Aquilina, Matteo; Budish, Eric B.; O'Neill, Peter

**Working Paper**

# Quantifying the High-Frequency Trading "Arms Race": A Simple New Methodology and Estimates

Working Paper, No. 300

**Provided in Cooperation with:**
George J. Stigler Center for the Study of the Economy and the State, The University of Chicago Booth School of Business


Chicago Booth Paper No. 20-16

IGM Working Paper No. 173

Stigler New Working Paper No. 45

# Quantifying the High-Frequency Trading "Arms Race": A Simple New Methodology and Estimates

Matteo Aquilina
Financial Conduct Authority and Financial Stability Board

Eric Budish
University of Chicago Booth School of Business

Peter O'Neill
Financial Conduct Authority

Fama-Miller Center for Research in Finance
Stigler Center for the Study of the Economy and the State
Initiative on Global Markets
The University of Chicago, Booth School of Business

This paper also can be downloaded without charge from the
Social Science Research Network Electronic Paper Collection:

http://ssrn.com/abstract=3636323

# Quantifying the High-Frequency Trading "Arms Race": A Simple New Methodology and Estimates*†

Matteo Aquilina‡, Eric Budish§, and Peter O'Neill¶

June 25, 2020

**Abstract**

We use stock exchange *message data* to quantify the negative aspect of high-frequency trading, known as "latency arbitrage." The key difference between message data and widely-familiar limit order book data is that message data contain *attempts* to trade or cancel that *fail*. This allows the researcher to observe both winners and losers in a race, whereas in limit order book data you cannot see the losers, so you cannot directly see the races. We find that latency-arbitrage races are very frequent (about one per minute per symbol for FTSE 100 stocks), extremely fast (the modal race lasts 5-10 millionths of a second), and account for a large portion of overall trading volume (about 20%). Race participation is concentrated, with the top 6 firms accounting for over 80% of all race wins and losses. Most races (about 90%) are won by an aggressive order as opposed to a cancel attempt; market participants outside the top 6 firms disproportionately provide the liquidity that gets taken in races (about 60%). Our main estimates suggest that eliminating latency arbitrage would reduce the market's cost of liquidity by 17% and that the total sums at stake are on the order of $5 billion annually in global equity markets.

*We thank Andrew Bailey, Markus Baldauf, Fabio Braga, Peter Cramton, Karen Croxson, Sean Foley, Joel Hasbrouck, Terrence Hendershott, Stefan Hunt, Anil Kashyap, Robin Lee, Donald MacKenzie, Paul Milgrom, Barry Munson, Brent Neiman, Talis Putnins, Alvin Roth, Edwin Schooling Latter, Makoto Seta, and John Shim for helpful discussions. We are extremely grateful to Matthew O'Keefe, Natalia Drozdoff, Jaume Vives, Jiahao Chen, and Zizhe Xia for extraordinary research assistance. Budish thanks the Fama-Miller Center, Initiative on Global Markets, Stigler Center and Dean's Office at Chicago Booth for funding.

†This paper circulated in January 2020 as a Financial Conduct Authority Occasional Paper, which is the FCA's working paper series. FCA policy is for academic research conducted by FCA staff to first be circulated as an Occasional Paper. The authors are then free, after incorporating feedback, to seek to publish the research in peer-reviewed academic journals. While Occasional Papers do not necessarily represent the position of the FCA, they are one source of evidence that the FCA may use while discharging its functions and to inform its views. Any errors and omissions are the authors' own. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

‡Financial Conduct Authority and Financial Stability Board, matteo.aquilina@fsb.org

§University of Chicago Booth School of Business and NBER, eric.budish@chicagobooth.edu

¶Financial Conduct Authority, peter.oneill@fca.org.uk

"The market is rigged." – Michael Lewis, Flash Boys (Lewis, 2014)

"Widespread latency arbitrage is a myth." – Bill Harts, CEO of the Modern Markets Initiative, an HFT lobbyist (Michaels, 2016)

# 1 Introduction

Flash Boys, in which the seemingly arcane topic of high-frequency trading became a #1 best seller in the hands of Michael Lewis, famously alleged that the U.S. stock market is "rigged for the benefit of insiders." The book's basic claim is that high-frequency trading firms (HFTs) use their speed advantage, combined with complex and opaque market practices, to make large amounts of nearly risk-free profits at the expense of ordinary investors. HFT advocates publicly disparaged the book as a "novel", i.e., a work of fiction, and dismissed speed-based arbitrage as a "myth".[1] In the years since the book's publication in 2014 the academic literature on high-frequency trading has continued to be quite active. While there have been data limitations, discussed in detail shortly, the evidence to date does not support the most alarmist or conspiratorial readings of Flash Boys, nor the notion that concerns about HFT are purely myth.[2] That said, regardless of one's view of the veracity of Flash Boys or HFT advocates, the importance of speed in modern financial markets is uncontroversial. By many estimates, HFT firms account for over 50% of trading volume. A speed race that just a decade ago was commonly measured in milliseconds (thousandths of a second) is now measured in microseconds (millionths) and even nanoseconds (billionths). HFT firms and other parties spend significant sums on microwave links between market centers (because information travels faster through air than glass), trans-oceanic fiber-optic cables (previous communications links were not in a straight line), putting trading algorithms onto hardware as opposed to software (hardware is significantly faster), co-location rights and proprietary data feeds from exchanges (to get updates faster and send trades faster), real estate adjacent to and even on the rooftops of exchanges, and, perhaps most importantly, high-quality human capital.[3]

[1] See Tabb (2014) and Narang (2014) for examples of prominent industry advocates calling Flash Boys a "novel" in print. The authors have heard the phrase "novel" used to refer to Flash Boys many other times in private conversations or at industry conferences. The Modern Markets Initiative used the language "latency arbitrage myth" again in a public statement in response to the first public draft of this paper; see Osipovich (2020).

[2] For surveys of the literature on HFT please see Jones (2013), Biais and Foucault (2014), O'Hara (2015), and Menkveld (2016). Papers with empirical evidence that relates to the benefits and costs of HFT include Hendershott, Jones and Menkveld (2011), Menkveld (2013), Brogaard, Hendershott and Riordan (2014), Brogaard et al. (2015), Budish, Cramton and Shim (2015), Foucault, Kozhan and Tham (2016), Shkilko and Sokolov (2016), Brogaard et al. (2018), Malinova, Park and Riordan (2018), Weller (2018), Van Kervel and Menkveld (2019), and Breckenfelder (2019). Theoretical models that relate to benefits and costs of HFT include Hoffmann (2014), Biais, Foucault and Moinas (2015), Du and Zhu (2017), Pagnotta and Philippon (2018), and Baldauf and Mollner (2020).

[3] Please see Laumonier (2014, 2019) and Laughlin, Aguirre and Grundfest (2014) regarding microwaves, CME Group, Inc. (2019) and Mulholland (2015) regarding the trans-atlantic Hibernia cable, Lockwood et al. (2012) for engineering details regarding the use of FPGA hardware for high-frequency trading, Investors' Exchange (2019) and Budish, Lee and Shim (2019) for details regarding co-location and proprietary data feeds, Baker and Gruley (2019) regarding the fight over real estate adjacent to the CME's Aurora data center, and Virtu Financial, Inc. (2019*b*) regarding the fight over access to the NYSE Mahwah data center's rooftop. Regarding human capital, Virtu's 2018 10-K filing reports average compensation costs of about $445,000 per employee (Virtu Financial, Inc., 2019*a*). Most other HFT firms are privately held but many firms report compensation for their European arms, for example Jump Trading International Limited (2018) implies compensation of $557,000 per employee.

Budish, Cramton and Shim (2015, henceforth BCS) provide a conceptual framework for the role of HFT and the importance of speed in modern financial markets. In the BCS model, the fastest traders endogenously choose to engage in two functions. The first, liquidity provision, is useful. The second, "sniping" stale quotes, also known as "latency arbitrage," is harmful. BCS show that the root cause of latency arbitrage is the design of modern financial exchanges, specifically the combination of (i) treating time as continuous (infinitely divisible) and (ii) processing requests to trade serially (one-at-a-time). These aspects of modern exchange design trace back to the era of human trading (e.g., trading pits, specialist markets), which also used versions of limit order books and price-time priority. But, to a computer, serial processing and time priority mean something much more literal than to a human. The consequence is that even *symmetric public information creates arbitrage rents*. We are all familiar with the idea that if you know something the rest of the market doesn't know, you can make money. BCS showed that even information seen and understood by many market participants essentially simultaneously—e.g., a change in the price of a highly-correlated asset or index, or of the same asset but on a different venue, etc.—creates arbitrage rents too. These rents lead to a never-ending arms race for speed, to be ever-so-slightly faster to react to new public information, and harm investors, because the rents are like a tax on market liquidity. BCS showed that the problem can be fixed with a subtle change to the underlying market design, specifically to discrete-time batch-process auctions; this preserves the useful function of algorithmic trading while eliminating latency arbitrage and the arms race.

Unfortunately, empirical evidence on the overall magnitude of the latency arbitrage problem has been scarce. BCS provide an estimate for one specific trade, S&P 500 futures-ETF arbitrage, and find that this specific trade is worth approximately $75 million per year. Aquilina et al. (2016) focus on stale reference prices in UK dark pools and estimate potential profits of approximately GBP4.2 million per year. The shortcoming of the approach taken in these studies is that it is unclear how to extrapolate from the profits in specific latency arbitrage trades that researchers know how to measure to an overall sense of the magnitudes at stake. Another notable study is Ding, Hanna and Hendershott (2014), who study the frequency and size of differences between prices for the same symbol based on exchanges' direct data feeds and the slower data feed in the U.S. known as the consolidated tape, which is sometimes used to price trades in off-exchange trading (i.e., dark pools). However, as the authors are careful to acknowledge, they do not observe which of these within-symbol price differences are actually exploitable in practice—not all are because of both noise in timestamps and physical limitations due to the speed at which information travels. Wah (2016) and Dewhurst et al. (2019) study the frequency and size of differences between prices for the same symbol across different U.S. equity exchanges. This is conceptually similar to and faces the same challenge as Ding, Hanna and Hendershott (2014), in that neither study observes which within-symbol price discrepancies are actually exploitable. For this reason, the magnitudes obtained in Wah (2016) and Dewhurst et al. (2019) are best understood as upper bounds on the within-symbol subset of latency arbitrage. Brogaard, Hendershott and Riordan (2014) and Baron et al. (2019) compute a large set of HFT firms' overall profits on specific exchanges (in NASDAQ data and Swedish data, respectively), and Baron et al. (2019) show that relatively faster HFTs earn significantly greater

profits, but neither paper provides an estimate for what portion of these firms' trading profits arise due to latency arbitrage.

In the absence of comprehensive empirical evidence, it is hard to know how important a problem latency arbitrage is and hence what the benefits would be from addressing it. Indeed, if the total magnitudes of latency arbitrage are sufficiently small then the HFT lobby's "myth" claim, while perhaps a bit exaggerated, is reasonable. Conversely, if the magnitudes are sufficiently large then "rigged", while perhaps a bit conspiratorial, may be appropriate. Notably, while numerous regulators around the world have investigated HFT in some capacity (e.g., the FCA, ESMA, SEC, CFTC, US Treasury, NY AG), and in a few specific instances have been required to rule specifically on speed bump proposals designed to address latency arbitrage, there are still different perspectives on what are the positive and negative aspects of HFT, and what if any regulatory rules or interventions are appropriate.[4]

This paper uses a simple new kind of data and a simple new methodology to provide a comprehensive measure of latency arbitrage. The data are the "message data" from an exchange, as distinct from widely familiar limit order book datasets such as exchange direct feeds or consolidated datasets like TAQ (Trades and Quotes) or the SEC's MIDAS dataset. Limit order book data provide the complete play-by-play of one or multiple exchanges' limit order books—every new limit order that adds liquidity to the order book, every canceled order, every trade, etc.—often with ultra-precise timestamps. But what is missing are the messages that *do not affect the state of the order book, because they fail*.[5]

For example, if a market participant seeks to snipe a stale quote but fails—their immediate or cancel (IOC) order is unable to execute so it is instead just canceled—their message never affects the state of the limit order book. Or, if a market participant seeks to cancel their order, but fails—they are "too late to cancel"—then their message never affects the state of the limit order book. But in both cases, there is an electronic record of the participant's *attempt* to snipe, or *attempt* to cancel. And, in both cases, there is an electronic record of the exchange's response to the failed message, notifying the participant that they were too late.

Our method relies on the simple insight that these failure messages are a direct empirical signature of speed-sensitive trading. If multiple participants are engaged in a speed race to snipe or cancel stale quotes, then, essentially by definition, some will succeed and some will fail. The essence of a race is that there are winners and losers—but conventional limit order book data doesn't have any record of the losers. This is why it has been so hard to measure latency arbitrage. You can't actually see the race in the available data.

We obtained from the London Stock Exchange (by a request under Section 165 of the Financial

[4]For regulatory investigations of HFT, please see Financial Conduct Authority (2018), Securities and Exchange Commission (2010), European Securities Market Authority (2014), Commodity Futures Trading Commission (2015), Joint Staff Report (2015), and New York Attorney General's Office (2014). Specific speed bump proposals include Cboe EDGA (2019), ICE Futures (2019), London Metals Exchange (2019), Chicago Stock Exchange (2016), and Investors' Exchange (2015).

[5]To our knowledge, ours is the first study to use exchange message data. All of the studies referenced above use limit order book data (either exchange direct feeds or consolidated datasets), in some cases with additional information such as participant identifiers.

Service and Markets Act) all message activity for all stocks in the FTSE 350 index for a 9 week period in Fall 2015.[6] The messages are time-stamped with accuracy to the microsecond (one-millionth of a second), and as we will describe in detail, the timestamps are applied at the right location of the exchange's computer system for measuring speed races (the "outer wall"). Using this data, we can directly measure the quantity of races, provide statistics on how long races take, how many participants there are, the diversity and concentration of winners/losers, etc. And, by comparing the price in the race to the prevailing market price a short time later, we can measure the economic stakes, i.e., how much was it worth to win.

Our main results are as follows:

- Races are frequent. The average FTSE 100 symbol has 537 latency-arbitrage races per day. That is about one race per minute per symbol.

- Races are fast. In the modal race, the winner beats the first loser by just 5-10 microseconds, or 0.000005 to 0.000010 seconds. In fact, due to small amounts of randomness in the exchange's computer systems, about 4% of the time the winner's message actually arrives to the exchange slightly later than the first loser's message, but nevertheless gets processed first.

- A large proportion of daily trading volume is in races. For the FTSE 100 index, about 22% of daily trading volume is in races.

- Races are worth small amounts per race. The average race is worth a bit more than half a tick, which on average comes to about 2GBP. Even at the 90th percentile of races, the races are worth just 3 ticks and about 7GBP.

- Race participation is concentrated. The top 3 firms win about 55% of races, and also lose about 66% of races. For the top 6 firms, the figures are 82% and 87%.

- The fastest firms disproportionately take, the remainder of market participants disproportionately provide the liquidity that gets taken. 90% of races are won by an aggressive order, i.e., a snipe attempt as opposed to a cancel attempt. The top 6 firms together take about 80% of liquidity in races while providing about 42%. Market participants outside the top 6 firms take about 20% of liquidity in races while providing about 58%. Thus, on net, much race activity consists of firms in the top 6 taking liquidity from market participants outside of the top 6. This taking is especially concentrated in a subset of 4 of the top 6 firms who account for a disproportionate share of stale-quote sniping relative to liquidity provision.

- In aggregate, these small races add up to a meaningful proportion of price impact, an important concept in market microstructure. We augment the traditional bid-ask spread decomposition suggested by Glosten (1987), which is widely utilized in the microstructure literature (e.g.,

[6]The FTSE 350 is an index of the 350 highest capitalization stocks in the UK. It consists of the FTSE 100, which are the 100 largest stocks, and roughly analogous to other countries' large-cap stock indices (e.g., the S&P 500 index), and the FTSE 250, which are the next 250 largest, and roughly analogous to other countries' small-cap stock indices (e.g., the Russell 2000 index).

Glosten and Harris, 1988; Hasbrouck, 1991*a*,*b*; Hendershott, Jones and Menkveld, 2011), to separately incorporate price impact from latency-arbitrage races and non-race trading. Price impact from trading in races is about 31% of all price impact, and about 33% of the effective spread.

- In aggregate, these small races add up to meaningful harm to liquidity. We find that the "latency-arbitrage tax," defined as the ratio of daily race profits to daily trading volume, is 0.42 basis points if using total trading volume, and 0.53 basis points if using only trading volume that takes place outside of races. The average value-weighted effective spread paid in our data is just over 3 basis points. We show formally that the ratio of the non-race latency arbitrage tax to the effective spread is the implied reduction in the market's cost of liquidity if latency arbitrage were eliminated; that is, if liquidity providers did not have to bear the adverse selection costs associated with being sniped.[7] This implies that market designs that eliminate latency arbitrage would reduce investors' cost of liquidity by 17%. As a complementary analysis, we also show that the liquidity provider's realized spread in races is significantly negative (i.e., they lose money), whereas it is modestly positive in non-race liquidity provision. This pattern holds whether or not the liquidity provider is one of the fastest firms. This is direct evidence that latency-arbitrage races impose a tax on liquidity provision.

- In aggregate, these small races add up to a meaningful total "size of the prize" in the arms race. The relationship between daily latency-arbitrage profits and daily trading volume is robust, with an $R^2$ of about 0.81, suggesting the latency-arbitrage tax on trading volume is roughly constant in our data.[8] Using this relationship, we find that the annual sums at stake in latency arbitrage races in the UK are about GBP 60 million. Extrapolating globally, our estimates suggest that the annual sums at stake in latency-arbitrage races across global equity markets are on the order of $5 billion per year.

**Discussion of Magnitudes** Whether the numbers in our study seem big or small may depend on the vantage point from which they are viewed. As is often the case in regulatory settings, the detriment per transaction is quite small: the average race is for just half a tick, and a roughly 0.5 basis point tax on trading volume certainly does not sound alarming. But these small races and this seemingly small tax on trading add up to significant sums. A 17% reduction in the cost of liquidity is undeniably meaningful for large investors, and $5 billion per year is, as they say, real money—especially taking into account the fact that our results only include equities, and not other

[7] More precisely, the ratio we take is latency arbitrage profits in GBP divided by non-race effective spread paid in GBP, or, equivalently, the "latency arbitrage tax" on non-race trading in basis points, divided by the non-race average effective spread paid in basis points. Please see Section 5.5 for full details of this decomposition and the price impact decomposition.

