterparty benefits. The first column shows a highly statistically significant and positive relationship between stale trades and price impact: consistent with latency arbitrage, aggressive benefit stale trades are more numerous than passive, and therefore overall their effect dominates.[36] The

[32] The rationale here is that HFT or co-located participants may be expected to infer information from the dark trade and cause price impact on the lit market in profiting from it.

[33] This index is similar to the VIX in the US, and 15-second intraday closing prices are used.

[34] Price impact is winsorized at 1% and 99%, results remain unchanged without winsorization.

[35] Results are qualitatively the same at 5 seconds and 1 minute.

[36] There are 3,409 trades that benefit the passive side and 25,117 that benefit the aggressor.

model estimates a positive overall price impact of midpoint dark pool trades, of 2.4 basis points (the value of the constant in the first column). Aggressive latency arbitrage increases this cost by 0.88 basis points (or 36%) — the coefficient on the variable *stale* in the second column. This effect is larger in size than the effect on price impact if the aggressor to a dark pool trade also executes against the full LSE best bid or ask (0.50 basis points, the coefficient on the variable *takebook* in the first column). Therefore, the effects of dark latency arbitrage seem to be larger than observed short term liquidity effects.

The variables in this model that identify trade participants — $aggrHFT$ and $aggrColo$ — indicator variables for HFT and co-located participant presence as aggressive counterparties to the trade, demonstrate a positive and statistically significant relationship with price impact. There are two potential explanations for these findings. First, HFT and co-located participants can react faster to information than non co-located participants. Second, they react faster to stale reference prices than non co-located participants, where the $aggrHFT$ variable acts as a proxy for stale trades that cannot be observed due to timestamp limitations. $passHFT$ is strongly related with reduced price impact. This may imply that HFT are adept at avoiding price impact that does not benefit them, as Brogaard, Hagströmer, Nordén, and Riordan (2015) finds for participants that choose co-location. *Spread* is positively related to $PriceImpact$, as it magnifies the effect of bid-ask bounce. We also observe that *takebook* is strongly correlated with higher price impact, as would be expected for trades that consume all BBO liquidity on the LSE at the same time as the dark trade. *Spread* is positively related to $PriceImpact$, as it magnifies the effect of bid-ask bounce. When the stale trades are split by whether the aggressive or passive counterparty benefits, the stale variable has a stronger positive relationship in the aggressive only column, than in the full sample, which pools aggressive and passive together. This implies that the opposing effects are masked in the aggregate. It also demonstrates participants with resting limit orders face higher costs, measured as positive price impact. Conversely, stale trades can also allow aggressive trade initiators to impose costs, by generating negative realized spreads.[37]

[37] Traditional market microstructure theoretical models such as Glosten and Milgrom (1985) model adverse selection as a cost that liquidity providers, or marketmakers, face.

Table 5: Regression of Price Impact and Stale Trades

This table reports coefficient estimates of trade by trade regressions of the price impact of dark trades in the sample using the specification in Formula 6. This model is estimated for three samples, the full sample of dark trades (non-stale and stale), only stale trades which benefit the aggressor counterparty along with non-stale and only stale dark trades that benefit the passive counterparty along with non-stale trades. $PriceImpact_{istd+m}$ is the price impact in basis points for trade i in stock s at time t on date d over the 100 millisecond period m. $Stale_{istd}$ takes the value of one if the dark trade is stale. $aggrHFT_{istd}$ and $aggrColo_{istd}$ take the value of one if the aggressive side of the trade is an HFT or a co-located participant, respectively. $passHFT_{istd}$ and $passColo_{istd}$ is the same for passive, respectively. $takebook_{istd}$ takes the value of of one if the aggressor counterparty consumes the LSE best bid or ask in the same direction. Standard errors are clustered on stock, date and venue.

	(1)	(2)	(3)
	Price Impact (100 Millisecond)		
VARIABLES	Full Sample	Aggresive Benefit	Passive Benefit
Stale	0.649***	0.883***	-1.095***
	(30.873)	(38.179)	(-18.490)
AggressiveHFT	1.511***	1.492***	1.499***
	(77.267)	(76.903)	(77.511)
AggressiveColo	0.466***	0.466***	0.462***
	(30.755)	(30.795)	(31.175)
PassiveHFT	-1.444***	-1.264***	-1.266***
	(-40.691)	(-35.124)	(-36.609)
PassiveColo	0.058***	0.057***	0.063***
	(5.255)	(5.173)	(5.676)
Consideration	-0.039***	-0.039***	-0.039***
	(-11.904)	(-11.820)	(-11.621)
Spread	0.016***	0.016***	0.015***
	(9.762)	(9.792)	(9.164)
VFTSE	-0.783***	-0.775***	-0.755***
	(-3.640)	(-3.610)	(-3.500)
Takebook	0.503***	0.515***	0.484***
	(7.197)	(7.399)	(6.861)
Constant	2.425***	2.402***	2.357***
	(4.457)	(4.419)	(4.318)
Observations	723,979	720,570	698,862
R-squared	0.135	0.134	0.126