[8] Daily volatility is also strongly related to daily latency-arbitrage profits, with an $R^2$ of about 0.66. Volume and volatility are highly correlated in our data, so adding volatility to the volume-only regression does not add much additional explanatory power. We present extrapolation results using both a volume-and-volatility model and a volume-only model, which is simpler; the results are very similar.

asset classes that trade on electronic limit order books such as futures, currencies, U.S. Treasuries, etc.

In this sense, our results are consistent with aspects of both the "myth" and "rigged" points of view. The latency arbitrage tax does seem small enough that ordinary households need not worry about it in the context of their retirement and savings decisions. Yet at the same time, flawed market design significantly increases the trading costs of large investors, and generates billions of dollars a year in profits for a small number of HFT firms and other parties in the speed race, who then have significant incentive to preserve the status quo.

**Organization of the Paper** The remainder of this paper is organized as follows. Section 2 describes the London Stock Exchange's systems architecture, to explain to the reader how our data are generated. Section 3 describes the message data in detail. Section 4 defines latency arbitrage and describes our methodology for detecting and measuring latency-arbitrage races. Section 5 presents the main results. Section 6 presents a number of sensitivity analyses. Section 7 extrapolates to an annual size of the prize for the UK and global equity markets. Section 8 concludes.

## 2 Inside a Modern Stock Exchange

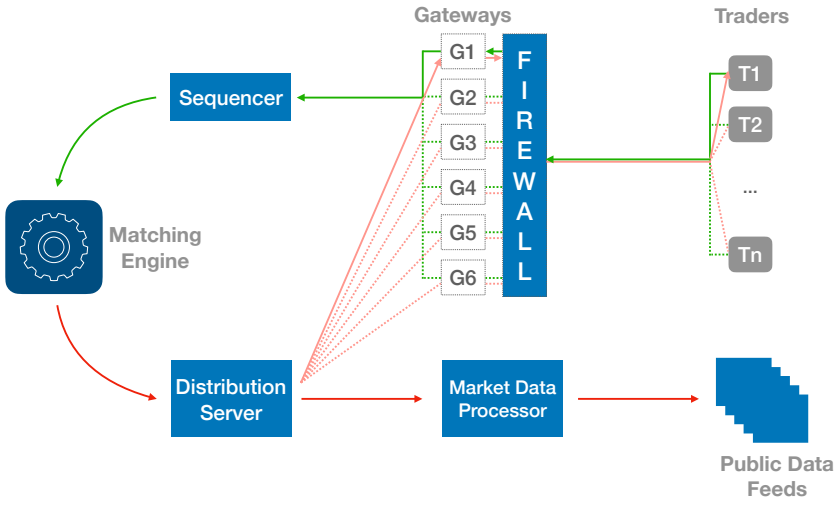
The continuous limit order book is at heart a simple protocol.[9] We guess that most undergraduate computer science students could code one up after a semester or two of training. Yet, modern electronic exchanges are complex feats of engineering. The engineering challenge is not the market design per se, but rather to process large and time-varying quantities of messages with extremely low latency and essentially zero system downtime.

In this section we provide a stylized description of a modern electronic exchange. We do this both because it is a necessary input for understanding our data (described in detail in Section 3), and because we expect it will be useful per se to both academic researchers and regulators who seek a better understanding of the detailed plumbing of modern financial markets.

Exchange operators do not typically disclose the full engineering details of their infrastructure, but some of them publicly disclose many of the relevant aspects. Our description in this section is

[9]We assume most readers are already familiar with the basics of a limit order book market but here is a quick primer for readers who need a refresher. The basic building block is a limit order, which consists of a symbol, price, quantity and direction (e.g., buy 100 shares of XYZ at 12.34). Market participants interact with the exchange by sending and canceling limit orders, and various permutations thereof (e.g., immediate-or-cancel orders, which are limit orders combined with the instruction to either fill the order immediately or to instead cancel it). Trade occurs whenever the exchange receives a new order to buy at a price greater than or equal to one or more outstanding orders to sell, or a new order to sell at a price less than or equal to one or more outstanding orders to buy. If this happens, the new order executes at the price of the outstanding order or orders, executing up to the new order's quantity, with the rest remaining outstanding. For example, if there are outstanding orders to sell 100 at 12.34 and another 200 at 12.35, a limit order to buy 600 at 12.35 would buy 100 at 12.34, buy another 200 at 12.35, and then the remaining 300 at 12.35 would "post" to the order book as a new outstanding order to buy. If there are multiple outstanding orders the new order could execute against, ties are broken based first on price (i.e., the highest offer to buy or lowest offer to sell) and then based on time (i.e., which outstanding order has been outstanding for the most time). Market participants may send new limit orders, or cancel or modify outstanding limit orders, at any moment in time. The exchange processes all of these requests, called "messages", continuously, one-at-a-time in order of receipt.

Figure 2.1: **Exchange Schematic**




**Notes:** Please see the text of Section 2.1 for supporting details for this figure.

based primarily on public documents published by the London Stock Exchange as well as discussions we had with the LSE in the process of conducting this study. We also have utilized public documents from other exchange families (e.g. Deutsche Börse, CME) and knowledge acquired through discussions with industry participants.[10]

## 2.1 A Stylized Description

### 2.1.1 The Matching Engine and Overall System Architecture

The core of a modern exchange (see Figure 2.1 for a schematic), and likely what most people think of as the exchange itself, is the *matching engine*. As the name suggests, this is where orders are matched and trades generated. A bit more fully, one should think of the matching engine as the part of the exchange architecture that executes the limit order book protocol. For each symbol, it processes messages serially in order of receipt, and, for each message, both economically processes the message and disseminates relevant information about the outcome of the message. For example, if the message is a new limit order, the matching engine will determine whether it can execute ("match") the order against one or more outstanding orders, or whether it should add the order to the book. It will then disseminate information back to the participant about whether their order posted, executed, or both; to any counterparties if the order executed; and to the public market data feeds about the updated state of the order book.

However, the matching engine is far from the only component of an exchange. Indeed, market participants do not even interact with the matching engine directly, in either direction. Rather, market participants interact with the exchange via what are known as *gateways*. Participants send

[10] See London Stock Exchange Group (2015*a*,*b*,*c*,*d*,*e*), Deutsche Börse Group (2018) and NYSE Group (2018).

messages to gateways, which in turn pass them on to a *sequencer*, which then passes the message to the matching engine for processing. The matching engine then transmits information back to a *distribution server*, which in turn passes private messages back to participants via the gateways, and public information to the market as a whole via the *market data processor*.

Before we describe each of these components, it is worth briefly emphasizing the overall rationale for this system architecture. The matching engine must, given the limit order book market design, process all messages that relate to a given symbol serially, in order of receipt. This serial processing is therefore a potential computational bottleneck. For a stark example, if a million messages arrived at precisely the same moment for the same symbol, the matching engine would have to process these million messages one-at-a-time.[11] Therefore, it is critical for latency to take as much of the work as possible "off of the shoulders" of the matching engine, and instead put it on to other components of the system.

### 2.1.2 Gateways

Gateways are the part of the exchange that participants directly interact with, in both directions. *Inbound*, participants send messages to the gateways using, in LSE's case, one of either two languages, called *interfaces*. One interface is called FIX,[12] which can be used widely across lots of different exchanges but, because it is not customized to LSE's system, is not optimized for speed. The other interface is called Native, because it is the "native" language of the LSE system; it is therefore faster.[13] Gateways receive messages from participants, verify their integrity, and then send them onwards. The verification includes things like checking that the message is of a valid length, all the required fields are populated and have valid parameters, etc., in addition to checking whether the message would violate the participant's risk threshold at an exchange, trying to detect erroneous "fat finger" trades, etc. If a message is verified, it is then, roughly speaking, "translated" into the language of the matching engine, and passed on.

*Outbound*, that is on the way back from the matching engine, gateways send messages back to participants informing them of the status of their order. For instance, that an outstanding order

[11] Computational backlogs associated with such bursts of messages were thought to play a role in the U.S. Treasury Market Flash Crash of October 15, 2014. See Joint Staff Report (2015)

[12] FIX is an acronym for Financial Information eXchange Protocol. See https://www.fixtrading.org/what-is-fix/.

[13] Incoming messages are organized as a stream of information. For a FIX message, this stream is delineated using field tags, <tag>=<value>. As an example, a new FIX limit order to buy 234 shares of Vodafone stock (which has instrument ID 133215) for £4.56 per share, submitted by traderID 789, with ClientOrderID 9452, Account 616, and Clearer 3113, would look like this: 8=FIX50SP2|9=156|35=D|49=789|56=FGW|34=10012|11=9452|48=VOD|22=8|40=2|54=1|38=234 |1138=234 |44=4.56|581=1|528=P|60=20150817-12:01:01.100|10=999|. A native message in binary format is not delimited and is sent as a string of binary bytes. The binary format protocol stipulates the order, and the starting and ending bytes of each parameter. There are no delimiters, as the length of each byte is used to delineate fields. The following is a stylized example which details the parameters the byte represents in sequence, so we do not reproduce the ones and zeroes. We have also included field delimiters ("|") to make it easier to interpret: 2|627|D|9452|789|616|3113|133215|0|0|2|0|20150817-13:01:01.100 |1|234|234|4.56|2|0|0|0|0|0|0|. The lack of delimiters makes the message shorter and quicker for the gateway to translate. Even the use of InstrumentID 133215 rather than VOD for Vodafone will be quicker for the exchange to read than converting the text. See London Stock Exchange Group (2015*c*,*d*)

was executed, or a new order posted to the book, or a cancel request failed. Additionally, if on the way in the gateway failed to verify a message, then the gateway will send an outbound message notifying the participant of the failure.

Notice, from a systems design perspective, how the gateway takes work off of the matching engine, and that much of the gateway function can be parallelized.[14] Most importantly, the gateway offloads from the matching engine the work of verifying the integrity of messages, of doing risk-checks, and of translating the message from the participant interface language into a language optimized for the matching engine.

#### 2.1.3 Sequencer

As emphasized above, it is valuable from a systems perspective to parallelize the gateway function, whereas the matching engine function intrinsically has to be serial (per symbol). The sequencer is essentially the bridge between the two. Its job is to receive input from all the gateways, and then, for each symbol, to pass on a single sequence of messages to the matching engine. From a systems perspective, this enables the matching engine to have to listen to only one input (per symbol) rather than many.

The details of the sequencer vary across exchanges. On the LSE, as well as many other exchanges including the New York Stock Exchange, the sequencer obtains messages from the gateways on a perpetual "round robin" basis, first obtaining a message from gateway 1 and then passing it to the matching engine, then obtaining a message from gateway 2, etc.[15] This means that it is possible that one message, say A, reaches its gateway before some other message, say B, reaches its gateway, yet B gets to the matching engine before A does. This will manifest in our data.

#### 2.1.4 Distribution Server

The matching engine, upon processing each order, sends output to the distribution server. The distribution server's job is then to further process the output for sending on (i) private messages to participants affected by the outcome, via the gateway; and (ii) public updates to subscribers to market data feeds (the Market Data Feed Server in our diagram). The public market data feeds typically contain information about all trades as well as all updates to the state of the limit order book.

Crucially for our study, not all information that is conveyed back in private messages to participants makes it to publicly available market data feeds. In particular, "too late to cancel" messages and "expired" (failed) immediate-or-cancel messages are both sent on to the relevant participants who either failed to cancel or failed to execute an immediate-or-cancel, but do not get sent on to public market data feeds because they do not affect the state of the order book. Similarly, such messages do not make it into academic data sets such as TAQ. Implicitly, these messages are viewed as "error messages", relevant to the participant but not relevant to market observers.

[14] In practice, this parallelization is achieved by assigning different participants to different gateways.
[15] See NYSE Group (2018).

# 3 Description of Data

As emphasized, the novel aspect of our data is that it includes all messages sent by participants to the exchange and by the exchange back to participants. Importantly, this includes messages that inform a participant that their request to trade or their request to cancel was not successful—such messages would not leave any empirical trace in traditional limit order book data. Also fundamental to our empirical procedure is the accuracy and location of the timestamps, which, as we will describe in detail below, are applied at the "outer wall" of the exchange's network and therefore represent the exact time at which a market participant's message reached the exchange. This timestamp location is ideal for measuring races, even more so than matching engine timestamps, as it represents the point at which messages are no longer under the control of market participants.[16]

We obtained these message data from the London Stock Exchange, following a request by the FCA to the LSE under Section 165 of the Financial Services and Markets Act. Our data cover the 44 trading days from Aug 17 – Oct 16 2015, for all stocks in the FTSE 350 index. We drop one day (Sept 7th) which had a small amount of corrupted data. This leaves us with 43 trading days and about 15,000 symbol-day pairs. In total, our data comprise roughly 2.2 billion messages, or about 150,000 messages per symbol-day.

## 3.1 Where and How Messages are Recorded and Timestamped

As described in Section 2, participants send messages to the exchange, and receive messages from the exchange, via gateways. Between the participants' own systems and the exchange's system is a firewall, through which all messages pass, in both directions. Our data are recorded and timestamped on the external side of this firewall using an optical TAP (traffic analysis point); please refer to Figure 3.1. This is the ideal timestamping location for measuring race activity because it records the time at which the participant's message reaches the "outer wall" of the exchange's system. Participant speed investments affect the speed with which their messages reach this outer wall, but once a message reaches the outer wall it is out of the participant's hands and in the exchange's hands. Therefore, the outer wall is the right way to think about what is the "finish line" in a race.

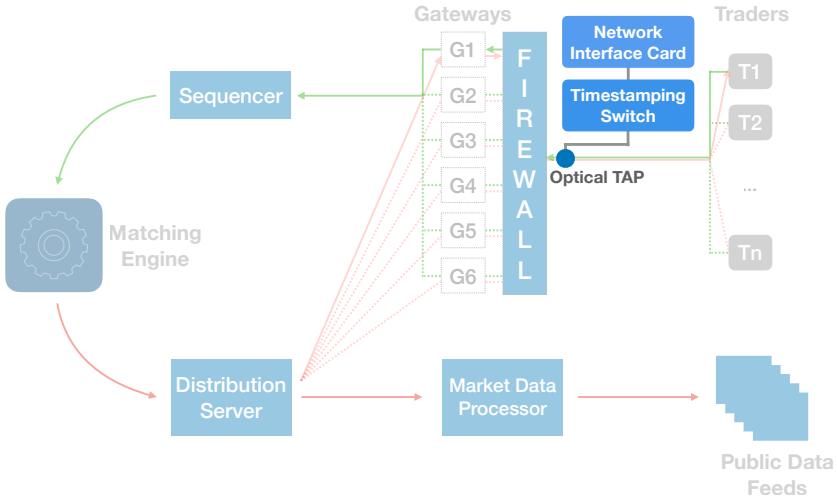
Messages are timestamped to 100 nanosecond (0.1 microsecond) precision, at this point of capture, by a hardware clock. Importantly, all messages are timestamped by a single clock. Therefore, while the clock may drift slightly over the course of the trading day, the relative timestamps of different messages in a race can be compared with extreme accuracy.

## 3.2 Contents of Messages

Any action by a market participant generates at least two messages: one on the way into the exchange, and one or more on the way out of the exchange. For example, a new limit order that

[16] We emphasize though that our methodology could be replicated in other contexts using matching engine timestamps, so long as the researcher had the full set of messages including failed cancels and failed IOCs. We think of the full message activity as a "must have" for the method and the precise location of the timestamps as more of a "nice to have."

Figure 3.1: **Exchange Schematic: Where the Message Data are Captured and Timestamped**




**Notes:** Please see the text of Sections 2.1 and 3.1 for supporting details for this figure.

both trades against a resting order and posts the remainder to the book will have a single inbound message with the new order, an outbound message to the user whose order was passively executed, and an outbound message to the user who sent the new limit order reporting both the quantity/price traded and the quantity/price that remains and is posted to the book. In this section we describe the contents of such inbound and outbound messages in detail.

### 3.2.1 Inbound Messages

Each inbound message contains the following kinds of information:[17]

**Identifiers.** These fields contain the symbol and date the message is associated with; the UserID of the participant who submitted the message; and a participant-supplied ID for the message. Additionally, if the message is a cancel or modification of an existing order, then the message often contains the matching-engine-supplied OrderID for the existing order (though the user is free to use just the participant-supplied ID they used previously for the order they are canceling).

**Timestamp.** As described above, each message has a timestamp down to 100 nanosecond granularity. For both inbound messages and outbound messages, the timestamp is applied at the optical capture point on the external side of the exchange firewall.

[17]There are some slight differences in how the information described below is organized in Native vs. FIX format messages (see Section 2 for more on Native vs. FIX). Since latency-sensitive participants essentially exclusively use Native format messages, our description focuses on Native and we do not note the small differences.

**Message Type Information.** Each message indicates what type of message it is, economically: for instance, a new limit order, a cancel, a cancel-replace, or an immediate-or-cancel order. This information is conveyed in a set of fields: a MessageType, which indicates whether it is a new order or a cancel or modification of an existing order; an OrderType, if it is a new order, which is typically set to indicate that it is a limit order, but could also be a market order, stop order, stop limit order, pegged order, etc.; and a Time in Force parameter, which indicates whether, for instance, a limit order is outstanding for the full day or whether it is immediate-or-cancel or fill-or-kill.

**Price/Quantity/Side Information.** Last, if a message is a new order or a modification of an existing order, it will of course indicate the price, quantity, and direction (buy/sell).

### 3.2.2 Outbound Messages

Each outbound message contains the following kinds of information:

**Identifiers.** These fields typically contain all of the same information as the inbound message, with the addition, for new orders, of a matching-engine-supplied OrderID. That is, for new orders, on the way in they just have the participant-supplied ID, but on the way out they contain both the participant-supplied ID and the matching-engine-supplied ID.[18]

**Timestamp.** As described above, both inbound messages and outbound messages are timestamped with 100 nanosecond granularity at the optical capture point on the external side of the exchange firewall. Note that in principle, the sequence of timestamps at this external border of the exchange's system can differ slightly from the actual sequence messages are executed in by the matching engine. We account for this issue in our method for maintaining the order book for a given symbol throughout the day, as described below in Section 3.4. Please note that neither the inbound nor outbound timestamps applied at this optical capture point are sent to market participants.

**Message Outcome Information.** Outbound messages contain information on the outcome of the message, as determined by the matching engine.[19] This outcome information is conveyed, primarily, in three fields. The first, ExecType, reports on what activity the matching engine just executed: a post to the book, a trade execution, a cancel, a cancel/replace, or an order expiration (in the event of a failed immediate-or-cancel order, for example). The second, OrderStatus, indicates the current status of the order: the main status options are new, filled, partially filled, canceled, and expired. The last, MessageType, is where we see if a cancel message failed.[20]

[18] An exception is Cancel Reject messages, which do not contain either the symbol or the matching engine OrderID (the order no longer exists in the matching engine); we infer both the symbol and the OrderID from the participant-supplied ID.

[19] A small subset of messages have an outcome which is instead determined by the gateway, wherein the gateway rejects the message as having invalid parameters before it reaches the matching engine. This could be caused by a participant error, for instance.

[20] In this case, the MessageType field will indicate that the message is a cancel reject, whereas for most other messages the MessageType field just tells us that the message is an execution report (with an ExecType and an

**Trade Execution Reports.** In the event of a successful trade (conveyed in the ExecType field described above), the outbound message will contain the executed price, quantity, and side. Note that if an order matches with multiple counterparties or at multiple prices, there will be a separate outbound message for each such match.

**Price/Quantity/Side Status Information.** Any outbound message that relates to an order that has not yet been fully executed or canceled will also report the order's price, side, and remaining quantity.

Full details on all of these fields and additional ones can be found in the online data appendix.[21]

### 3.3 Event Classification

As described above, any action by any market participant is associated with one inbound message from that participant, one or more outbound messages back to that participant, and, if applicable, outbound messages to other participants whose orders were passively executed. An important piece of our code is to classify sets of such messages into what we call order book events—for instance, a "new order - executed in full" event, or a "resting order - passive execution" event.

In our code, we loop through each user and each order (using the information from both the participant-supplied IDs and the matching-engine supplied IDs) to classify each message according to what order book event it is a part of. We give a special designation to the first such message in each event—typically, the inbound message that initiates the event and utilize this message's timestamp for the purpose of race detection (described below). The only exception is if the first message in an event is a passive fill, in which case we use the outbound message timestamp to account for the fact that the inbound message associated with that fill could have reached the exchange a long time before the event. Table 3.1 gives the pattern of inbound and outbound message activity for the most important order book events.

### 3.4 Maintaining the Order Book

Observe that neither inbound nor outbound messages contain the state of the limit order book — i.e., the prices and quantities at the best bid and offer, and at other levels of the order book away from the best bid and offer. This is because conveying the state of the order book in each message, while convenient, would mean larger and hence slower messages. We thus have to build and maintain the state of the limit order book ourselves.[22]

We maintain the state of the limit order book, for each symbol-date, on *outbound* messages. That is, whenever there is an outbound message reporting that any event occurred that updates the

OrderStatus).

[21] Our codebase and a user guide will be made publicly available upon publication. Regulators and researchers interested in obtaining this codebase and user guide prior to publication should contact the authors.

[22] The familiar TAQ (trades-and-quotes) data contains information about the state of the order book. But, studies that have utilized direct-feed data from exchanges, such as Budish, Cramton and Shim (2015) and others, must build and maintain the order book themselves.

Table 3.1: **Classifying Inbound and Outbound Messages Into Events**

| Event Name | Inbound Message Type | Outbound Message Type |
|---|---|---|
| New order posted to book | New Order (Limit) | New Order Accepted |
| New order aggressively executed in full | New Order (Limit) | Full Fill (Aggressive) |
| | New Order (IOC) | Partial Fill (Aggressive) - multiple such orders that sum to the full quantity |
| New order aggressively executed in part | New Order (Limit) | Partial Fill (Aggressive) - one or more that sum to less than the full quantity |
| | New Order (IOC) | Order Expire - for IOCs, not Limits which will post the remainder |
| Order passively executed in part | - | Partial Fill (Passive) |
| Order passively executed in full | - | Full Fill (Passive) |
| Cancel accepted | Cancel | Cancel Accept |
| Failed cancel | Cancel | Cancel Reject |
| Failed IOC | New Order (IOC) | Order Expire |

**Notes:** Please see the text of Section 3.3 for a description of Event Classification. Please see Section 3.2 for a description of the contents of inbound and outbound messages.

state of the limit order book—a new limit order is added to the book, a resting order is passively filled, a resting order is canceled, etc.—we update the state of the order book. We do this on outbound messages rather than on inbounds because outbound messages report what the matching engine actually did. In the instances where multiple inbound messages arrive very close together in time, it is possible that the matching engine executes messages in a different sequence from what we would have expected given their inbound message timestamps (as we will see below in Figure 5.1, this occurs in about 4% of races; see Section 2.1 above for the systems architecture reason for this). Hence, we use the actual outbound executions to update the book.

We include limit orders submitted before the market open if they are not labeled good for auction, i.e., if they are valid to rest on the book after the opening auction. During this period the order book may cross, i.e., there may be offers to buy that exceed offers to sell. Any orders that trade in the opening auction we remove accordingly from the book (and similarly orders that are canceled prior to the open).

A technical issue that affects how we maintain the order book is that our data is subject to a small amount of packet loss.[23] Packet loss only affects the data recorded by the optical capture point (used for an LSE internal reporting solution) and not the messages sent to market participants. The LSE states that the occurrence of packet loss is extremely low. Packet loss can cause our calculated state of the limit order book to be different from the actual state. We take two steps to address this issue.

First, we build checks into our code that builds the order book that corrects the state of the

[23]Packet loss is the term for when a computer network recording device records strictly less than 100.0% of all activity.

order book in the event that we observe a matching engine event that contradicts our current state of the order book. For example, if we think the state of the book is bid 10 – ask 11, and then observe a trade where the aggressor buys at 12 (but not 11), we update the book to eliminate the asks at 11 which we know must no longer be present in the book; either the passive fills associated with trades at 11 were lost or cancels of the orders at 11 were lost.[24]

Second, we then produce audit statistics on both (i) the magnitude of the corrections, and (ii) the % of time that our order book state performs as expected. In a high-volume symbol (Vodafone) on a typical-volume day (09-23-2015), we are correct 99.95% of the time about whether a new limit order should trade against the book versus post to the book. On the highest-volume day of our sample (08-24-2015), which contained a mini-flash-crash and was noticeably an outlier on many measures relative to the other days, we are correct in this manner 99.82% of the time. Also reassuring, most of the time that we had to execute an order book correction, the correction concerned just a single level of the book, and involved a number of shares that was less than the mean depth at the top level of the book.

One other related note is that when we compute race statistics that rely on the order book, we always utilize the state of the order book as of the first message in the race. Thus, even if the burst of activity associated with races leads to a larger proportion of order book data issues, this should not affect our measures. Reassuringly, our measures of race profits based on depth in the order book at the start of the race are very similar to our measures of race profits based on the actual quantity traded and canceled in the race.

# 4 Defining and Measuring Latency Arbitrage Races

In this section we give the details for our method of measuring latency arbitrage activity using exchange message data. Section 4.1 provides a review of the relevant theory that motivates our approach. Section 4.2 describes the empirical method utilizing exchange message data. Section 4.3 provides supporting analysis regarding some of the specific time parameters we utilize.

We note that the method detailed in Section 4.2 is meant to be generalizable—that is, researchers or regulators who obtain message data from other exchanges should be able to follow the method described in 4.2 as a reasonably direct blueprint for their own analysis—whereas the timing parameter analysis in 4.3 is specific to the London Stock Exchange circa the time of our data.

## 4.1 Theory of Latency Arbitrage

Budish, Cramton and Shim (2015) develop a model of trading on a continuous limit order book market that both provides a theoretical definition of latency arbitrage and articulates the economics of the high-frequency trading speed race. We base our empirical strategy on the main insights of that model. Therefore, it will be useful to provide a brief summary of the main features of the BCS

[24] We do two kinds of state corrections. One uses matching engine actions that contradict our understanding of the state of the book. The second uses a field in outbound messages called PriceDifferential which, for limit orders that post to the book, indicates whether they are at the best bid or offer or if not how many levels away they are.

model of continuous trading and what the model implies for the questions we are trying to answer in this study.

Readers familiar with the BCS model may skip to Section 4.2 without loss.

### 4.1.1 Setup of the Model

BCS study a market where a single security, denoted $x$, trades on a continuous limit order book market.[25] There is a public signal, denoted $y$, about the fundamental value of this security which can be observed by all market participants. This public signal can be interpreted as a metaphor for information that comes from correlated financial instruments (e.g., a change in the FTSE 100 index, or activity in the option market for a given stock or vice versa), information that comes from trade in the same security but on another venue (e.g., another exchange or a dark-pool), or public news announcements.

There are two types of agents in the model. First, *investors* who have an exogenous demand to buy or sell $x$. They exogenously arrive to market and behave essentially mechanically, either buying or selling at the prevailing best offer or best bid immediately upon their arrival. In the BCS model investors have no private information, i.e., they can be interpreted as noise traders or liquidity traders.

Second, *trading firms* who have no intrinsic demand to either buy or sell $x$, but rather seek to buy $x$ at prices lower than $y$ and sell $x$ at prices higher than $y$. BCS first analyze the case of an exogenous number of trading firms, each with exactly the same speed technology—that is, in the event $y$ changes or there is some order book activity, all trading firms observe this information at exactly the same time. They then consider a model in which trading firms can endogenously choose to invest in speed technology, and those who invest are faster than those who do not.

Investors provide an incentive for trading firms to make markets, that is, to have orders resting on the book to buy at prices lower than $y$ and sell at prices higher than $y$. If an investor arrives, the trading firm who provided liquidity to the investor—i.e., the trading firm whose resting bid or ask the investor traded against—earns a profit equal to the difference between their quoted price and the fundamental value $y$. In equilibrium, the bid and ask will be symmetric around the fundamental value, and therefore a trading firm who provides liquidity to an investor earns half the bid-ask spread.

It is straightforward to enhance the model to also have informed traders of the sort modeled in Glosten and Milgrom (1985) and the large literature thereafter. For this extension please see Budish, Lee and Shim (2019), equation (3.1), and the surrounding text. In this extension, some innovations in the signal $y$ are publicly observed and some innovations are privately observed.

[25] Readers unfamiliar with the continuous limit order book should consult footnote 9. Other terms for this market design are continuous-time limit order book, centralized limit order book and electronic limit order book. These all mean the same thing.

### 4.1.2 Latency Arbitrage

If there is a publicly observed jump in the signal $y$, and this jump is more than half the bid-ask spread, there will be a race to "snipe" the resulting stale quotes. If the jump in $y$ is positive and exceeds the half-spread, the race will be to snipe the now-stale offers, and if the jump in $y$ is negative and exceeds the half-spread, the race will be to snipe the now-stale bids. If the provider of the stale quotes is fast they will also be part of the race, seeking to cancel their stale quotes before they are sniped. If the provider of the stale quotes is not fast then whether or not they attempt to cancel is irrelevant, either way they will get sniped.[26]

A conceptual insight of BCS is that even in the case where all trading firms have *exactly* the same technology, and *exactly* the same information, such public information creates arbitrage rents—because of the serial processing nature of the continuous limit order book. Even if multiple firms respond to new public information at *exactly* the same time, one of them earns a rent. These rents then induce a never-ending speed race: if any firm is even a tiny bit faster than the others in the race, they win. In practice, this never-ending speed race means that different firms may respond at different speeds to different kinds of public signals.

BCS thus define *latency arbitrage* as arbitrages in races to respond to public information, as opposed to the rents from private information that are at the heart of classic models in market microstructure, such as Kyle (1985) and Glosten and Milgrom (1985). In the simple generalization of BCS's model referenced above, which also includes informed traders, both latency arbitrage from public information and traditional adverse selection arising from private information play a role in equilibrium. Both are costs of liquidity provision that in equilibrium affect the bid-ask spread and market depth.

We emphasize that while in a theoretical model it is possible to draw a sharp line between races to respond to symmetric public information and trading based on asymmetric private information, and hence between latency arbitrage and traditional adverse selection, in practice the dividing line is not sharp. Our empirical method will attempt to account for this in two ways as described below in Section 4.2.4.

### 4.1.3 Key Theoretical Results from BCS

We briefly list the theoretical results from BCS that inform our study.

First, when there is a large-enough jump in a public signal, the activity should consist of fast trading firms attempting to snipe any stale quotes, and, if any of the stale quotes belong to fast trading firms, attempts to cancel the stale quotes. The total latency arbitrage prize includes both the profits in cases where a stale quote is sniped, and, in the case where a liquidity provider wins the

[26] While BCS focus on equilibria in which only fast firms provide liquidity (pgs. 1588-1590), there also exist, under slightly more precise modeling formalities introduced in Budish, Lee and Shim (2019), equilibria in which either slow firms provide all liquidity or in which liquidity is provided by a mixture of fast and slow firms. The bid-ask spread and latency-arbitrage prize are identical across all of these equilibria, and each fast firm gets the same total rent (equal to $\frac{1}{N}$ of the total sniping prize), whether they earn it via sniping or liquidity provision. The equilibria in which both slow and fast firms provide liquidity seems most empirically relevant given our results in Section 5. For additional discussion of theoretical details please see Appendix B.1.

race with a successful cancel, the value of the avoided loss. The reason is that this loss avoidance profit is the way that a fast trader who provides liquidity is compensated for the opportunity cost of not instead being a sniper. As we will see empirically in Section 5, however, loss avoidance is relatively rare; about 90% of races are won by snipers.

Second, the size of the latency arbitrage prize for a particular security depends on the probability of and size-distribution of jumps in y, and the bid-ask spread and market depth which themselves depend on the level of investor demand for the security. Hence, both the volume of trade and the volatility of the security are closely related to the size of the latency-arbitrage prize.

Third, latency arbitrage increases the cost of liquidity provision. Liquidity providers choose their equilibrium price and quantity of liquidity endogenously, and this choice will factor in the cost of latency arbitrage, just like it factors in the cost of traditional adverse selection. This holds whether the liquidity provider is fast or slow — fast trading firms are sometimes able to successfully cancel whereas slow firms never are, but these successful cancels are compensation for fast firms' opportunity cost of not instead trying to snipe. In equilibrium, the latency arbitrage prize ultimately comes out of the pockets of investors via a higher-than-otherwise cost of liquidity.

Finally, in the version of the model with endogenous investment in speed, the latency arbitrage prize is dissipated by such investments. These investments could take the form of communications links, hardware, human capital, etc. In the model, there is an equivalence among (i) the latency arbitrage prize; (ii) socially wasteful investment in speed; and (iii) the cost to investors in the form of higher cost of liquidity.

### 4.2 Method for Measuring Latency Arbitrage Using Exchange Message Data

The theory described above suggests that the empirical signature of a BCS-style latency arbitrage race, as distinct from Glosten-Milgrom-style informed trading, is that:

1. Multiple market participants act on the same security, side, and price level or levels ...
2. ... at least some of whom are aggressing (i.e., sniping stale quotes), and potentially one or more of whom are canceling (i.e., canceling stale quotes) ...
3. ... some succeed, some fail ...
4. ... all at the "same time."

For each of these 4 characteristics we provide a baseline definition and alternatives.

Items #1-#3 are each relatively straightforward to define. We structure the analysis so that our baseline is likely to be inclusive of all races and the alternatives filter down to more-conservative subsets of races.

Item #4 is conceptually more difficult. We structure the analysis so that the baseline method is conservative and then consider a wide range of sensitivity analyses.

Note that throughout, when we describe either actions or timestamps, we refer to the *inbound* messages and timestamps, enhanced with the event classification information described above in

Section 3 using subsequent outbound messages. For example, if we refer to a failed IOC, we are referring to the inbound IOC message and its timestamp, having inferred from subsequent outbound messages that the IOC failed to execute.

#### 4.2.1 Characteristic #1: *Multiple market participants act on the same security, side, and price level or levels*

**Baseline.** The "same security, side, and price level or levels" aspect is straightforward. Every limit order message (including IOC's, etc.) includes the symbol, price, and side of the order. We interpret a limit or IOC order to buy at $p$ as relevant to any race at price $p$ or lower, and similarly a limit or IOC order to sell at $p$ as relevant to any race at price $p$ or higher. Cancel messages can be linked to the price and side information of the order that the message is attempting to cancel. We count a cancel order of a quote at price $p$ as relevant to races at price $p$ only.[27]

Our baseline definition of "multiple market participants" is 2+ unique UserIDs. Note that a particular trading firm might use different UserIDs for different trading desks. Our approach treats distinct trading desks within the same firm as potentially distinct competitors in a latency-sensitive trading opportunity.

**Alternatives.** For alternatives, we also consider

- Larger minimum requirements for the number of participants in the race, such as 3+
- Requiring that the FirmIDs are unique, not just UserIDs.

#### 4.2.2 Characteristic #2: *at least some of whom are aggressing (i.e., HFTs sniping stale quotes), and potentially one or more of whom are canceling (i.e., HFTs canceling stale quotes)*

**Baseline.** For our baseline, we require that at least one of the multiple market participants is aggressing at $p$. Thus, a baseline race can consist of either 1+ aggressors and 1+ cancelers, or 2+ aggressors and 0 cancelers.

Defining a message as aggressing at $p$ is straightforward. For a race at an ask of $p$, a limit order or IOC is aggressive if it is an order to buy at $p$ or higher, and similarly for a race at a bid of $p$, a limit order of IOC is aggressive if it is an order to sell at $p$ or lower.

**Alternatives.** For alternatives we also consider

- Requiring 2+ aggressors. (That is, excluding races with 1 aggressor and 1+ canceler).
- Requiring that there are 1+ aggressors and 1+ cancelers. (That is, excluding races with 2+ aggressors and 0 cancelers).

[27] For example, if we observed an IOC to buy at 20 and a cancel of an ask at 21 at the same time, we would not want to count that as a race at 20. Whereas, if we observed an IOC to buy at 21 and a cancel of an ask at 20 at the same time, we potentially would want to count that as a race at 20.

- Requiring that there are 2+ aggressors and 1+ cancelers.

#### 4.2.3 Characteristic #3: *some succeed, some fail*

For our baseline, we require 1+ success and 1+ fail, defined as follows.