Robust t-statistics in parentheses

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

5.3 Alternate Measures of the Cost of Liquidity

Our results thus far have shown that dark latency arbitrageurs are able to impose near-instantaneous price impacts on participants who are willing to supply liquidity in midpoint dark pools. The granular nature of our data, however, allows us to construct a variety of liquidity measures which can shed further light on the impact arbitrageurs have on the efficiency of prices — that is, whether the latency arbitrageurs are truly informed about the future direction of prices. As discussed in Huang and Yueshen (2021), traders can invest in both the acquisition of superior fundamental information, and/or the acquisition of superior speed of execution. Investments in speed are expected to hasten price discovery, but also to reduce the informational efficiency of prices. Prices can fluctuate for both fundamental and non-fundamental reasons, however only price changes due to fundamental reasons will persist. If latency arbitrageurs are informed, we expect their trades against stale prices not only to generate immediate profits, but also to result in higher long-run profitability.

We construct two separate measures to shed light on the type of information dark latency arbitrageurs trade upon. The first is labelled *Dark Profit*. This measures the immediate profit (or loss) a trader would make from buying(selling) at the dark midpoint, and *immediately* unwinding their position at the prevailing ask(bid). More formally:

$$\text{DarkProfit } t_{it} = \begin{cases} \text{Aggressive Buy: } Bid_{LSE,i} - P_{MP,it} \\ \text{Aggressive Sell: } P_{MP,it} - Ask_{LSE,i} \end{cases} \tag{7}$$

In situations of dark latency arbitrage, this measure will be inherently positive, while in situations where the midquote remains stable, no profit will be possible, and the *Dark Profit* will remain negative at half the quoted spread.

Our ability to observe the aggressor in dark trades further allows us to calculate the realized spread for dark midpoint orders. Realized spread is a standard market microstructure measure which computes the returns to liquidity provision over a specified holding period. For example, Conrad and Wahal (2020) show that the introduction of high-frequency traders has compressed the appropriate holding period to assess realized spread for U.S. equities into the sub-second range. Following this evidence, and given the low-latency nature of the trading we are trying to examine, we construct the realized spread after 100 milliseconds, 500 milliseconds, and 1 minute horizons, using the following formula:

$$RealizedSpread_{it} = q_{it}(p_{i,t} - m_{i,t+m})/m_{i,t} \tag{8}$$

Our results are presented in Table 6 and show significant impacts of stale dark trades. Trades executed at stale midpoints result in significantly positive immediate profits (*Dark Profit*), with realized spreads showing significantly reduced realized spreads at horizons of 100 and 500 milliseconds. After one minute, we observe no statistically significant impact of stale trades on the rewards to liquidity provision. This indicates that the latency arbitrageurs focus only on speed acquisition, and are indifferent to fundamental information. Such evidence would suggest that acquiring speed is more of a substitute than a complement for traders, when assessed relative to the theoretical framework of Huang and Yueshen (2021).

Further evidence of the important role HFT's play in this race is provided with the significant increase(decrease) in dark profits when an HFT is aggressive(passive). Consistent with an HFT's ability to adversely select providers of liquidity, the realized spread on orders in which an HFT is aggressive(passive) experience significantly lower(higher) realized spreads at the shorter 100-500 millisecond horizons. Co-located participants are also able to generate significant immediate dark profits, and to impose significant costs on resting dark orders at the 100-500 millisecond horizons, similar to that of HFTs. While there is a highly statistically significant result for both HFTs and co-located participants, co-located participants only appear to capture 20-33% as much of the benefit attributable to HFTs. This could be due to the effects of 'speed races' (documented by Aquilina, Budish, and O'Neill (2022) that HFTs have invested in winning. If HFTs *on average* execute against stale dark orders first, they will exhibit stronger effects than the less specialized co-located traders.