**Baseline: Fails.** A cancel attempt is a fail if the matching engine responds with a too-late-to-cancel error message. An immediate-or-cancel limit order is a fail if the matching engine responds with an "expired" message, indicating that the IOC order was canceled because it was unable to execute immediately. Note that an IOC order that trades any positive quantity will not count as a fail, even if the traded quantity is significantly less than the desired quantity.[28]

In our baseline, we count a limit order as a fail in a race at price $p$ if it was priced aggressively with respect to $p$ (i.e., is an order to buy at $\geq p$ or an order to sell at $\leq p$) but obtains zero quantity at $p$. That is, it either executes at a price strictly worse than $p$ (e.g., it buys at $> p$), or it posts to the book at $p$ or worse (e.g., instead of buying at $p$ it becomes the new bid at $p$). While most sniping attempts in our data are IOCs (over 90% in the baseline race analysis), in a race it can make sense to use limit orders instead of IOCs for two reasons. First, by using a limit order instead of an IOC, the participant posts any quantity he does not execute to the book, which in principle may yield advantageous queue position in the post-race order book. Second, at the LSE, there was a small (0.01 GBP per message) fee advantage to using plain-vanilla limit orders instead of IOC orders.[29] This difference means that, technically, IOCs are often dominated by "synthetic IOCs" created by submitting a plain-vanilla limit order followed by a cancellation request.[30]

That said, limit orders that obtain zero quantity at $p$ and instead post to the book may represent post-race liquidity provision reflecting the post-race value, as opposed to a failed attempt to snipe. For that reason, we also consider and will frequently emphasize the following alternative:

**Alternatives: Fails.**

- Not allowing non-IOC limit orders to count as fails. That is, only failed IOCs and failed cancel attempts count as fails.

[28] To be conservative, we do not allow for fill-or-kill orders to count as fails. FOK orders are rare (whereas IOCs are common) and do not make sense to use in a latency arbitrage race (whereas IOCs do make sense). For example, if there are 10,000 shares outstanding at a stale price, a sniper should attempt to take all 10,000, but should still want to take the rest even if some liquidity provider succeeds in canceling some small order (say for 1,000 shares, leaving 9,000 remaining) before the sniper's order is processed.

[29] At the time of our data, the LSE assessed an "Order management charge" of 0.01 GBP for non-persistent orders such as IOCs, whereas there was no order management charge for plain-vanilla limit orders (London Stock Exchange Group, 2015*f*). These order management charges are the same in the LSE's most recently posted fee schedule as of this writing.

[30] An exception is if the trader has triggered the "High usage surcharge" by having an order-to-trade ratio of at least 500:1; such traders must pay a fee of 0.05 GBP per message, so the synthetic IOC would be nearly twice as expensive as an IOC (London Stock Exchange Group, 2015*f*). However, our understanding is that triggering this surcharge is very rare.

**Baseline: Successes.** For our baseline, we consider an IOC or a limit order to be successful in a race at price $p$ if it is priced aggressively with respect to $p$ (i.e., is an order to buy at $\geq p$ or an order to sell at $\leq p$) and obtains positive quantity at a price $p$ or better (i.e., it buys positive quantity at a price $\leq p$ or sells positive quantity at a price of $\geq p$). We consider a cancel to be successful in a race at price $p$ if the order being canceled is at price $p$ and the cancel receives a cancel-accept response.

We note that this requirement is inclusive in two senses. First, it counts an IOC or a limit order as successful even if it trades only part of its desired quantity. However, the fact that an IOC or limit order trades only part of its desired quantity, in conjunction with the requirement that some *other message fails*—i.e., some other participant tried to cancel and received a too-late-to-cancel message, or some other participant tried to aggress at $p$ but executed zero quantity—will typically mean that the full quantity available at price level $p$ was contested and there were genuine winners and losers of the race. The possible exception is a successful IOC or limit for a subset of the available liquidity at price $p$, in conjunction with a failed cancel for part of that same subset of the available liquidity at price $p$. This case should be rare and we will attempt to filter it out with an alternative below.

Second, it counts a cancel as a success even if it cancels just a small quantity relative to the full quantity available at price level $p$. However, if the only success is a cancel, then since we also require a fail and 1+ aggressor, this implies that the full quantity available at price level $p$ was contested and there were genuine winners and losers of the race.

As alternatives, therefore, we also consider:

**Alternative: Successes.**

- Requiring that 100% of depth at the race price is cleared in the race. This can be satisfied either by observing a failed IOC at the race price $p$, a limit order at the race price $p$ that posts to the book at least in part, or by observing quantity traded plus quantity canceled of 100% of the displayed depth at the start of the race.
- Requiring that at least 50% of depth at the race price is cleared in the race.

#### 4.2.4 Characteristic #4: *all at the "same time."*

Of the 4 characteristics, this last one is conceptually the hardest. In a theory model there can be a precise meaning of "at the same time", but in practice and in the data no two things happen at *exactly* the same time, if time is measured precisely enough. Indeed, even if a regulatory authority or exchange *intends* for market participants to receive a piece of information at exactly the same time, and even if the market participants have *exactly* the same technology and choose *exactly* the same response, there will be small measured differences in when they receive the information, and when they respond to the information, if time is measured finely enough.[31]

[31] Try to blink your left eye and right eye at exactly the same time, measured to the nanosecond. You will fail! Computers are better at this sort of task than humans are, but even they are not perfect. See, e.g., MacKenzie

We consider two different approaches to this issue.

**Baseline Method: Information Horizon.** Our baseline approach, which we call the Information Horizon method, requires that the difference in inbound message timestamps between the first and second participants in a race is small enough that we are essentially certain that the second participant is not reacting to the action of the first participant. Concretely, we measure the information horizon as:

$$\begin{aligned} \textit{Information Horizon} = \textit{Actual Observed Latency} : M1\ \textit{Inbound} \rightarrow M1\ \textit{Outbound} \\ + \textit{Minimum Observed Reaction Time} : M1\ \textit{Outbound} \rightarrow M2\ \textit{Inbound} \end{aligned}$$

where: M1 refers to the first message in a race; M2 refers to the second message in the race; Actual Observed Latency M1 Inbound → M1 Outbound refers to the actual measured time between M1's inbound message's timestamp and its outbound message's timestamp, and Minimum Observed Reaction Time M1 Outbound → M2 Inbound refers to the minimum time it takes a state-of-the-art high-frequency trader to respond to a matching engine update, as measured from the outbound message's time stamp to the response's inbound message time stamp.

Given this formula, if M2's inbound message has a timestamp that follows M1's inbound message by strictly less than the information horizon, then the sender of M2 logically cannot be responding to information about the outcome of M1. Whereas, if M2's inbound message has a timestamp that follows M1 by more than the information horizon, it is logically possible that M2 is a response to M1. In this method, such a response would not be interpreted as the same time.

In our data we compute the Minimum Observed Reaction Time as 29 microseconds,[32] and the median Actual Observed Latency is about 150 microseconds (90th percentile: about 300 microseconds). We provide further details in Section 4.3. We also decided, in consultation with FCA supervisors, to place an upper bound on the information horizon of 500 microseconds. That is, if the sum of the observed matching engine latency and the minimum observed reaction time exceeds 500 microseconds, we use 500 microseconds as the race horizon instead. The reason for this upper bound is that our assumption that M1 and M2 are responses to the same (or essentially same) information set becomes strained if the observed matching engine latency is sufficiently long, because even though the sender of M2 would not be able to see M1, the sender of M2 might have seen new data from other symbols or from other exchanges. We would expect all of these parameters to be potentially different for different exchanges or different periods in time.

**Alternative Method: Sensitivity Analysis.** Our second approach to defining what it means for multiple participants to act at the "same time" is more agnostic. For a range of choices of $T$,

(2019).

[32] This 29 microseconds reflects a combination of the minimum time it takes an HFT to react to a privately-received update from an outbound message, plus the difference in data speed between a private message sent to a particular market participant (M1 outbound) and data obtained from the LSE's proprietary data feed, which is different from our message data. In fact, our analysis suggests that the 29 microseconds is comprised of about 17 microseconds from the first component and about 12 microseconds from the second component, as we will describe in Section 4.3.

we define "same time" as no further apart than $T$. Clearly, if we choose $T$ to be the finest amount of time observable in our data (100 nanoseconds) there will be essentially no races, whereas if we choose $T$ to be too long the results will be meaningless. We will present these results for $T$ ranging from 50 microseconds to 3 milliseconds. What $T$'s would be of interest we would expect to evolve over time as technology evolves.

#### 4.2.5 A Note on Code Structure and Multi-Level Races

Depending on the size of the jump in value (i.e., $y$ in the theory model), a latency-arbitrage race could occur on one level of the book or on multiple levels. We structure our code so that it identifies races that satisfy the four characteristics described above at one price level at a time. That is, if $p$ and $p'$ are separate price levels in a multi-level race, our code will detect two single-level races, one at $p$, starting at say time $t$, and one at $p'$ starting at say time $t'$.

A related code structure issue to mention is that once we observe a race at a price level of $p$ starting at time $t$, we do not look for other races at $p$ until at least either the information horizon or $T$ amount of time has passed. That is, we do not allow for "overlapping" races at a single price level.

### 4.3 Computing the Information Horizon

As described in Section 4.2.4, there are three elements of our Information Horizon calculation:

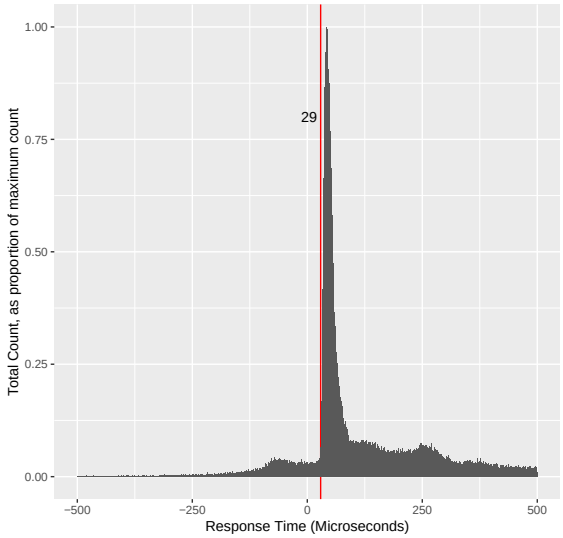
1. Actual Observed Latency: M1 Inbound → M1 Outbound
2. Minimum Observed Reaction Time: M1 Outbound → M2 Inbound
3. Upper bound on maximum possible information horizon

We can compute the Actual Observed Latency: M1 Inbound → M1 Outbound directly in our data, for each inbound message. This is obtained by taking the difference between the inbound message's timestamp and its outbound message's timestamp. The median response time is 157 microseconds, and there is considerable variation: the 10th percentile is 108 microseconds and the 90th percentile is 303 microseconds.[33]

To compute the Minimum Observed Reaction Time: M1 Outbound → M2 Inbound, we perform the following analyses. First, we look at instances of the specific sequence of events where M1 outbound is a new limit order that adds liquidity at some price level, and M2 inbound is an aggressive order (i.e., take) from a different UserID at the same price level. In this sequence of events, M2 may be responding to the new liquidity at the price level by taking it. Clearly, sometimes this sequence of events will happen by chance, but sometimes this sequence of events will happen because M2 is responding to M1. Figure 4.1 reports the distribution of the difference in time between these two events.

[33]These figures are based on the M1 Inbound → M1 Outbound response time over all messages that are the first message in a race.

Figure 4.1: **Distribution of Time between M1 Outbound New Limit Order → M2 Inbound Takes Liquidity**




**Notes:** Over all regular-hour messages from four high-volume symbols, BP, GLEN, HSBA, VOD, we obtain all cases where some outbound message confirms a new order added to the book and subsequently gets filled at least in part. We then obtain the first outbound message that is an execution against this new order, obtain the inbound message associated with this outbound execution message, and compute the difference in the message timestamp between the first order's (M1) outbound message and the second order's (M2) inbound message. Note that this difference can be negative if M2's inbound is sent by the participant before M1's outbound is sent by the outbound gateway. The distribution depicted is a microsecond-binned histogram truncated at -500 microseconds and +500 microseconds. As described in the text, we compute the start of the spike (29 microseconds) by computing the mean and standard deviation of the distribution in the period -100 microseconds to 0 microseconds, and then finding the first microsecond after 0 that is at least 5 standard deviations above this pre-0 mean.

As can be seen, this distribution spikes upwards a bit to the right of 0. We interpret the beginning of this spike as the minimum amount of time it takes the fastest market participants to respond to such an M1 with such an M2, as measured from the outbound time stamp to the inbound time stamp. Note that it need not be the case that the market participant is responding literally to the outbound message sent to the participant who sent M1; rather, the market participant is likely responding to their own receipt of information about the state of the order book from the LSE's proprietary data feed, sent through the message server as depicted earlier in Figure 2.1. Using the simple statistical criterion of looking for the start of the spike by asking what is the first microsecond at which the density is more than 5 standard deviations above the distribution in the 100 microseconds leading up to time 0, we determine that the spike starts at 29 microseconds.

We also examined the case where M1 is a partial fill, and M2 is a successful cancel. In this case, the participant might be responding to their own privately-received message—so we might expect this to be faster than what we saw above for the Add-Take sequence. Here (see Appendix Figure A.1), the spike starts at around 17 microseconds. An interpretation is that the 17 microseconds is the minimum response time to a privately-observed outbound message, and the additional 12 microseconds is the minimum difference in latency between a private message sent to a particular

market participant and the LSE's broadly disseminated proprietary data feed.[34]

Last, the upper bound on the information horizon that we utilize, 500 microseconds, was determined in consultation with supervisors at the Financial Conduct Authority. This was based on the discussions they had with fast market participants on their reaction times, differences in the speeds of competing microwave connectivity providers, the variance in arrival times across long distances (such as Chicago to London), the geographical distance between the LSE's data center and other UK exchanges' data centers, and the judgment of supervisory experts to establish an amount of time short enough for our assumption that M2 is not reacting to M1 to be reasonable. This 500 microsecond truncation of the information horizon binds in just under 4% of cases.

# 5 Main Results

This section presents all of our main results under the baseline specification as described in Section 4. In the following section (Section 6) we will explore various alternative specifications and sensitivity analyses. Section 5.1 presents results on race frequency, duration, and trading volume. Section 5.2 presents results on race participation patterns. Section 5.3 presents results on profits per race. Section 5.4 presents results on aggregate profits and the "latency arbitrage tax." Section 5.5 presents two spread decompositions that explore what proportion of the cost of liquidity is the latency arbitrage component versus the traditional adverse selection component.

## 5.1 Frequency and Duration of Latency-Arbitrage Races

### Races Per Day

The average FTSE 100 symbol in our sample has 537 races per day. Over an 8.5 hour trading day, this corresponds to a race roughly once per minute per symbol. There are fewer races for FTSE 250 symbols: the average FTSE 250 symbol has 70 races, or roughly one per 7 minutes. Also, while all FTSE 100 symbols have daily race activity (the minimum is 76 races per day), the bottom quartile of FTSE 250 symbols have zero or hardly any race activity. See Table 5.1, Panel A.

Across all symbols in our data, there are on average about 71,000 races per day, of which 54,000 are FTSE 100 and 17,000 are FTSE 250. This total number of races per day ranges from a min of 48,000 to a max of 144,000. See Table 5.1, Panel B.

### Race Durations

The average race duration in our data, as measured by the time from the first success message to the first fail message, is 79 microseconds, or 0.000079 seconds. Table 5.2 and Figure 5.1 depict the distribution of race durations. The mode of the distribution is between 5-10 microseconds, and the median is 46 microseconds. There is then steady mass in the distribution up until about 150

[34] A similar difference between the speed with which private messages are received versus book updates from proprietary data feeds has been a recurring source of controversy at the Chicago Mercantile Exchange. See Patterson, Strasburg and Pleven (2013) and Osipovich (2018).

Table 5.1: **Races Per Day**

Panel A: Number of races per day across symbols

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 537.24 | 473.26 | 132 | 184 | 240 | 352 | 619 | 1,134 | 2,067 |
| FTSE 250 | 70.05 | 93.53 | 0 | 0 | 2 | 44 | 104 | 166 | 404 |
| Full Sample | 206.03 | 340.73 | 0 | 1 | 14 | 87 | 239 | 511 | 1,814 |

Panel B: Number of races per day across dates

| Description | Mean | sd | Min | Pct10 | Pct25 | Median | Pct75 | Pct90 | Max |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 54,261 | 15,660 | 35,174 | 40,490 | 44,036 | 51,361 | 60,632 | 70,588 | 117,370 |
| FTSE 250 | 17,232 | 3,856 | 11,536 | 13,444 | 14,800 | 16,125 | 19,404 | 23,326 | 26,613 |
| Full Sample | 71,493 | 19,223 | 48,175 | 54,264 | 58,698 | 64,516 | 79,429 | 93,914 | 143,752 |

**Notes:** Please see Section 4.2 for a detailed description of the baseline race-detection criteria and Section 3 for details of the message data including how we classify inbound messages and how we maintain the order book. This table reports the distribution of the number of races detected at the symbol level (Panel A) and at the date level (Panel B). The symbol level averages across all dates for each symbol. The date level sums across all symbols for each date.

microseconds, the 90th percentile is about 200 microseconds, and there is a tail up to our truncation point of 500 microseconds.

### Sometimes the "Wrong" Message Wins

Interestingly, in Figure 5.1, there is a small amount of mass to the left of zero; that is, the first fail message arrives before the first success message. Recall from Section 3.1 that our timestamps are obtained at the outer wall of the exchange's system. It is therefore possible, if two race messages arrive to different gateways at nearly the same time, that they reach the matching engine in a different order from the order at which they reached the exchange's outer perimeter. Thus, the "wrong" message wins the race about 4% of the time in our data.

We do not think the fact that the wrong message wins is necessarily that economically interesting; it is akin to one shopper choosing a slightly faster queue than another shopper at the supermarket. Rather, we think of the result as reinforcing just how fast races are: they are so fast that randomness in exchange gateway processing is sometimes the difference between winning and losing.[35]

[35] Please also see a recent essay of MacKenzie (2019) on various aspects of randomness in high-frequency trading races.

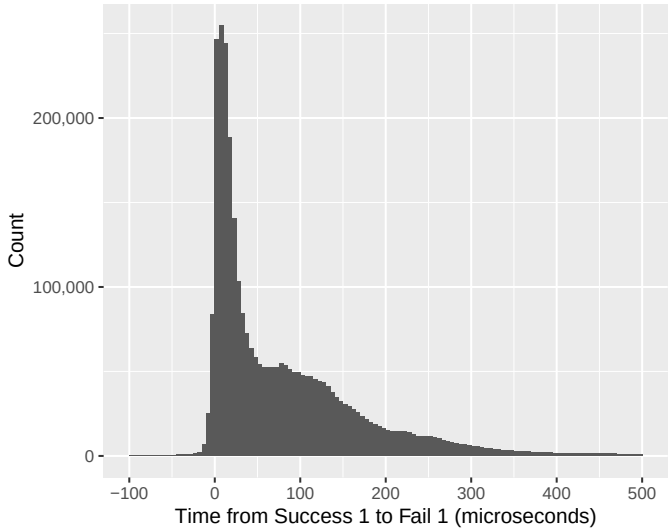
Table 5.2: **Race Duration**

Time from S1 to F1 (microseconds)

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 80.81 | 92.14 | -9.00 | 3.70 | 12.60 | 48.50 | 123.70 | 207.50 | 402.80 |
| FTSE 250 | 71.85 | 80.84 | -4.40 | 4.30 | 12.80 | 37.10 | 111.70 | 185.60 | 338.00 |
| Full Sample | 78.65 | 89.63 | -7.90 | 3.80 | 12.70 | 45.60 | 120.90 | 201.90 | 390.20 |

**Notes:** For each race detected by our baseline method (see Section 4.2 for detailed description) we compute the difference in message timestamps between the first inbound message in the race that is a success and the first inbound message in the race that is a fail (success and fail are defined in Section 4.2.3). Denote these messages S1 and F1, respectively. This table reports the distribution of F1's timestamp minus S1's timestamp in microseconds, that is, by how long did the first successful message in the race beat the first failed message.

Figure 5.1: **Duration of Races**




**Notes:** The figure plots the distribution of F1's timestamp minus S1's timestamp in microseconds, as defined in Table 5.2, for the full sample. The histogram has a bin size of 5 microseconds.

### Significant Trading Volume in Races

For the average FTSE 100 symbol, races take up a total of 0.043 seconds per day, or about 0.0001% of the trading day. This is based on the 537 races per day reported in Table 5.1 and the 81 microsecond race duration reported in Table 5.2 (537 * 0.000081 = 0.043 seconds).

During this tiny slice of the trading day, an average of 21% of FTSE 100 trades take place corresponding to 22% of FTSE 100 daily trading volume (value-weighted). Please see Table 5.3.

For the average FTSE 250 symbol, races take up about 0.00002% of the trading day. During this time 17% of trades take place constituting 17% of daily trading volume.

Table 5.3: **Volume and Trades in Races**

Panel A: Percentage of volume (value-weighted) in races across dates

| Description | Mean | sd | Min | Pct10 | Pct25 | Median | Pct75 | Pct90 | Max |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 22.15 | 1.90 | 17.84 | 20.09 | 21.15 | 22.02 | 23.11 | 24.85 | 26.08 |
| FTSE 250 | 16.90 | 1.78 | 11.58 | 14.73 | 15.71 | 17.07 | 18.19 | 19.21 | 20.13 |
| Full Sample | 21.46 | 1.75 | 17.63 | 19.70 | 20.50 | 21.41 | 22.53 | 24.02 | 25.02 |

Panel B: Percentage of number of trades in races across dates

| Description | Mean | sd | Min | Pct10 | Pct25 | Median | Pct75 | Pct90 | Max |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 20.69 | 1.59 | 16.91 | 18.62 | 19.83 | 20.80 | 21.58 | 22.93 | 23.51 |
| FTSE 250 | 16.96 | 1.50 | 13.29 | 15.24 | 16.01 | 17.01 | 18.07 | 18.91 | 19.31 |
| Full Sample | 19.70 | 1.42 | 16.07 | 18.04 | 18.94 | 19.65 | 20.68 | 21.73 | 22.22 |

**Notes:** For each symbol-date in our dataset, we obtain all outbound messages in regular-hours trading that are aggressive fills, i.e., that report a trade execution to a just-received new order that aggressed against a previously-received resting order. We then obtain the inbound message associated with each such outbound aggressive fill, and check whether the inbound is part of a race (see notes for Table 5.1). For Panel A, for each date, we then sum the quantity in GBP associated with all aggressive fills that are part of races, divided by the quantity in GBP associated with all aggressive fills, whether or not in race. We do this separately for the FTSE 100 (i.e., both the numerator and denominator sum across all symbols in the FTSE 100), the FTSE 250, and the full sample. For Panel B, for each date, we then sum the number of trades associated with all aggressive fills that are part of races, divided by the number of trades associated with all aggressive fills, whether or not in race.

## 5.2 Race Participation

### Number of Participants

Table 5.4, Panel A provides data on the number of participants in races. Since the information horizon varies across races depending on the matching engine's processing lag, to keep the measure consistent across races we report the distribution for varying amounts of time $T$ after the start of the race, ranging from 50 microseconds to 1 millisecond. Note that 50 microseconds is shorter than the information horizon for nearly all races and 1 millisecond is longer than the information horizon for all races (which is capped at 500 microseconds). Focusing on the 500 microseconds row, the average race has about 3.3 participants; the median has 3 participants; the 25th percentile has 2 participants; and there is a right tail with a 99th percentile of 9 participants and a max of 23 participants.

Comparing the 500 microseconds row to the 50 and 100 microseconds rows, we see that at shorter time horizons there are fewer participants. This is consistent with heterogeneity in speed, whether across firms or across different kinds of public signals. In the sensitivity analyses in Section 6, we will specifically consider using only races with at least a certain level of participation very quickly, and we will also consider less restrictive definitions of races that allow for participation over longer periods (up to a maximum of 3 milliseconds).

### Number of Takes and Cancels

Panels B and C of Table 5.4 provide the distribution of the number of take messages and cancel messages in races, respectively. Focusing initially on the 500 microseconds row, we see that the 3.27

Table 5.4: **Number of Participants and Messages in Races**

Panel A: Number of participants

| Description | Mean | sd | Min | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Participants within 50us | 1.77 | 0.86 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 5 | 12 |
| Participants within 100us | 2.08 | 0.97 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 5 | 13 |
| Participants within 200us | 2.56 | 1.13 | 1 | 1 | 2 | 2 | 2 | 3 | 4 | 6 | 16 |
| Participants within 500us | 3.27 | 1.56 | 2 | 2 | 2 | 2 | 3 | 4 | 5 | 9 | 23 |
| Participants within 1000us | 3.64 | 1.94 | 2 | 2 | 2 | 2 | 3 | 4 | 6 | 11 | 26 |

Panel B: Number of take messages

| Description | Mean | sd | Min | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Takes within 50us | 1.66 | 0.97 | 0 | 0 | 1 | 1 | 1 | 2 | 3 | 5 | 14 |
| Takes within 100us | 1.93 | 1.08 | 0 | 0 | 1 | 1 | 2 | 2 | 3 | 5 | 15 |
| Takes within 200us | 2.37 | 1.30 | 0 | 1 | 1 | 1 | 2 | 3 | 4 | 7 | 17 |
| Takes within 500us | 3.07 | 1.78 | 1 | 1 | 1 | 2 | 3 | 4 | 5 | 9 | 29 |
| Takes within 1000us | 3.45 | 2.19 | 1 | 1 | 1 | 2 | 3 | 4 | 6 | 11 | 40 |

Panel C: Number of cancel messages

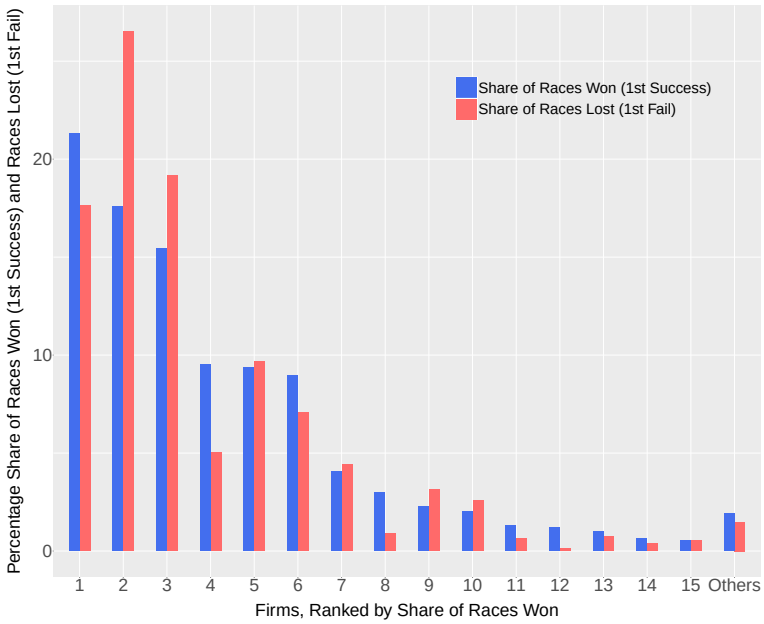
| Description | Mean | sd | Min | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cancels within 50us | 0.17 | 0.41 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 8 |
| Cancels within 100us | 0.22 | 0.47 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 8 |
| Cancels within 200us | 0.30 | 0.56 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 2 | 12 |
| Cancels within 500us | 0.40 | 0.70 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 14 |
| Cancels within 1000us | 0.44 | 0.78 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 21 |

**Notes:** For each race detected by our baseline method (see Section 4.2 for detailed description) we obtain the timestamp of the first inbound message and the price and side of the race. We then use the message data to count the number of messages within the next T microseconds, for different values of T as depicted in the table, that are race relevant, defined as either new orders that are aggressive at the race price and side (i.e., if the race is to buy at p, then new orders to buy at $\geqslant$ p, if the race is to sell at p, then new orders to sell at $\leqslant$ p), or cancels at exactly the race price (i.e., if the race is to buy at p, cancels of offers to sell at p, and vice versa). Panel A depicts the distribution of the number of participants with at least one race-relevant message. Panel B depicts the distribution of the number of race-relevant take messages and Panel C depicts the distribution of race-relevant cancel messages.

participants per race send an average of 3.47 messages of which 3.07 are takes and 0.40 are cancels. These figures tell us that in most races most of the activity is aggressive. This is consistent with equilibria of the BCS model in which the fastest traders primarily engage in sniping as opposed to liquidity provision, and substantial liquidity is provided by participants who are not the very fastest participants in the market (see Appendix B.1 for theoretical discussion of these equilibria). We will return to this pattern shortly.

Of these 3.07 take attempts, the large majority, 2.81, are immediate-or-cancel orders (IOCs) that are marketable at the race price, with the remainder, 0.25, being ordinary limit orders that are marketable at the race price. Please see Appendix Table A.4 for this and additional participation data. In Section 6 we will consider a sensitivity analysis that does not allow ordinary limit orders to count as losers of a race, since they may reflect an intention to provide liquidity at the new price rather than sniping liquidity at the old price. (Ordinary limit orders that execute at the race price will still count as winners of course, and indeed there can be a tiny economic advantage to sniping with an ordinary limit order relative to an IOC, as discussed in Section 4.2.3).

Figure 5.2: **Percentage of 1st Successful and 1st Failed Messages by Firm (FTSE 100 Races)**




**Notes:** For each race detected by our baseline method (see Section 4.2 for detailed description) we obtain the FirmID of the participant who sends the first success message and the first fail message (i.e., S1 and F1, respectively, in Table 5.2). We then compute, over all races for FTSE 100 symbols, for each FirmID that appears, the portion of races in which that FirmID is the first success message, and the portion of races in which that FirmID is the first fail message. The table sorts FirmIDs based on the proportion of races won. The "Others" bar sums all FirmIDs outside of the top 15.

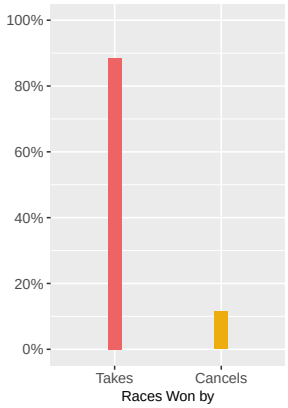
### Pattern of Winners and Losers

Figure 5.2 displays data on the pattern of winners and losers across races, focusing on races for symbols in the FTSE 100. The figure is sorted by firm based on the proportion of races in which they are the first successful message (S1). As can be seen, the top 3 firms are each either S1 or F1 (i.e., the first fail message) in over one-third of races, with firm 1 winning 21% of races while losing another 18% of races, firm 2 winning 18% of races while losing 27%, and firm 3 winning 15% of races while losing 19%. The next 3 firms then each win about another 9% of races each, and then there are another 4 firms that win between 2-4% of races each.

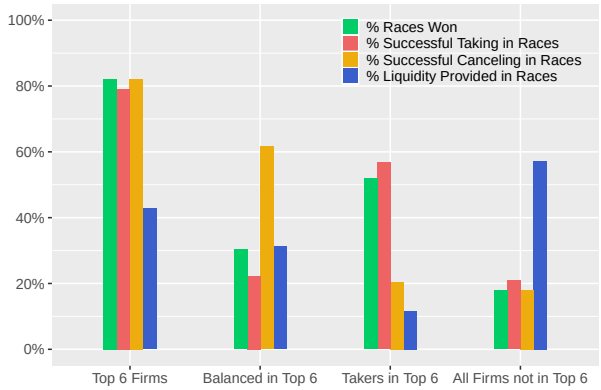
It is notable that there is clear concentration of winners, with the top 3 firms winning 54% of races, and the top 6 firms winning 82% of races. Yet, these same firms who win a lot of races also lose a lot of races. The top 3 winning firms lose 63% of races, and the top 6 lose 85%. These patterns are consistent with the BCS model in two ways. First, as the model suggests, fast trading firms "sometimes win, sometimes lose," and indeed in any particular race who wins may be a bit random. Second, as the model suggests, firms not at the cutting edge of speed should essentially never be competitive in a race. Put differently, these facts are consistent with the idea that there is an arms race for speed, and that, at least in UK equity markets circa 2015, there are a relatively

Figure 5.3: **Pattern of Takes, Cancels, and Liquidity Provision**

Panel A: Races Won by Takes vs. Cancels




Panel B: Analysis by Firm Group




**Notes:** Panel A: For each FTSE 100 race detected by our baseline method (see Section 4.2 for detailed description) we obtain whether the first successful message (i.e., S1) is a take or a cancel. Panel B: The first bar, % Races won, reports the data depicted in Figure 5.2 aggregated by firm group, with the firm groups as described in the text. The second bar, % Successful Taking in Races, is computed by taking all trading volume in all FTSE 100 races detected by our baseline method, and utilizing the FirmID associated with the aggressive order in each trade. For each bar, the numerator is the total quantity taken in races by firms in that group, in GBP, and the denominator is the total quantity traded across all races in GBP. The third bar, % Successful Canceling in Races, is computed by taking all successful cancels in FTSE 100 races detected by our baseline method, and utilizing the FirmID associated with the cancel attempt. For each bar, the numerator is the total quantity canceled in races by firms in that group, in GBP, and the denominator is the total quantity canceled across all races in GBP. The fourth bar, % Liquidity Provided in Races, is computed by taking all trading volume in all FTSE 100 races detected by our baseline method, and utilizing the FirmID associated with the passive side of each trade, i.e., the resting order that was taken by the aggressive order utilized in the % Successful Taking bar. For each bar, the numerator is the total quantity provided in races by firms in that group, in GBP, and the denominator is the total quantity traded across all races in GBP.

small number of firms competitive in this race.[36]

### Pattern of Takes, Cancels, and Liquidity Provision

Figure 5.3 Panel A shows that about 90% of races are won with a take (i.e., aggressive order or snipe attempt) with the remaining 10% won by a cancel. This makes sense in light of the data in Table 5.4 which showed that most of the message activity in races is take attempts as opposed to cancel attempts.

Figure 5.3 Panel B provides data on the pattern of successful takes, successful cancels, and liquidity provision across firms. The top 6 firms, as defined by the proportion of races won as shown in Figure 5.2, account for about 80% each of race wins, liquidity taken in races, and liquidity successfully canceled in races. In contrast, these 6 firms account for about 42% of all liquidity provided in races — that is, of all of the trading volume in races, 42% is volume where the resting order had been provided by one of the top 6 firms.

Within these top 6 firms there are two distinct patterns of race participation. 2 of the top 6 firms

[36] Around this time, a US high-frequency trading CEO described to one of the authors of this study that, in the US, there were 7 firms in what he called the "lead lap" of the speed race.

Table 5.5: **Liquidity Taker-Provider Matrix**

% of Race Volume by Taker-Provider Combination

| | | **Provider** | | |
|---|---|---|---|---|
| | | Takers in Top 6 | Balanced in Top 6 | Non-Top 6 |
| **Taker** | Takers in Top 6 | 5.7 | 17.2 | 34.3 |
| | Balanced in Top 6 | 2.5 | 6.4 | 13.3 |
| | Non-Top 6 | 3.2 | 7.4 | 10.1 |

**Notes:** For each race detected by our baseline method (see Section 4.2 for detailed description) we obtain all executed trades, and for each executed trade we obtain the FirmID of the participant who sent the take message that executed and the FirmID of the participant whose resting order was passively filled. The FirmIDs are classified into firm groups as described in the text. Each cell of the matrix reports the percentage of GBP trading volume associated with that particular combination of taker firm group and liquidity provider firm group.

together account for 28% of race wins, 22% of liquidity taken, 61% of successful cancels in races, and 31% of all liquidity provided in races. These data suggest that these 2 firms engage in meaningful quantities of both stale-quote sniping and liquidity provision; their ratio of liquidity taken in races to liquidity provided in races is about 2:3. The remaining 4 of the top 6 firms together account for 54% of race wins, 57% of liquidity taken, 21% of successful cancels, and just 11% of all liquidity provided in races. These data suggest that these 4 firms engage in significantly more stale-quote sniping than liquidity provision; their ratio of liquidity taken in races to liquidity provided in races is 5:1. We therefore denote these two groups of firms as "Balanced in Top 6" and "Takers in Top 6", respectively.[37]

Market participants outside of the top 6 firms account for about 20% each of race wins, liquidity taken in races, and liquidity successfully canceled in races. Where they stand out is that they account for 58% of all liquidity provided in races; that is, they provide nearly 3 times as much liquidity in races as they take.

Thus, on net, much race activity consists of firms in the top 6 taking liquidity from market participants outside of the top 6. This taking is especially concentrated in a subset of the fastest firms who account for a disproportionate share of stale-quote sniping relative to liquidity provision. The modal trade in our race data consists of a Taker in Top 6 firm taking from a market participant outside the top 6 (34.3% of all race volume). There is also significant race activity that consists of the fastest firms taking from each other. This volume is especially likely to consist of a Taker in Top 6 firm sniping a Balanced in Top 6 firm (17.2%). Please see Table 5.5 for a matrix of race trading volume organized by such taker-provider combinations.

[37] Previous studies that document heterogeneity across HFT firms with respect to their taking and liquidity provision behavior include Benos and Sagade (2016) and Baron et al. (2019). Benos and Sagade (2016) report that the most aggressive group of firms in their sample has an aggressiveness ratio of 82%, which means that 82% of their overall trading volume is aggressive, with the remaining 18% passive. Baron et al. (2019) report that the 90th percentile of firms in their sample has an aggressiveness ratio of 88%.