Table 6: Alternative Measures of Dark Trade Profitability

This table reports coefficient estimates of trade by trade regressions of alternative measures of dark trades profitability for robustness purposes. This model is estimated for the full sample of dark trades (non-stale and stale). $DarkProfit$ is calculated as the best bid price on the LSE, less the dark trade price for a buyer-initiated dark trade and the dark trade price less the ask price in stock s at time t on date d, in basis points. Realized spread is calculated as the dark trade price, less the midpoint price m seconds later, in basis points. $Stale_{istd}$ takes the value of one if the dark trade is stale. $aggrHFT_{istd}$ and $aggrColo_{istd}$ take the value of one if the aggressive side of the trade is an HFT or a co-located participant, respectively. $passHFT_{istd}$ and $passColo_{istd}$ is the same for passive, respectively. $takebook_{istd}$ takes the value of of one if the aggressor counterparty consumes the LSE best bid or ask in the same direction. Standard errors are clustered by stock, date and venue.

VARIABLES	(1) Dark Profit (bps) Immediate Full Sample	(2) Realized Spread (bps) 100 milliseconds Full Sample	(3) Realized Spread (bps) 500 milliseconds Full Sample	(4) Realized Spread (bps) 1 minute Full Sample
Stale	0.233***	-0.676***	-0.615***	223.208
	(4.705)	(-25.347)	(-20.429)	(1.065)
AggressiveHFT	1.163***	-1.588***	-1.460***	-317.820
	(30.727)	(-61.001)	(-55.121)	(-1.363)
AggressiveColo	0.243***	-0.496***	-0.416***	-67.510
	(6.847)	(-24.463)	(-19.399)	(-0.594)
PassiveHFT	-1.076***	1.729***	1.674***	147.849
	(-10.726)	(28.623)	(27.145)	(1.210)
PassiveColo	0.090***	-0.061***	-0.084***	72.000
	(4.190)	(-4.969)	(-6.424)	(0.856)
Consideration	-0.063***	0.042***	0.044***	-24.840
	(-7.527)	(9.618)	(9.333)	(-0.628)
Spread	-0.350***	-0.022**	-0.028**	14.267*
	(-17.082)	(-1.978)	(-2.568)	(1.735)
VFTSE	-1.027*	0.891***	0.926***	1,900.788
	(-1.682)	(3.213)	(3.152)	(1.215)
Takebook	0.084	-0.689***	-0.804***	257.787
	(0.445)	(-7.039)	(-7.432)	(1.112)
Constant	1.265	-2.631***	-2.737***	-4,542.608
	(0.827)	(-3.630)	(-3.632)	(-1.243)
Observations	723,979	723,979	723,979	723,979
R-squared	0.403	0.098	0.077	0.001

Robust t-statistics in parentheses

*** p<0.01, ** p<0.05, * p<0.1

6 Eliminating Latency Arbitrage - Market Design Solutions

In the previous sections we have demonstrated that reference price latency is costly for liquidity providers. Unsurprisingly, European exchanges sought to innovate to reduce the impact of dark latency arbitrage on market participants. This Section examines two interventions to reduce reference price latency arbitrage — a randomized uncross of the dark midpoint, and the introduction of a frequent batch auction. Both mechanisms seek to minimize or eliminate the ability of fast dark-latency arbitrageurs to snipe resting dark orders by 'slowing down' the dark matching process. Specifically, executions only occur on the time-scale of seconds, with a randomized duration.

Unfortunately both of these events occur outside of the sample set of regulatory data we have available. Absent the regulatory data used in the prior sections, we rely on public data provided by Thompson Reuters Tick History (TRTH). Given a variety of inferiorities associated with the TRTH data in comparison to regulatory data, we use two alternative proxies to identify the impact of these regulatory changes on dark latency arbitrage.[38] The first is price impact calculated as the absolute value of the price change, (removing q), because the data consists of public trade reports which do not carry initiator flags. Such a measure of instantaneous price impact is well accepted by both academics and practitioners working at high frequency, with industry examples including ITG (Saraiya and Mittal, 2009) and IEX (Aisen, 2015), and academic research including Boni, Brown, and Leach (2013) who examine volatility ahead of dark pool trades, and Nimalendran and Ray (2014) who use changes in the spread. The following difference in differences model is implemented

[38]TRTH provides millisecond time stamps to their data. However, this timestamp is applied by TRTH on receipt. Transmission between the exchange matching engine and the TRTH facility occurs via the public internet. In tests of data quality, we identify this 'jitter' to be in the magnitude of +/- 20 milliseconds, which makes the identification of dark latency arbitrage infeasible.