### Expected Number of Races By Chance

We can use the arrival rate of messages that could potentially be part of a race to compute the number of races we would expect to observe by chance if messages arrived randomly. We say that a message is potentially-race-relevant if the message is either a marketable limit order (including marketable IOCs) or is a cancel of a message at the best bid or offer. For each symbol-date, we compute the total number of such potentially-race-relevant messages per day to get an average arrival rate; to fix ideas, the average arrival rate for FTSE 100 symbols is a bit over 1 potentially-race-relevant message per second. We then use these arrival rates to compute the number of times per day we would expect to observe $N$ such messages within $T$ time on the same side of the order book. For the mean FTSE 100 symbol-date, the number of times per day we should expect to see $N = 2$ such messages on the same side of the order book within $T = 500$ microseconds, the upper bound of the information horizon, is just 3.55, in contrast with 537 races per symbol per day in our data. Increasing $T$ to 1 millisecond increases the expected number to 7.09. For the FTSE 250, the number of times per day we should expect to see $N = 2$ such messages within $T = 500$ microseconds is just 0.04, in contrast with 70 races per symbol per day in our data. The number of times we would expect to see $N = 3$ or more such messages arrive by chance is essentially zero. For the mean FTSE 100 symbol-date, the expected number of instances per day we would expect to see $N = 3$ or more messages within $T = 1$ millisecond by chance is 0.003, and for the mean FTSE 250 symbol-date, the figure is 0.000. (For full details, please see Appendix Table A.5).

Keep in mind as well that all of these figures are *upper bounds* on the number of $N$-participant races that would occur by chance, because occurrences of messages on the same side of the order book at the same time only constitute a race if our other race criteria are satisfied (in particular, at least one message must fail).

The bottom line is that the number of races we would observe by chance is de minimis.

## 5.3 Race Profits

### Profits Per-Race

Table 5.6 presents statistics on per-race profits. As in BCS, we compute profits as the signed difference between the price in the race and the midpoint in the near future, which has the interpretation of the mark-to-market value for the asset in the race.[38] Our main results use the midpoint 10 seconds out, and we will report figures for horizons ranging from 1 millisecond to 100 seconds shortly.[39]

[38]Note that while successful snipers must "cross the spread" in the trade that snipes a stale quote, they need not cross the spread in unwinding this position. This is both because trading firms that engage in sniping often also engage in liquidity provision, and because sniping opportunities are equally likely to be buys versus sells. Also note that it is appropriate to ignore trading fees in computing the size of the latency arbitrage prize, as long as exchanges' marginal costs of processing trades are zero, because trading fees assessed on latency-arbitrage trades simply extract some of the sniping prize.

[39]Since our data include firm identifiers, it would seem possible to use the actual trades made by participants to realize their profits rather than using mark-to-market profits at a range of time horizons. However, in addition to concerns about exploring specific firms' trading strategies in more detail than is necessary for this study, given that this is a privileged regulatory dataset obtained under a Section 165 request, there are two key limitations to this idea. First, we only have data from the London Stock Exchange, so do not observe when positions are closed by trades on other venues (see also Carrion (2013) who notes the same concern). Second, firms may not unwind positions after each race, but may instead manage inventory risk on a portfolio basis (see, for example, Korajczyk and Murphy (2019)).

Table 5.6: **Detail on Race Profits (Per-Share and Per-Race) Marked to Market at 10s**

Panel A: FTSE 100

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| Per-share profits (ticks) | 0.48 | 4.17 | -7.00 | -1.50 | -0.50 | 0.00 | 1.00 | 2.50 | 10.00 |
| Per-share profits (GBX) | 0.16 | 1.61 | -2.50 | -0.50 | -0.05 | 0.00 | 0.25 | 1.00 | 3.50 |
| Per-share profits (basis points) | 1.20 | 7.75 | -13.95 | -4.02 | -1.18 | 0.00 | 3.42 | 6.31 | 20.32 |
| Per-race profits displayed depth (GBP) | 1.95 | 17.87 | -22.99 | -3.29 | -0.42 | 0.00 | 2.37 | 7.99 | 45.50 |
| Per-race profits qty trade/cancel (GBP) | 1.84 | 17.07 | -20.74 | -3.06 | -0.40 | 0.00 | 2.23 | 7.46 | 41.92 |

Panel B: FTSE 250

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| Per-share profits (ticks) | 0.77 | 2.99 | -4.50 | -1.00 | -0.50 | 0.50 | 1.50 | 3.00 | 11.00 |
| Per-share profits (GBX) | 0.20 | 0.99 | -1.50 | -0.25 | -0.05 | 0.05 | 0.25 | 0.75 | 3.50 |
| Per-share profits (basis points) | 3.09 | 11.07 | -18.12 | -5.14 | -1.70 | 1.37 | 6.12 | 13.28 | 38.78 |
| Per-race profits displayed depth (GBP) | 1.55 | 9.63 | -9.13 | -1.52 | -0.20 | 0.09 | 1.67 | 5.25 | 27.68 |
| Per-race profits qty trade/cancel (GBP) | 1.48 | 9.34 | -8.48 | -1.40 | -0.19 | 0.09 | 1.55 | 4.94 | 26.40 |

Panel C: Full Sample

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| Per-share profits (ticks) | 0.55 | 3.92 | -6.50 | -1.50 | -0.50 | 0.50 | 1.00 | 3.00 | 10.00 |
| Per-share profits (GBX) | 0.17 | 1.48 | -2.00 | -0.50 | -0.05 | 0.01 | 0.25 | 1.00 | 3.50 |
| Per-share profits (basis points) | 1.66 | 8.71 | -15.00 | -4.26 | -1.29 | 0.50 | 3.89 | 7.98 | 27.02 |
| Per-race profits displayed depth (GBP) | 1.85 | 16.27 | -20.00 | -2.76 | -0.34 | 0.00 | 2.15 | 7.27 | 41.50 |
| Per-race profits qty trade/cancel (GBP) | 1.76 | 15.57 | -18.13 | -2.56 | -0.32 | 0.00 | 2.02 | 6.78 | 38.44 |

**Notes:** For each race detected by our baseline method (see Section 4.2 for detailed description) we obtain the race price and side, the quantity in the book at that price and side as of the last outbound message before the initial race message, and the quantity traded and canceled in the race. Per-share profits in ticks, pence (GBX), and basis points are computed by comparing the race price to the midpoint price 10 seconds after the first race message (i.e., as of the last outbound message before 10 seconds after the timestamp of the first race message). Per-race profits are computed by multiplying per-share profits in GBX, times 1/100 to convert to GBP, times either the quantity displayed or the quantity traded and canceled. Panel A shows the distribution for all races for FTSE 100 symbols, Panel B for FTSE 250 symbols, and Panel C for the full sample.

The average FTSE 100 race is worth about half a tick per share (0.48 ticks), or about 1.20 basis points. This comes to about 2 GBP per race, measured either using all of the displayed depth at the start of the race (1.95 GBP) or all of the quantity traded or canceled during the race (1.84 GBP). For the FTSE 250, the figures are 0.77 ticks, 3.09 basis points, and GBP 1.55 per race based on displayed depth, and GBP 1.48 per race based on quantity traded or canceled. For the full sample, the figures are 0.55 ticks, 1.66 basis points, GBP 1.85, and GBP 1.76.

There is of course significant variation in profitability across races. This reflects both that some races are more profitable ex ante than others, i.e., reflect larger jumps in public information, and that over a 10 second horizon other information can materialize, either positively or negatively, that affects realized race profits ex post. Across our full sample, a 90th percentile race is worth 3.00 ticks and 7.98 basis points; a 99th percentile race is worth 10 ticks and 27.02 basis points.

Table 5.7 presents statistics on average per-race profits for different mark-to-market time horizons. As can be seen, average per-race profits increase with the time horizon, eventually flattening out at around 10 seconds for the FTSE 100 and at around 60 seconds for the FTSE 250. Our finding

Table 5.7: **Average Race Profits (Per-Share and Per-Race) for Different Mark to Market Horizons**

Panel A: FTSE 100

| Description | 1ms | 10ms | 100ms | 1s | 10s | 30s | 60s | 100s |
|---|---|---|---|---|---|---|---|---|
| Mean per-share profits (ticks) | 0.08 | 0.24 | 0.31 | 0.39 | 0.48 | 0.49 | 0.50 | 0.51 |
| Mean per-share profits (GBX) | 0.05 | 0.09 | 0.11 | 0.14 | 0.16 | 0.16 | 0.16 | 0.16 |
| Mean per-share profits (basis points) | 0.31 | 0.68 | 0.83 | 1.01 | 1.20 | 1.23 | 1.24 | 1.25 |
| Mean per-race profits displayed depth (GBP) | 0.40 | 1.14 | 1.42 | 1.72 | 1.95 | 1.89 | 1.86 | 1.82 |
| Mean per-race profits qty trade/cancel (GBP) | 0.43 | 1.10 | 1.35 | 1.62 | 1.84 | 1.78 | 1.74 | 1.70 |

Panel B: FTSE 250

| Description | 1ms | 10ms | 100ms | 1s | 10s | 30s | 60s | 100s |
|---|---|---|---|---|---|---|---|---|
| Mean per-share profits (ticks) | -0.10 | 0.12 | 0.24 | 0.43 | 0.77 | 0.94 | 1.04 | 1.06 |
| Mean per-share profits (GBX) | -0.01 | 0.05 | 0.08 | 0.12 | 0.20 | 0.24 | 0.26 | 0.26 |
| Mean per-share profits (basis points) | -0.26 | 0.64 | 1.09 | 1.78 | 3.09 | 3.74 | 4.14 | 4.24 |
| Mean per-race profits displayed depth (GBP) | -0.09 | 0.41 | 0.65 | 0.97 | 1.55 | 1.79 | 1.92 | 1.93 |
| Mean per-race profits qty trade/cancel (GBP) | -0.06 | 0.41 | 0.64 | 0.93 | 1.48 | 1.71 | 1.84 | 1.85 |

Panel C: Full Sample

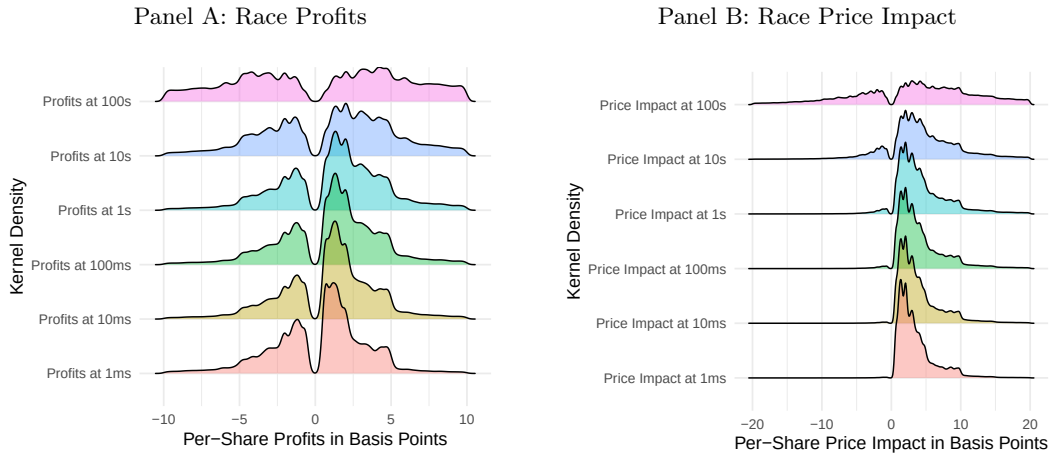
| Description | 1ms | 10ms | 100ms | 1s | 10s | 30s | 60s | 100s |
|---|---|---|---|---|---|---|---|---|
| Mean per-share profits (ticks) | 0.03 | 0.21 | 0.29 | 0.40 | 0.55 | 0.59 | 0.63 | 0.64 |
| Mean per-share profits (GBX) | 0.03 | 0.08 | 0.10 | 0.13 | 0.17 | 0.18 | 0.18 | 0.18 |
| Mean per-share profits (basis points) | 0.18 | 0.67 | 0.89 | 1.20 | 1.66 | 1.83 | 1.94 | 1.97 |
| Mean per-race profits displayed depth (GBP) | 0.28 | 0.96 | 1.24 | 1.54 | 1.85 | 1.86 | 1.88 | 1.84 |
| Mean per-race profits qty trade/cancel (GBP) | 0.31 | 0.94 | 1.18 | 1.45 | 1.76 | 1.76 | 1.77 | 1.74 |

**Notes:** For each race detected by our baseline method (see Section 4.2 for detailed description), and for each race profits measure described in Table 5.6, we re-compute the profits measure for different mark to market horizons, ranging from 1 millisecond to 100 seconds. That is, for each measure, we compute race profits by comparing the price and side in the race to the midpoint price T later, for T ranging from 1 millisecond to 100 seconds (Table 5.6 used T = 10 seconds). We then report the mean at each horizon.

that it takes non-zero time for race profits to materialize, and that with this time comes noise as well, is consistent with both discussions with practitioners as well as empirical evidence in Conrad and Wahal (2019) on what they call the "term structure of liquidity."

Figure 5.4 complements Table 5.7 by presenting the distribution of race profits and price impact at different time horizons. The difference between the two measures is that race profits are the difference between the price paid in the race and the midpoint price in the future, whereas price impact compares the midpoint at the time of the first inbound message in the race (i.e., just prior to its effect on the order book) to the midpoint price in the future (i.e., price impact does not charge the winner of the race the half bid-ask spread). Focus first on 1ms. At this relatively short time horizon, many races have profits that are either a small positive amount or small negative amount per share, whereas nearly all races have weakly positive price impact. This pattern reflects that, at the moment of a first success in a race, the mark-to-market profits of the winner are typically negative. For example, if the market is at bid 10 – ask 12, so the midpoint is 11, and there is positive public news triggering a race to buy at 12, then a successful sniper buys at 12 while the midpoint is still 11 (or, if the market becomes bid 10 – ask 13, the midpoint becomes 11.5)—for a

Figure 5.4: **Race Profits and Price Impact Distributions at Different Time Horizons**




**Notes:** For each race detected by our baseline method (see Section 4.2 for detailed description) we obtain per-share profits and price impact in basis points at different mark to market horizons ranging from 1 millisecond to 100 seconds. Profits at horizon $T$ are defined as the signed difference between the race price and the midpoint price at time $T$, while price impact at horizon $T$ is the signed difference between the midpoint price at the time of the first inbound message of the race (i.e., before that message affects the order book) and the midpoint price at time $T$. The figure plots the kernel density of the distribution of per-share profits (Panel A) and per-share price impact (Panel B), each in basis points, at different time horizons. To make the distributions readable, we drop all of the mass at exactly zero profits or price impact.

small mark-to-market loss. The figure shows that even by 1 millisecond, many races are profitable on a mark-to-market basis. As the figure progresses from 1 millisecond to 1 second, you can see visually that mass shifts to the right of the distribution (Table 5.7 reports the means), though there remains a meaningful mass of races with negative mark-to-market profits. Up to 1 second, nearly all races have weakly positive price impact.[40] By 100 seconds, as can be seen in both the race profits figure and the price impact figure, there is meaningful noise.

## 5.4 Aggregate Profits and the "Latency Arbitrage Tax"

Table 5.8 presents statistics on the total daily race profits in our sample. Panel A reports statistics at the symbol level, and Panel B reports statistics aggregated across all symbols in the FTSE 100, FTSE 250, and full sample. Note that all of these numbers are daily race profits in our data from the London Stock Exchange; we will extrapolate from these numbers to the full UK equities market and to global equities markets in Section 7.

Referring to Panel A, we see that the average symbol in the FTSE 100 has daily race profits of GBP 1,047, and the 99th percentile symbol has daily race profits of GBP 3,432. For the FTSE 250 the average and 99th percentile are GBP 108 and GBP 606, respectively.

Referring to Panel B, we see that the average day in our data set has race profits of GBP 105,734

[40] In principle, races with negative mark-to-market profits could either be spurious races that our method picks up but are not profitable, or they could be races based on public signals that multiple market participants expected to be profitable but turned out not to be profitable ex-post. Given the low likelihood of spurious races as discussed in Section 5.2 and reported in Appendix Table A.5, we suspect the latter interpretation is more quantitatively important.

Table 5.8: **Daily Profits in GBP**

Panel A: Daily Profits by Symbol

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 1,046.9 | 729.6 | 199.7 | 340.5 | 526.6 | 909.3 | 1,410.5 | 1,967.2 | 3,431.8 |
| FTSE 250 | 108.3 | 134.1 | -0.7 | 0.5 | 7.6 | 67.1 | 160.8 | 257.2 | 606.3 |
| Full Sample | 381.5 | 590.7 | -0.6 | 1.5 | 26.7 | 135.1 | 466.2 | 1,184.5 | 2,273.8 |

Panel B: Daily Profits by Date

| Description | Mean | sd | Min | Pct10 | Pct25 | Median | Pct75 | Pct90 | Max |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 105,734 | 32,852 | 62,980 | 78,777 | 87,038 | 93,074 | 117,979 | 153,712 | 223,187 |
| FTSE 250 | 26,643 | 8,592 | 14,667 | 19,501 | 21,376 | 23,100 | 30,392 | 40,100 | 49,066 |
| Full Sample | 132,378 | 40,266 | 82,391 | 99,363 | 108,706 | 116,636 | 147,814 | 183,227 | 272,253 |

**Notes:** For each race detected by our baseline method (see Section 4.2 for detailed description) we take per-race profits in GBP based on displayed depth with prices marked to market at 10 seconds (see notes for Table 5.6). We then compute daily profits for each symbol-date, by summing all races for that symbol on that date. In Panel A, for each symbol, we compute its average daily race profits, and report the distribution across symbols. In Panel B, for each date, we compute total daily race profits summed across all symbols, and report the distribution across dates. For each Panel, we perform the analysis separately for FTSE 100, FTSE 250, and full sample.

for the FTSE 100, GBP 26,643 for the FTSE 250, and GBP 132,378 for the full sample.

These aggregate profits numbers are difficult to interpret in isolation. A more interpretable measure is obtained by dividing race profits by daily trading volume, with both measures in GBP. We refer to this ratio as the "Latency Arbitrage Tax," since, following the theory in BCS, the prize in latency arbitrage races is like a tax on overall market liquidity. We consider two versions of this measure, the first based on all trading volume, and the second based on all non-race trading volume. The version based on all trading volume is both simpler to describe and more appropriate for out-of-sample extrapolation. However, the version based on all non-race trading volume more closely corresponds to the theory, which shows that latency arbitrage imposes a tax on non-race trading (both noise trading and non-race informed trading).