for each event study:

$$\begin{aligned} PriceImpact_{sdv+m} &= \alpha + \beta_1 event_d + \beta_2 treated_{sv} + \beta_3 eventTreated_{sdv} \\ &+ \beta_4 Spread_{sd} + \beta_5 VIX_d + \beta_6 consideration_{sdv} + \beta_7 FE_{sd} + \epsilon_{sdv} \end{aligned} \tag{9}$$

Where $PriceImpact_{sdv}$ measures the natural log of the consideration-weighted average price impact for stock s date d, for exchange venue v, $m = 100$ milliseconds before and after the trade, in basis points. This is the absolute value of the difference between the NBBO midpoint m before the trade and m after the trade. In addition we use an alternative measure of price impact, which we label $\%PriceMove_{sdv}$ which is the proportion of trades in stock s, date d, venue v for which a change in the midpoint occurs over the same time period m (being 100 milliseconds before and after the trade). This is useful as a significant proportion of dark trades occur with no price movement - without movement in the midquote, we can unambiguously state that there has been no dark latency arbitrage. $Event_d$ is a dummy variable which takes the value of one on and after the event in question, and zero otherwise.[39] $Treated_{sv}$ is a dummy variable which takes the value of one for stocks traded on venues which are subject to the market design change, and zero otherwise. $EventTreated_{sdv}$ is the interaction term for venues which are subject to the treatment, after the event occurs. $Spread_{sd}$ is the stock's time weighted average quoted spread. $Consideration_{sdv}$ is the natural log of the sum of trading consideration in British Pounds or US dollars. Stock and date fixed effects are also included. VIX_d is the natural log of the mean value of the VIX volatility index, or the VFTSE (for UK event studies), measured at 15 seconds intervals throughout the day.

6.1 Turquoise Random Uncross - Event Study

The first event study examines the introduction of a 'random uncross' by the UK based exchange 'Turquoise'. This order type was introduced alongside the existing dark midpoint, effectively creating a separate pool of liquidity. It works by delaying the matching of dark orders, undertaking an 'uncross' of accumulated midpoint

[39] This takes the value of one for the dark venues.

orders at random periods of between 5 and 45 seconds throughout the day, depending on the stock's underlying level of liquidity (Turquoise, 2016). As the uncrossing period is unknown, aggressive participants cannot race reference price update messages. This feature was intended to reduce or eliminate arbitrage costs arising from such races. We use the 'relaunch' of this facility in September 2013 as our event. Turquoise's CEO stated that volumes in uncross were 'up more than three times since the start of September 2013', (Times, 2013). Therefore, this event can be used to determine if adverse selection decreases in relation to other comparable dark venues, using the same sample of stocks in a difference in differences model. The untreated sample are dark trades on the BATS and Chi-X dark pools. We use the same 57 FTSE 100 and 57 FTSE 250 stocks as in our previous analysis. Data is taken from the 1^{st} of June 2013 to the 31^{st} of August 2013 for the pre-event window, and from the 1^{st} of October 2013 to the 30^{th} of December 2013 for the post-event window. Table 7 sets out the results which show a statistically significant reduction in adverse selection after the relaunch of the random uncross. This is roughly a 12% reduction in price impact, and a 4.1% reduction in the proportion of dark trades which experience adverse movements in prices after the trade. These results indicate that this market design intervention is effective at reducing adverse selection driven by reference price latency.

Table 7: Avoiding Latency Arbitrage — Turquoise Uncross and BATS Batch
This table reports coefficient estimates for an event study of the introduction of Turqouise uncross in the first two columns, and estimates from trade-level regressions in the last two columns. The regression uses the specification in Formula 9. The dependent variable, $AdverseSelection_{sdv+m}$ is (a) the trade value weighted price impact in basis points for stock s on date d for venue v over the time interval m (100 milliseconds) before to after the trade and (b) the % of dark trades for which there is a change in the midpoint over the 100 millisecond period before and after the trade, for stock s on date d for venue v over the time interval m. $Event_d$ is a variable which takes the value of one on and after the event date, and zero otherwise. $Treated_{sv}$ is a dummy variable which takes the value of one for stocks traded using Turqouise uncross, and zero otherwise. $EventTreated_{sdv}$ is the interaction term for Turqouise after the introduction of the randomized uncross. In our trade level regressions, $uncross_{sv}$ takes the value of one for trades in the Turquoise dark uncross, and $batch_{sv}$ takes the value of one for trades that are flagged as BATS periodic batch auction trades. Standard errors are clustered by stock, date and venue.