Table 5.9 reports that for the average symbol in the FTSE 100, the latency arbitrage tax is 0.492 basis points based on the all-volume measure, and 0.675 basis points based on the non-race-volume measure. For the average FTSE 250 symbol, the latency arbitrage tax is 0.562 based on the all-volume measure and 0.692 basis points based on the non-race-volume measure. Higher-volume symbols tend to have lower latency arbitrage taxes, so the overall value-weighted average daily latency arbitrage tax, for all symbols in the FTSE 350, is 0.419 basis points using the all-volume measure and 0.534 basis points using the non-race-volume measure.

An interpretation of the first figure is that for every GBP 1 billion that is transacted in the market overall, latency arbitrage adds GBP 41,900 to trading costs. An interpretation of the second figure is that for every GBP 1 billion that is transacted by participants not in latency-arbitrage races, latency arbitrage adds GBP 53,400 to trading costs.

Table 5.9: **Latency Arbitrage Tax**

Panel A: Distribution Across Symbols

Sub-Panel (i): Measure 1, Latency Arbitrage Tax based on All Trading Volume (basis points)

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 0.492 | 0.235 | 0.163 | 0.236 | 0.292 | 0.454 | 0.627 | 0.827 | 1.035 |
| FTSE 250 | 0.562 | 0.393 | -0.022 | 0.022 | 0.267 | 0.565 | 0.817 | 1.043 | 1.540 |
| Full Sample | 0.542 | 0.356 | -0.014 | 0.054 | 0.283 | 0.519 | 0.774 | 0.960 | 1.508 |

Sub-Panel (ii): Measure 2, Latency Arbitrage Tax based on Non-Race Trading Volume (basis points)

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 0.675 | 0.362 | 0.200 | 0.303 | 0.387 | 0.587 | 0.870 | 1.180 | 1.595 |
| FTSE 250 | 0.692 | 0.504 | -0.028 | 0.024 | 0.287 | 0.678 | 1.029 | 1.304 | 2.042 |
| Full Sample | 0.687 | 0.466 | -0.020 | 0.057 | 0.345 | 0.651 | 0.995 | 1.275 | 2.032 |

Panel B: Distribution Across Dates

Sub-Panel (i): Measure 1, Latency Arbitrage Tax based on All Trading Volume (basis points)

| Description | Mean | sd | Min | Pct10 | Pct25 | Median | Pct75 | Pct90 | Max |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 0.383 | 0.053 | 0.286 | 0.329 | 0.345 | 0.381 | 0.415 | 0.456 | 0.516 |
| FTSE 250 | 0.663 | 0.099 | 0.495 | 0.552 | 0.591 | 0.653 | 0.725 | 0.790 | 0.912 |
| Full Sample | 0.419 | 0.053 | 0.313 | 0.360 | 0.382 | 0.416 | 0.450 | 0.495 | 0.537 |

Sub-Panel (ii): Measure 2, Latency Arbitrage Tax based on Non-Race Trading Volume (basis points)

| Description | Mean | sd | Min | Pct10 | Pct25 | Median | Pct75 | Pct90 | Max |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 0.493 | 0.075 | 0.351 | 0.418 | 0.443 | 0.487 | 0.533 | 0.603 | 0.656 |
| FTSE 250 | 0.800 | 0.133 | 0.577 | 0.653 | 0.712 | 0.788 | 0.899 | 0.969 | 1.136 |
| Full Sample | 0.534 | 0.076 | 0.384 | 0.454 | 0.481 | 0.531 | 0.581 | 0.652 | 0.680 |

**Notes:** Panel A. For each symbol, we compute total race profits in GBP, summed over all dates in our sample, using per-race profits in GBP based on displayed depth with prices marked to market at 10 seconds (see notes for Table 5.6). We then compute total regular-hours trading volume in GBP, and total non-race regular-hours trading volume in GBP (see notes for Table 5.3). Panel A(i) reports the distribution across symbols of race profits divided by all trading volume. Panel A(ii) reports the distribution across symbols of race profits divided by non-race trading volume. Panel B is the same except at the date level (with race profits, all volume and non-race volume each summed across all symbols) instead of at the symbol level. All analyses are conducted separately for FTSE 100, FTSE 250, and full sample.

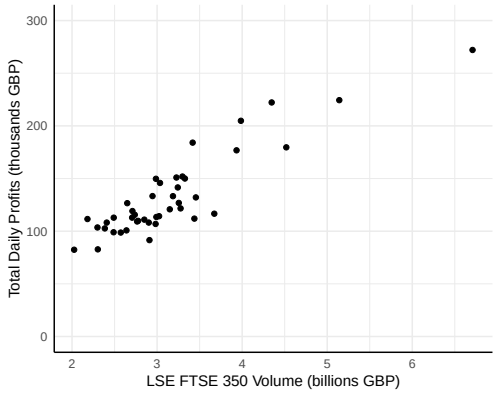
### Relationship between Profits, Volume and Volatility

Figure 5.5 presents scatterplots of latency arbitrage profits against trading volume (Panel A) and 1-minute realized volatility (Panel B). Each dot represents one day of our data. As can be seen, latency arbitrage profits are highly correlated to both volume and volatility. The $R^2$ of the relationship between profits and volume is 0.811 and the $R^2$ of the relationship between profits and 1-minute volatility is 0.661. These relationships are consistent with the theory in BCS, which suggests that the size of the latency arbitrage prize should be related to both volume and volatility.

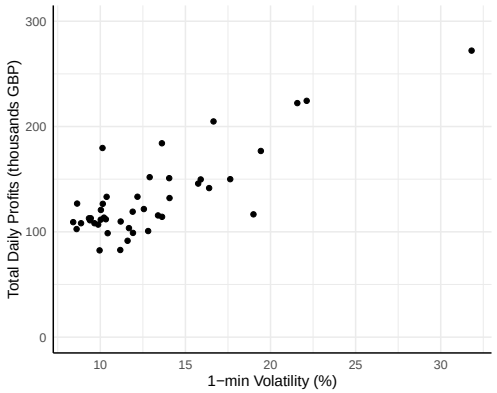
Figure 5.6 presents scatterplots of the latency arbitrage tax (Measure 1, all volume) against these same measures: trading volume (Panel A) and 1-minute realized volatility (Panel B). The

Figure 5.5: **Latency Arbitrage Profits Correlation with Volume and Volatility**

Panel A: Profit vs. Volume




Panel B: Profits vs. Volatility

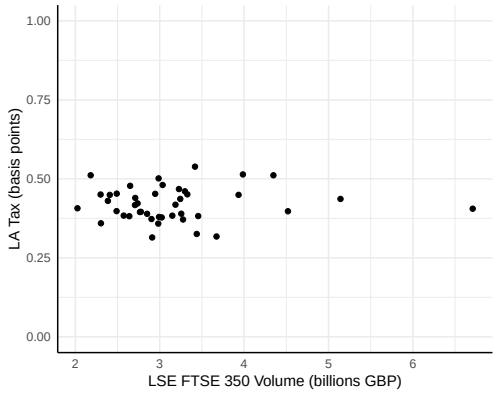



**Notes:** Panel A presents a scatterplot of daily race profits for the full sample, computed as in Table 5.8 (Panel B), against daily regular-hours trading volume (see notes for Table 5.3). Panel B presents a scatterplot of daily race profits for the full sample, against daily realized 1-minute volatility for the FTSE 350 index, computed using Thomson Reuters Tick History (TRTH) data.

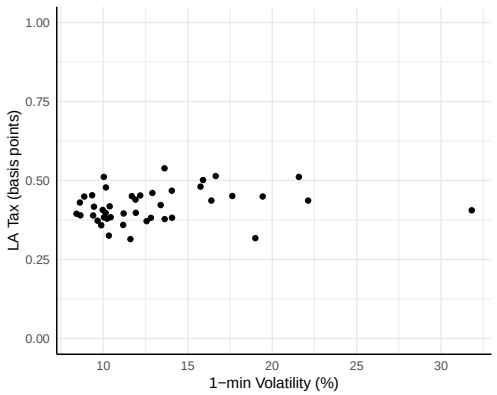
figures show that once we divide latency arbitrage profits by daily trading volume, to obtain the latency arbitrage tax in basis points, the result is relatively flat across the days in our sample. We will report further details on these relationships in Section 7, where they will be used for the purpose of out-of-sample extrapolation.

Figure 5.6: **Latency Arbitrage Tax Correlation with Volume and Volatility**

Panel A: LA Tax vs. Volume




Panel B: LA Tax vs. Volatility




**Notes:** Panel A presents a scatterplot of the daily latency arbitrage tax, defined as daily race profits for the full sample divided by daily regular-hours trading volume, against regular-hours trading volume. Panel B presents a scatterplot of the daily latency arbitrage tax against daily realized 1-minute volatility for the FTSE 350 index. Please see the notes for Figure 5.5 which is closely related.

## 5.5 Latency Arbitrage's Share of the Market's Cost of Liquidity

In this sub-section we quantify latency arbitrage as a proportion of the market's overall cost of liquidity. We present two distinct approaches.

### 5.5.1 Approach #1: Traditional Bid-Ask Spread Decomposition

An influential decomposition of the bid-ask spread (e.g., Glosten, 1987; Stoll, 1989; Hendershott, Jones and Menkveld, 2011) is:

$$\textit{EffectiveSpread} = \textit{PriceImpact} + \textit{RealizedSpread} \tag{5.1}$$

where *EffectiveSpread* is defined as the value-weighted difference between the transaction price and the midpoint at the time of the transaction, *PriceImpact* is defined as the value-weighted change between the midpoint at the time of the transaction and the midpoint at some time in the near future (e.g., 30 seconds), and *RealizedSpread* is the remainder. *EffectiveSpread* is typically interpreted as the revenue to liquidity providers from capturing the bid-ask spread, *PriceImpact* as the cost of adverse selection, and *RealizedSpread* as revenues net of adverse selection.

The theory of latency arbitrage as discussed in Section 4.1 suggests two refinements to (5.1). First, we can decompose the price impact component of the spread into two components: one that reflects latency arbitrage and one that reflects traditional private information. Specifically, for each symbol-day, we sum the value-weighted price impacts for all trades that are part of a latency arbitrage race, and we sum the value-weighted price impacts for all trades that are not part of a latency arbitrage race. Second, the theory shows that the equilibrium bid-ask spread also reflects the value of "losses avoided" by fast liquidity providers who successfully cancel in a latency arbitrage race. The intuition is that fast liquidity providers must earn a rent in equilibrium for being fast that is equal to the rent earned by fast traders who try to snipe; i.e., they earn the "opportunity cost of not sniping."

Formally, we start with equation (3.1) of Budish, Lee and Shim (2019), which gives the equilibrium bid-ask spread in the continuous limit order book (CLOB) market as

$$\lambda_{invest}\frac{s^{CLOB}}{2} = (\lambda_{public} + \lambda_{private}) \cdot L(\frac{s^{CLOB}}{2}), \tag{5.2}$$

with the notation defined as follows. $\lambda_{invest}$, $\lambda_{public}$ and $\lambda_{private}$ are, respectively, the Poisson arrival rates of investors who trade and thus pay the half-spread to a liquidity provider, publicly observed jumps in the fundamental value which cause a sniping race, and privately observed jumps in the fundamental value which lead to Glosten and Milgrom (1985) adverse selection. $s^{CLOB}$ denotes the equilibrium bid-ask spread. $L(\frac{s^{CLOB}}{2})$ denotes the expected loss to a liquidity provider, at this spread, if there is a jump in the fundamental value and they get sniped or adversely selected. In Appendix B.2 we show formally that equation (5.2) implies the spread decomposition:

$$\textit{EffectiveSpread} = \textit{PriceImpact}_{Race} + \textit{PriceImpact}_{NonRace} + \textit{LossAvoidance} + \textit{RealizedSpread} \tag{5.3}$$

with terms defined as follows. *EffectiveSpread* is defined in the standard way, as the value-weighted absolute difference between the price paid in trades and the midpoint at the time of the trade (i.e., the value-weighted half-spread). $PriceImpact_{Race}$ and $PriceImpact_{NonRace}$ are, respectively, the value-weighted change between the midpoint at the time of the trade and the midpoint at some time in the near future (we will use 10 seconds), for trades in latency-arbitrage races and trades not in latency-arbitrage races. That is we take the usual definition of *PriceImpact* and decompose it into two components, for trades in and not in races, respectively, so that $PriceImpact = PriceImpact_{Race} + PriceImpact_{NonRace}$. Last, *LossAvoidance* is defined as the value-weighted change between the race price and the midpoint in the near future for successful cancels in latency arbitrage races. Note that *LossAvoidance* is calculated as race price to midpoint, whereas $PriceImpact_{Race}$ is calculated as midpoint to midpoint. This difference reflects the fact that *LossAvoidance* measures trades that a fast liquidity provider *avoided*, so no liquidity taker paid the effective spread; in contrast, in races won by an aggressor, the aggressor paid the effective spread and the liquidity provider's losses are price impact less this effective spread they collected.

Table 5.10 gives details for decomposition (5.3) at the symbol level. For the average symbol in the FTSE 100, averaged over the days of our data set, the overall effective spread is 3.27 basis points, of which price impact is 3.62 basis points, loss avoidance is 0.01 basis points, and realized spread is -0.36 basis points. That price impact slightly exceeds the effective spread, so that the realized spread is slightly negative, is relatively common in modern markets, as noted in O'Hara (2015), and documented in Battalio, Corwin and Jennings (2016); Malinova, Park and Riordan (2018); Baron et al. (2019). That loss avoidance is small is consistent with our finding earlier that most race activity is aggressive.

The FTSE 100 overall effective spread of 3.27 basis points reflects relatively similar effective spreads in races and outside of races, at 3.18 and 3.29 basis points, respectively. Price impact, in contrast, is meaningfully higher in races than not in races: 5.11 basis points versus 3.15 basis points. Consequently, the realized spread is -1.93 basis points in races versus +0.15 basis points not in races.[41] This result suggests that liquidity provision is modestly profitable in non-race trading but loses significant money in races. Note as well that this negative realized spread in races obtains even at the 99th percentile of FTSE 100 symbols (-0.88 basis points), which suggests that the finding is robust in the cross section of symbols.

Aggregated over all trading volume, price impact in races accounts for about 37% of the effective spread and 33% of all price impact in FTSE 100 stocks. Since price impact is an object of per se interest to market microstructure researchers, the finding that a substantial percentage of price impact occurs in latency arbitrage races is potentially of interest for the literature.

For symbols in the FTSE 250,[42] overall effective spreads are higher, at 8.06 basis points, realized

[41] Note that the realized spread in races, multiplied by the roughly 22% of trading volume in races as reported in Table 5.3, corresponds roughly to the all-volume latency-arbitrage tax as reported in Table 5.9. (The relationship is not exact due to loss avoidance, which we count as part of the latency-arbitrage prize but does not count towards realized spreads, and some small differences in how the data are aggregated). Conceptually, the negative realized spread in races and the latency-arbitrage tax are two very similar ways of expressing the harm to liquidity providers.

[42] This table conditions on the symbol having at least 100 races in the sample period, or a bit more than 2 per

Table 5.10: **Spread Decomposition**

Panel A: FTSE 100 by Symbol

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| Effective spread paid - overall (bps) | 3.27 | 1.22 | 1.22 | 1.75 | 2.28 | 3.18 | 4.13 | 4.91 | 5.79 |
| Effective spread paid - in races (bps) | 3.18 | 1.22 | 0.99 | 1.70 | 2.21 | 3.17 | 4.05 | 4.89 | 5.98 |
| Effective spread paid - not in races (bps) | 3.29 | 1.22 | 1.25 | 1.78 | 2.30 | 3.17 | 4.15 | 4.96 | 5.71 |
| Price impact - overall (bps) | 3.62 | 1.36 | 1.40 | 1.92 | 2.52 | 3.56 | 4.52 | 5.55 | 6.99 |
| Price impact - in races (bps) | 5.11 | 1.83 | 2.02 | 2.85 | 3.48 | 4.90 | 6.50 | 7.56 | 8.81 |
| Price impact - not in races (bps) | 3.15 | 1.16 | 1.21 | 1.66 | 2.21 | 3.17 | 3.97 | 4.67 | 5.99 |
| Loss avoidance (bps) | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.03 |
| Realized spread - overall (bps) | -0.36 | 0.32 | -1.07 | -0.76 | -0.55 | -0.35 | -0.17 | 0.01 | 0.39 |
| Realized spread - in races (bps) | -1.93 | 0.70 | -3.72 | -2.83 | -2.40 | -1.79 | -1.42 | -1.11 | -0.88 |
| Realized spread - not in races (bps) | 0.15 | 0.30 | -0.35 | -0.20 | -0.05 | 0.08 | 0.34 | 0.56 | 0.90 |
| PI in races / PI total (%) | 33.16 | 6.09 | 19.99 | 24.88 | 29.53 | 32.13 | 37.23 | 41.72 | 44.72 |
| PI in races / Effective spread (%) | 36.90 | 7.18 | 19.79 | 27.73 | 33.06 | 36.59 | 41.97 | 46.44 | 51.67 |

Panel B: FTSE 250 by Symbol

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| Effective spread paid - overall (bps) | 8.06 | 3.81 | 2.65 | 4.63 | 5.59 | 7.14 | 9.84 | 13.10 | 19.11 |
| Effective spread paid - in races (bps) | 6.74 | 3.03 | 2.42 | 4.32 | 4.97 | 6.08 | 7.63 | 9.96 | 15.62 |
| Effective spread paid - not in races (bps) | 8.22 | 3.87 | 2.72 | 4.70 | 5.72 | 7.31 | 9.94 | 13.34 | 19.55 |
| Price impact - overall (bps) | 8.09 | 3.54 | 2.64 | 4.96 | 5.71 | 7.10 | 9.40 | 12.95 | 19.91 |
| Price impact - in races (bps) | 12.22 | 6.19 | 4.04 | 7.17 | 8.82 | 10.72 | 13.75 | 18.12 | 33.42 |
| Price impact - not in races (bps) | 7.50 | 3.52 | 2.36 | 4.37 | 5.09 | 6.40 | 8.79 | 12.39 | 19.39 |
| Loss avoidance (bps) | 0.01 | 0.02 | -0.02 | 0.00 | 0.00 | 0.01 | 0.01 | 0.02 | 0.07 |
| Realized spread - overall (bps) | -0.04 | 1.14 | -2.30 | -1.02 | -0.53 | -0.14 | 0.34 | 0.96 | 2.67 |
| Realized spread - in races (bps) | -5.48 | 3.68 | -20.22 | -9.36 | -6.14 | -4.43 | -3.44 | -2.73 | -1.62 |
| Realized spread - not in races (bps) | 0.72 | 1.07 | -0.97 | -0.13 | 0.20 | 0.59 | 1.07 | 1.76 | 3.14 |
| PI in races / PI total (%) | 21.60 | 9.50 | 1.79 | 6.00 | 14.89 | 22.98 | 28.19 | 32.16 | 39.60 |
| PI in races / Effective spread (%) | 22.50 | 10.92 | 1.58 | 5.62 | 14.84 | 23.57 | 30.44 | 34.79 | 47.67 |