	Uncross		Turqoise and BATS	
VARIABLES	Price Impact	% PriceMove	Price Impact	% PriceMove
	(a)	(b)	(a)	(b)
Event	1.105	0.103		
	(1.498)	(1.212)		
Treated	-0.248***	-0.091***		
	(-12.734)	(-27.544)		
Treated_event	-0.126***	-0.041***		
	(-5.237)	(-9.709)		
Uncross			-1.135***	-0.379***
			(-29.567)	(-93.118)
Batch			-0.090**	-0.163***
			(-1.986)	(-14.317)
Consideration	-0.053**	-0.022***	-0.048***	-0.014***
	(-2.319)	(4.709)	(-6.290)	(-11.921)
Spread	0.000*	-0.000**	0.016***	0.000
	(1.692)	(-2.203)	(8.741)	(0.899)
VIX	5.181*	0.584*	-1.564**	-0.400***
	(1.649)	(1.717)	(-2.209)	(-2473)
Constant	-12.510	-0.593	4.827***	1.444***
	(-1.483)	(-0.676)	(2.840)	(3.715)
Fixed Effects	Stock	Stock	Stock	Stock
	Date	Date	Date	Date
Observations	37,921	39,684	48,079	54,496
Adj. R-squared	0.143	0.246	0.288	0.324

Robust t-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

6.2 Uncross and Batch Auctions: Trade Level Evidence

The second event study exploits the existence of trade-by-trade level data, allowing us to estimate the impact on individual trades execution quality. For this experiment, we examine trading in two distinct mechanisms, both of which are designed to mitigate dark latency arbitrage. The first mechanism is the Turquoise Uncross described in the previous section. The second mechanism is the BATS Europe Batch Auction. This new order type consists of a separate orderbook that performs intra-day auctions with randomized uncross periods throughout the day. These auctions have price, time and size priority, but are only partially pre-trade transparent. There is an indicative uncrossing price and executable volume published in the call period prior to the uncross, but this is only published if the full minimum acceptable quantity of orders is met (BATS Europe, 2017). Whilst this is described as an auction, for stocks which are tick-constrained (the vast majority of our sample), only midpoint executions were allowed.[40] The randomization of the uncrossing of the batch auction is the mechanism which is designed to mitigate the potential for latency arbitrage, similar to the randomized uncross in Turquoise.

A three month period in 2017 (May to July inclusive) is examined, a period where the Turquoise uncross flag is revealed in Bloomberg trade data, and the BATS Periodic Batch auctions are identified by a flag in TRTH data. The identification of individual trades which chose to utilize the Turquoise Randomized Uncross or the BATS Batch Auction allow us to determine whether these mechanisms are effective at reducing the costs associated with dark latency arbitrage, in comparison to other dark orders executed on these venues (and also on Chi-X) which did not make use of these mechanisms. Given the unpredictability of the uncross prevents race conditions after lit market updates, one would expect lower price impact for both Turquoise uncross and BATS batch auctions.

Turquoise trades from Bloomberg are matched to TRTH to obtain millisecond timestamps. Using trade identifiers, price impact is compared for batch trades against dark non-batch trades using the same control variables as in Equation 9.

[40] For more information on tick sizes in Europe, see Foley, Meling, and Ødegaard (2023).

The dummy variables $batch_{sdv}$ and $uncross_{sdv}$ take the value of one for the consideration weighted mean of adverse selection for the stock, date, venue observation for batch and uncross trades respectively. This sample includes all dark trades for FTSE 350 stocks in Chi-X and batch and dark trades in BATS, as well as dark trades and uncross trades in Turquoise.[41] Our results are reported in Table 7. There are statistically significant negative relationships between $batch_{sdv}$ and the adverse selection measures. We observe a significant reduction in price impact of 1.135 basis points for trades using Turquoise uncross, and a 37.9% reduction in the proportion of midpoint changes 100 milliseconds before and after the trade for batch auction trades in comparison to dark trades not using the uncross feature. A similar statistically significant reduction in price impact and adverse price movements is observed for BATS periodic batch auctions, but these are less economically significant at 0.09 basis points and a 16.3% reduction in midpoint changes. These results are consistent with the prediction that the randomized uncross feature renders latency arbitrate strategies ineffective, reducing adverse selection. Given that the randomized uncross introduces a small delay in trading (from instantaneous execution to 5-45 seconds) it would seem that the significant reduction in latency arbitrage costs is well worth this sub-minute increase in trade latency.

7 Conclusion

As the automation and speed of markets has increased, so has the importance of latency in markets, and in particular, latency in reference prices. This study examines the prevalence of stale reference prices, their effect on the cost of liquidity provision and participant strategies with respect to stale dark trades for a representative sample of UK exchange operated midpoint dark pools. Trades are frequently observed at stale reference prices, resulting in asymmetric outcomes across participant types, where faster (HFT) participants impose costs on slower participants. Stale reference prices are shown to play a significant role in increasing price impact and hence the

[41] The results are qualitatively unchanged when Chi-X trades are excluded.

cost for passive liquidity provision by 2.4 basis points.

Using a new measure of liquidity in dark pools, we characterize dark liquidity provision in the UK by different participant classes, finding that HFTs virtually never provide resting (marketable) liquidity. We further contribute to the literature by analyzing the effectiveness of two types of market design innovations to resolve the negative consequences of reference price latency. We conduct two event studies and show that removing the ability for aggressive participants to race the market data feed by making the timing of dark executions unpredictable (through random uncross mechanisms and batch auctions) is effective in reducing the cost of liquidity provision.

Our evidence has significant policy implications. Reference price determination is important to the fairness of dark pool pricing. Regulators should focus on the adoption of alternative market designs like random uncross features and batch auctions. Indeed, it may be the desire to avoid these 'sharks in the dark' which has led to the endogenous adoption of speed bumps and batch auctions by an increasing number of exchange venues around the world.

References

Aisen, Daniel, 2015, The Genesis of an Order Type, Blog.

Alexander, Jeff, Linda Giordano, and David Brooks, 2015, Dark Pool Execution Quality: A Quantitative View, .

Anderson, Lisa, Baiju Devani, and Yifan Zhang, 2016, The Hidden Cost: Reference Price Latencies, Trading Review And Analysis Investment Industry Regulatory Organization of Canada.

Aquilina, Matteo, Eric Budish, and Peter O'Neill, 2022, Quantifying the High-Frequency Trading "Arms Race"*, *The Quarterly Journal of Economics* 137, 493–564.

Aquilina, Matteo, and Carla Ysusi, 2016, Are high-frequency traders anticipating the order flow? Cross-venue evidence from the UK market., *FCA Occasional Papers* 16.

ASIC, 2015, Review of high-frequency trading and dark liquidity, Discussion Paper 452 Australian Securities and Investments Commission.

Aspris, Angelo, Sean Foley, and Peter O'Neill, 2020, Benchmarks in the spotlight: The impact on exchange traded markets, *Journal of Futures Markets* 40, 1691–1710 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/fut.22120.

Baldauf, Markus, and Joshua Mollner, 2020, High-frequency trading and market performance, *The Journal of Finance* 75, 1495–1526.

Banks, E., 2014, *Dark Pools: Off-Exchange Liquidity in an Era of High Frequency, Program, and Algorithmic Trading* (Springer).

Baron, Matthew, Jonathan Brogaard, Björn Hagströmer, and Andrei Kirilenko, 2019, Risk and return in high-frequency trading, *Journal of Financial and Quantitative Analysis* 54, 993–1024.

Bartlett, Robert P., and Justin McCrary, 2019, How rigged are stock markets? Evidence from microsecond timestamps, *Journal of Financial Markets*.

BATS Europe, 2017, BATS Europe Guidance Note: Periodic Auctions Book, .

BATS Europe, 2017, Trading Price List, Discussion paper BATS Europe.

Boni, Leslie, David C. Brown, and J. Chris Leach, 2013, Dark pool exclusivity matters, *Available at SSRN 2055808.*

Brogaard, Jonathan, Björn Hagströmer, Lars Nordén, and Ryan Riordan, 2015, Trading Fast and Slow: Colocation and Liquidity, *The Review of Financial Studies* 28, 3407–3443.

Budish, Eric, Peter Cramton, and John Shim, 2015, The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response, *The Quarterly Journal of Economics* 130, 1547–1621.

Chau, Ching, Angelo Aspris, Sean Foley, and Hamish Malloch, 2021, Quote-Based manipulation of illiquid securities, *Finance Research Letters* 39, 101556.

Chen, Haoming, Sean Foley, Michael A. Goldstein, and Thomas Ruf, 2016, The Value of a Millisecond: Harnessing Information in Fast, Fragmented Markets, SSRN Scholarly Paper ID 2860359 Social Science Research Network Rochester, NY.

CME, 2014, CME Group Customer Forum Q2 2014, .

Conrad, Jennifer, and Sunil Wahal, 2020, The term structure of liquidity provision, *Journal of Financial Economics* 136, 239–259.

Deutsche Bank, 2016, SuperX EMEA: Order type, matching logic and operational summary, Quick Guide.

Ding, Shengwei, John Hanna, and Terrence Hendershott, 2014, How Slow Is the NBBO? A Comparison with Direct Exchange Feeds, *Financial Review* 49, 313–332.

Dyhrberg, Anne H., Sean Foley, and Jiri Svec, 2022, When Bigger is Better: The Impact of a Tiny Tick Size on Undercutting Behavior, *Journal of Financial and Quantitative Analysis* pp. 1–30 Publisher: Cambridge University Press.

Egginton, Jared F., Bonnie F. Van Ness, and Robert A. Van Ness, 2016, Quote stuffing, *Financial Management* 45, 583–608 Publisher: Wiley Online Library.

Eurex, 2016, Insights into trading system dynamics: Eurex Exchange's T7, .

Europe, KCG, 2015, KCG Europe Limited 2015 Financial Statements, Financial Statements KCG Europe.

Fidessa, 2017, Fidessa Fragmentation Index, .

Foley, Sean, Xiaolu Hu, Haozhi (Rachel) Huang, and Jiang Li, 2023, Should Underwriters Be Trusted? Reducing Agency Costs Through Primary Market Supervision, .

Foley, Sean, William Krekel, Vito Mollica, and Jiri Svec, 2023, Not so fast: Identifying and remediating slow and imprecise cryptocurrency exchange data, *Finance Research Letters* 51, 103401.

Foley, Sean, Tom G Meling, and Bernt Arne Ødegaard, 2023, Tick Size Wars: The Market Quality Effects of Pricing Grid Competition, *Review of Finance* 27, 659–692.

Foley, Sean, and Tālis J. Putniņš, 2016, Should we be afraid of the dark? Dark trading and market quality, *Journal of Financial Economics* 122, 456–481.

Foucault, Thierry, Johan Hombert, and Ioanid Roşu, 2016, News Trading and Speed, *The Journal of Finance* 71, 335–382.

Foucault, Thierry, Roman Kozhan, and Wing Wah Tham, 2016, Toxic Arbitrage, *The Review of Financial Studies* 30, 1053–1094.

Frino, Alex, Gbenga Ibikunle, Vito Mollica, and Tom Steffen, 2018, The impact of commodity benchmarks on derivatives markets: The case of the dated Brent assessment and Brent futures, *Journal of Banking & Finance* 95, 27–43.

Glosten, Lawrence R., and Paul R. Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 14, 71–100.

GSEC, 2016, SIGMA X FAQ, Discussion Paper Version 8 Goldman Sachs Execution & Clearing, L.P. ("GSEC").

Guiso, Luigi, Paola Sapienza, and Luigi Zingales, 2008, Trusting In Stock Markets, *The Journal of Finance* 63, 2557–2600.

Hagströmer, Björn, and Lars Nordén, 2013, The diversity of high-frequency traders, *Journal of Financial Markets* 16, 741–770.

Hasbrouck, Joel, 2018, High-frequency quoting: Short-term volatility in bids and offers, *Journal of Financial and Quantitative Analysis* 53, 613–641 Publisher: Cambridge University Press.

Hope, Bradley, 2014, Goldman Agrees to $800,000 Fine Over Dark Pool, *Wall Street Journal*.

Hu, Edwin, 2019, Intentional access delays, market quality, and price discovery: Evidence from iex becoming an exchange, Discussion paper New York University School of Law.

Huang, Shiyang, and Bart Zhou Yueshen, 2021, Speed Acquisition, *Management Science* 67, 3492–3518 Publisher: INFORMS.

IEX, 2015, Investors Exchange Form 1 Submission, Filing to Register as Exchange 3235-0017 Investors' Exchange LLC.

International, Jump Trading, 2015, Jump Trading International 2015 Financial Statements, Financial Statements Jump Trading International.

Kirilenko, Andrei, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun, 2017, The flash crash: High-frequency trading in an electronic market, *The Journal of Finance* 72, 967–998.

Kirilenko, Andrei A., and Gui Lamacie, 2015, Latency and Asset Prices, SSRN Scholarly Paper ID 2546567 Social Science Research Network Rochester, NY.

Lewis, Michael, 2014, *Flash Boys: A Wall Street Revolt.* (W. W. Norton and Company).

LSEG, 2017, Dark Midpoint Order book, .

Malinova, Katya, and Andreas Park, 2016, “Modern” Market Makers, *Working Paper*.

Menkveld, Albert J., 2016, The economics of high-frequency trading: Taking stock, *Annual Review of Financial Economics* 8, 1–24 Publisher: Annual Reviews.

——— , 2018, High-frequency trading as viewed through an electron microscope, *Financial Analysts Journal* 74, 24–31 Publisher: Taylor & Francis.

——— , and Bart Zhou Yueshen, 2019, The Flash Crash: A Cautionary Tale About Highly Fragmented Markets, *Management Science* 65, 4470–4488 Publisher: INFORMS.

Menkveld, Albert J., and Marius A. Zoican, 2017, Need for Speed? Exchange Latency and Liquidity, *The Review of Financial Studies* 30, 1188–1228.

Neumeier, Christian, Arie Gozluklu, Peter Hoffmann, Peter O'Neill, and Felix Suntheim, 2021, Occasional Paper No. 60: Banning Dark Pools: Venue Selection and Investor Trading Costs, .

Nimalendran, Mahendrarajah, and Sugata Ray, 2014, Informational linkages between dark and lit trading venues, *Journal of Financial Markets* 17, 230–261.

Rosenblatt, 2014, Monthly Dark Liquidity Tracker European Edition, Discussion Paper September 2014 Rosenblatt Securities Limited.

Saraiya, Nigam, and Hitesh Mittal, 2009, Understanding and Avoiding Adverse Selection in Dark Pools, .

SEC, 2016, ADMINISTRATIVE PROCEEDING File No. 3-17079, .

Shkilko, Andriy, and Konstantin Sokolov, 2020, Every cloud has a silver lining: Fast trading, microwave connectivity, and trading costs, *The Journal of Finance* 75, 2899–2927 Publisher: Wiley Online Library.

Standard and Poors Capital IQ, 2015, Real Time Solutions Global Network Map, .

Swanson, Eric, 2015, Investors' Exchange, LLC ("IEX"); Notice of Filing of Application, as Amended, for Registration as a National Securities Exchange under Section 6 of the Securities Exchange Act of 1934; Release No. 34-75925, SEC Filing 10- 222. BATS.

Times, Asset Servicing, 2013, Trading Platforms: Azure waters ahead, *Asset Servicing Times.*

Turquoise, 2016, Turquoise Plato Uncross Factsheet, .

Zhu, Haoxiang, 2014, Do dark pools harm price discovery?, *Review of Financial Studies* 27, 747–789.

Previous volumes in this series

1114 August 2023	The term structure of inflation forecasts disagreement and monetary policy transmission	Alessandro Barbera, Fan Dora Xia and Xingyu Sonya Zhu
1113 August 2023	To Lend or Not to Lend: the Bank of Japan's ETF purchase program and securities lending	Mitsuru Katagiri, Junnosuke Shino and Koji Takahashi
1112 July 2023	Trust bridges and money flows	Tobias Adrian, Rodney Garratt, Dong He, and Tommaso Mancini-Griffoli
1111 July 2023	How much do firms need to satisfy the employees? – Evidence from credit spreads and online employee reviews	Koji Takahashi and Sumiko Takaoka
1110 July 2023	Fiscal sources of inflation risk in EMDEs: the role of the external channel	Ryan Banerjee, Valerie Boctor, Aaron Mehrotra and Fabrizio Zampolli
1109 July 2023	Original sin redux: role of duration risk	Carol Bertaut, Valentina Bruno and Hyun Song Shin
1108 July 2023	Innovation convergence	Bryan Hardy and Can Sever
1107 July 2023	Financial heterogeneity and monetary union	Simon Gilchrist, Raphael Schoenle, Jae Sim and Egon Zakrajsek
1106 July 2023	Global public goods, fiscal policy coordination, and welfare in the world economy	Pierre-Richard Agénor and Luiz A Pereira da Silva
1105 June 2023	The demand for government debt	Egemen Eren, Andreas Schrimpf and Fan Dora Xia
1104 June 2023	The Crypto Multiplier	Rodney Garratt and Maarten R C van Oordt
1103 June 2023	Privacy regulation and fintech lending	Sebastian Doerr, Leonardo Gambacorta, Luigi Guiso and Marina Sanchez del Villar
1102 May 2023	MPC Heterogeneity and the Dynamic Response of Consumption to Monetary Policy	Miguel Ampudia, Russell Cooper, Julia Le Blanc and Guozhong Zhu
1101 May 2023	Insights into credit loss rates: a global database	Li Lian Ong, Christian Schmieder, and Min Wei

All volumes are available on our website www.bis.org.