Panel C: Full Sample by Date

| Description | Mean | sd | Min | Pct10 | Pct25 | Median | Pct75 | Pct90 | Max |
|---|---|---|---|---|---|---|---|---|---|
| Effective spread paid - overall (bps) | 3.17 | 0.27 | 2.74 | 2.92 | 3.06 | 3.12 | 3.22 | 3.38 | 4.52 |
| Effective spread paid - in races (bps) | 2.99 | 0.13 | 2.64 | 2.84 | 2.90 | 2.99 | 3.06 | 3.16 | 3.28 |
| Effective spread paid - not in races (bps) | 3.22 | 0.32 | 2.77 | 2.95 | 3.09 | 3.17 | 3.29 | 3.44 | 4.90 |
| Price impact - overall (bps) | 3.38 | 0.19 | 2.96 | 3.19 | 3.23 | 3.38 | 3.52 | 3.61 | 3.80 |
| Price impact - in races (bps) | 4.82 | 0.24 | 4.35 | 4.53 | 4.66 | 4.79 | 4.99 | 5.07 | 5.55 |
| Price impact - not in races (bps) | 2.99 | 0.19 | 2.57 | 2.79 | 2.86 | 2.95 | 3.13 | 3.29 | 3.38 |
| Loss avoidance (bps) | 0.01 | 0.00 | -0.01 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 |
| Realized spread - overall (bps) | -0.22 | 0.23 | -0.62 | -0.38 | -0.31 | -0.26 | -0.15 | -0.09 | 1.08 |
| Realized spread - in races (bps) | -1.83 | 0.17 | -2.43 | -2.01 | -1.92 | -1.81 | -1.74 | -1.64 | -1.51 |
| Realized spread - not in races (bps) | 0.23 | 0.26 | -0.17 | 0.05 | 0.14 | 0.20 | 0.29 | 0.34 | 1.68 |
| PI in races / PI total (%) | 30.58 | 2.64 | 22.91 | 27.88 | 29.88 | 30.81 | 31.93 | 33.39 | 35.81 |
| PI in races / Effective spread (%) | 32.82 | 3.73 | 17.38 | 29.92 | 31.60 | 33.66 | 34.70 | 36.54 | 39.52 |

**Notes:** Please see the text of Section 5.5 for definitions of Effective Spread, Price Impact (PI), Loss Avoidance, and Realized Spread. Panel A reports the distribution of these statistics by symbol, for all symbols in the FTSE 100. Panel B reports the distribution for all symbols in the FTSE 250. We only include symbols that have at least 100 races summed over all dates; this drops about one-quarter of FTSE 250 symbols and does not drop any FTSE 100 symbols. Panel C reports the distribution of these statistics by date for the full sample.

spreads are a bit less negative at -0.04 basis points, and loss avoidance remains small (0.01 basis points). Effective spreads are noticeably a bit narrower in races versus not in races, at 6.74 basis points in races versus 8.22 basis points outside of races.[43] As with FTSE 100 stocks, price impact is significantly higher in races than in non-race trading (12.22 basis points versus 7.50 basis points), and consequently the realized spread is modestly positive in non-race trading (0.72 basis points) and meaningfully negative in races (-5.48 basis points). Aggregated over all trading volume, price impact in races acounts for about 22% each of the effective spread and of all price impact in FTSE 250 stocks.

In the full sample, value-weighted, the effective spread is 3.17 basis points, the realized spread is -1.83 basis points in races versus +0.23 basis points not in races, and price impact in races accounts for 30.58% of all price impact and 32.82% of the overall effective spread.

**The Realized Spread is Negative in Races for Both Fast and Slow Firms** Importantly, this negative realized spread in races does not appear to discriminate much by firm speed. For the top 6 firms as defined by the proportion of races won (see Figure 5.2) the realized spread in races is -1.699 basis points, versus -1.930 basis points for firms outside the top 6. The difference between the Takers and Balanced firms in the top 6 is small as well: -1.493 basis points versus -1.775 basis points. Please see Table 5.11.

Similarly, both fast and slow firms earn a modestly positive realized spread in non-race liquidity provision. For the top 6 firms the realized spread in non-race liquidity provision is 0.347 basis points versus 0.152 basis points for firms outside the top 6.

There is a more significant difference between faster and slower firms in their canceling behavior. The top 6 firms attempt to cancel in races about 35% of the time within the race horizon, and about 39% of the time within 1 millisecond of the starting time of the race. Within these top 6 firms, the maximum cancel rate is 66% within the race-horizon and 68% of the time within 1 millisecond. Firms outside of the top 6 attempt to cancel just 7.57% of the time within races and 9.47% of the time within 1 millisecond of the starting time of the race. If we look beyond 1 millisecond to include any failed cancel attempts of quotes taken in a race, the top 6 cancel attempt rate goes up to 40% and the cancel rate for firms outside of the top 6 goes up to 13.35%.[44] Thus, fast firms are about five times more likely to attempt to cancel in a race than are slower firms.

Together, these results reinforce the idea that latency arbitrage imposes a tax on liquidity provision — it is expensive to be the liquidity provider who gets sniped in a race. The fastest

day, to ensure that the comparisons between races and non-races is meaningful. This drops a bit over a quarter of FTSE 250 symbols. The dropped symbols have noticeably wider effective spreads than the FTSE 250 symbols with non-trivial race activity.

[43]The narrower spread in FTSE 250 races versus in non-race trading activity could reflect an investor or trading firm triggering a race by submitting a limit order that sufficiently narrows the spread, as in models of Foucault, Kozhan and Tham (2016) (part of what the paper calls nontoxic arbitrage) and Li, Wang and Ye (2020). The results for the FTSE 100 suggest that this is not an important empirical phenomenon in FTSE 100 stocks.

[44]For firms in the top 6 essentially all of the incremental failed cancels come within 3 milliseconds after the race start (98.57% of all cancel attempts are within 3ms of the race start). For firms outside the top 6 the large majority of the incremental failed cancels come by 3 milliseconds after the race start (85.73%), and essentially all come by 1 second after the race start (99.43%).

Table 5.11: **Realized Spreads in Races by Firm Group**

| | Realized Spread (bps) | | | Cancel Attempt Rate (%) | | |
|---|---|---|---|---|---|---|
| **Firm Group** | Overall | Non-Race | Race | In Race | Within 1ms | Ever |
| All Firms | −0.209 | 0.236 | −1.833 | 19.29 | 21.89 | 24.53 |
| Fast vs. Slow | | | | | | |
| Top 6 | −0.086 | 0.347 | −1.699 | 35.35 | 38.94 | 39.88 |
| All Others | −0.302 | 0.152 | −1.930 | 7.57 | 9.47 | 13.35 |
| Within Fast | | | | | | |
| Takers in Top 6 | 0.016 | 0.455 | −1.493 | 45.16 | 47.56 | 47.82 |
| Balanced in Top 6 | −0.120 | 0.311 | −1.775 | 30.97 | 35.09 | 36.33 |

**Notes:** Firm groups are as in Figure 5.3. The realized spread is calculated as described in the text and reported in Table 5.10. To calculate the cancel attempt rates we first compute, for each firm, the number of races in which they have a cancel attempt within the race horizon, the number of races in which they either have a cancel attempt within the race horizon or a cancel attempt within 1 millisecond of the start of the race for an order taken in the race, the number of races in which they either have a cancel attempt within the race horizon or a cancel attempt anytime after the race horizon for an order taken in the race, and the number of races in which they either have a successful cancel or provide liquidity (each is measured at the relevant price and side for the race). We then aggregate into the firm-group cancel rates by, for the numerator, summing the number of races with cancel attempts over all firms in the group (possibly counting the same race multiple times), and for the denominator, summing the number of races with either cancel attempts or liquidity provision over all firms in the group (possibly counting the same race multiple times).

firms are better than slower firms at avoiding this cost, but even they get sniped with significant probability if their quotes become stale.

### 5.5.2 Approach #2: Implied Reduction of the Bid-Ask Spread if Latency Arbitrage Were Eliminated

Our second approach asks what would be the proportional reduction in the market cost of liquidity if there were no latency arbitrage. Formally, we seek to empirically measure:

$$\frac{\frac{s^{CLOB}}{2} - \frac{s^{FBA}}{2}}{\frac{s^{CLOB}}{2}} \tag{5.4}$$

where $s^{CLOB}$ is the bid-ask spread under the continuous limit order book (CLOB) and $s^{FBA}$ is the bid-ask spread under a counterfactual market design, frequent batch auctions (FBA), which eliminates latency arbitrage. To turn (5.4) into something empirically measurable, we take the following steps. First, we multiply the numerator and denominator of (5.4) by $(\lambda_{invest} + \lambda_{private})$. Second, we use (5.2) to solve out for $\lambda_{invest}\frac{s^{CLOB}}{2}$ in the numerator. Third, we use equation (5.1) of Budish, Lee and Shim (2019),

$$\lambda_{invest}\frac{s^{FBA}}{2} = \lambda_{private} \cdot L(\frac{s^{FBA}}{2}) \tag{5.5}$$

where $L(\frac{s^{FBA}}{2})$ is the loss to the liquidity provider if there is a privately-observed jump of at least $\frac{s^{FBA}}{2}$ and they get adversely selected, to solve out for $\lambda_{invest}\frac{s^{FBA}}{2}$ in the numerator of (5.4). Observe that the difference between the equilibrium bid-ask spread characterization for

frequent batch auctions, (5.5), and the equilibrium bid-ask spread for continuous trading, (5.2), is the $\lambda_{public}L(\cdot)$ term; if there is a publicly-observed jump a liquidity provider in an FBA does not get sniped, unlike in the continuous market.

These manipulations and some algebra, included in Appendix B.3 for completeness, shows that equation (5.4) can be re-expressed as:

$$\frac{\frac{s^{CLOB}}{2} - \frac{s^{FBA}}{2}}{\frac{s^{CLOB}}{2}} = \frac{\lambda_{public}L(\frac{s^{CLOB}}{2})}{(\lambda_{invest} + \lambda_{private})\frac{s^{CLOB}}{2}} \tag{5.6}$$

Both the numerator and denominator of the right-hand-side of (5.6) are directly measurable. The numerator is simply latency arbitrage profits (including both races where an aggressor wins and races where a cancel wins). The denominator is the non-race portion of the effective spread; that is, it is all of the bid-ask spread revenue collected by liquidity providers outside of latency arbitrage races. These objects can be measured either in GBP terms, or, by dividing both numerator and denominator by non-race trading volume, in basis points terms. Thus, we have the relationship:

$$\begin{aligned}\text{Proportional Reduction in Liquidity Cost} &= \frac{\text{Race Profits (GBP)}}{\text{Non-Race Effective Spread (GBP)}} \\ &= \frac{\text{Latency Arbitrage Tax (Non-Race Volume)}}{\text{Non-Race Effective Spread (bps)}}\end{aligned} \tag{5.7}$$

Table 5.12 presents our computation of (5.7). For the average symbol in the FTSE 100, eliminating latency arbitrage would reduce the cost of liquidity by 19.95%. For the FTSE 250, the figure is 11.93%. Even though race profits are higher as a proportion of trading volume for the FTSE 250 (per Table 5.9), bid-ask spreads are several times wider for FTSE 250 symbols than for FTSE 100 symbols, so eliminating latency arbitrage would reduce the overall cost of liquidity by less for the FTSE 250 than for the FTSE 100.

For the market as a whole, value-weighted and averaging over all dates in our sample, eliminating latency arbitrage would reduce the cost of liquidity by 16.73%.

# 6 Sensitivity Analysis

In this section we present sensitivity analyses for the main results presented in Section 5.

Section 6.1 explores sensitivity to the race horizon, i.e., to the definition of what counts as "at the same time." Section 6.2 explores sensitivity to the number of race participants, e.g., requiring 3+ participants at the same time rather than 2+. Section 6.3 explores sensitivity to requiring cancel attempts in the race, i.e., to not counting races that contain only aggressive orders, and also explores stricter requirements on the number of aggressive orders. Section 6.4 explores varying the definition of what counts as a success and a fail. Together, then, Sections 6.1-6.4 explore sensitivity to the four components of our race definition: multiple participants, at the same time, at least some of

Table 5.12: **Percentage Reduction in Liquidity Cost, if Latency Arbitrage Eliminated**

Panel A: Symbol level

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 19.95 | 5.29 | 8.87 | 13.30 | 16.79 | 19.69 | 23.58 | 26.50 | 32.54 |
| FTSE 250 | 11.93 | 6.31 | 0.58 | 3.12 | 8.05 | 11.91 | 15.33 | 18.58 | 31.31 |
| Full Sample | 14.77 | 7.09 | 0.70 | 5.55 | 10.03 | 14.55 | 19.41 | 24.10 | 32.22 |

Panel B: Date level

| Description | Mean | sd | Min | Pct10 | Pct25 | Median | Pct75 | Pct90 | Max |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 19.06 | 3.29 | 7.49 | 16.53 | 17.53 | 18.97 | 21.48 | 22.25 | 25.40 |
| FTSE 250 | 11.39 | 1.66 | 8.27 | 9.43 | 10.22 | 11.17 | 12.45 | 13.36 | 16.18 |
| Full Sample | 16.73 | 2.57 | 7.88 | 14.57 | 15.19 | 16.82 | 18.66 | 19.17 | 21.58 |

**Notes:** For each symbol, we implement equation (5.7) by dividing total race profits in GBP, across all dates, and dividing by total non-race Effective Spread paid in GBP, across all dates. Race profits in GBP are as described in Table 5.8 and Effective Spread paid in GBP is as described in Table 5.10. Analogously, for each date, we implement equation (5.7) by dividing total race profits in GBP, across all symbols, and dividing by total non-race Effective Spread paid in GBP, across all symbols. We do both exercises separately for FTSE 100, FTSE 250, and full sample. As in Table 5.10, the symbol-level measures drop symbols with fewer than 100 races summed across all dates in our sample.

whom are aggressive, and at least some of whom succeed and some of whom fail. In Section 6.5 we combine the insights from all of the sensitivity analyses to discuss lower and upper bounds on our measures of race profits and the harm to liquidity provision.

## 6.1 Sensitivity to Race Horizon

As a reminder, our baseline method requires that messages satisfying the baseline race requirements (i.e., 2+ messages from distinct users, 1+ aggressing, 1+ success, and 1+ fail) arrive within the "information horizon" of the first message of the race or 500 microseconds, whichever is smaller. The information horizon, which is the window of time such that we can be essentially certain that inbound messages in the race are not responding to earlier messages' outbound reports (see Section 4.3) has a median of 186 microseconds in our data. The 500 microsecond truncation binds 4% of the time.

Table 6.1 presents sensitivity analysis for changes to the race horizon. The first column of the table re-presents our main results from Section 5 for this baseline specification, to facilitate comparison. The next set of columns presents these same results using fixed race horizons of varying lengths, from 50 microseconds to 3 milliseconds. That is, instead of using the information horizon method, under which the race window will vary with the observed lag in information processing by the LSE's matching engine, we just fix a time window, and consider a wide range of such windows. The 50 microsecond window roughly corresponds to the minimum observed information horizon (which is 43 microseconds), the 200 microsecond window roughly corresponds to the median observed information horizon, and 500 microseconds corresponds to the upper bound on the information horizon we determined in consultation with FCA supervisory experts. The horizons beyond that are included to capture races among firms of varying technological sophistication that could still be

considered racing one another. For instance, the threshold should be wide enough to include a firm that is not utilizing the fastest connections to exchanges in the United States or elsewhere, but is using the next-fastest.[45] We consulted with HFT industry contacts and FCA supervisors to agree on an appropriate horizon. Following these discussions, we determined 3 milliseconds would capture most of these additional potential races, though for races originating from signals far from London (e.g., Chicago) differences in speed between cutting-edge HFTs and relatively sophisticated firms could easily exceed that number. The last set of columns runs a sensitivity analysis specifically on the choice of the truncation parameter for the information horizon method.

Focus first on the number of races per day per symbol in the FTSE 100, the first row of the table. In the baseline there are 537 races per symbol per day. In the 50 microsecond column, this number is reduced to 297. As the race horizon increases, so does the number of races detected. The growth is especially steep up to 500 microseconds, reaching 793 races per symbol per day, and then tapers off, with 870 races at a horizon of 1 millisecond and 946 races at a horizon of 3 milliseconds. Varying the truncation parameter for the information horizon method does not yield much additional insight beyond what is already learned from the baseline and the fixed horizon columns. Using a 100 microsecond truncation parameter yields results that are very similar to the 100 microsecond fixed race horizon, which makes sense since this truncation parameter will bind most of the time. Using a 1 millisecond truncation parameter yields results that are similar to the baseline with the 500 microsecond truncation parameter, which again makes sense because neither truncation parameter will bind very much.

Turn next to the measures of per-race profits. Interestingly, per-race profits, whether measured per-share (ticks, pence (GBX), basis points) or in GBP per-race (either displayed depth or quantity actually traded/canceled), are relatively similar across these different specifications. This tells us that the additional races being picked up by the longer race horizons are, on average, of similar profitability to the races being picked up at shorter race horizons. This will not be the case for some of the subsequent sensitivities.

As a result, the latency arbitrage tax measures are all increasing with the race horizon. At a 50 microsecond race horizon, the FTSE 350 latency arbitrage tax, using the all-volume measure, is 0.20 basis points, versus 0.42 basis points in our baseline specification. At the 3 millisecond race horizon, the latency arbitrage tax is 0.81 basis points, or 4 times higher, roughly proportional to the increase in the number of races. The effect on the second measure of the latency arbitrage tax, based on non-race trading volume, is even larger, because as the numerator (race profits) is increasing, the denominator (non-race volume) is also shrinking. This figure increases from 0.22 basis points at 50 microseconds, to 0.53 basis points in our baseline specification, all the way up to 1.55 basis points at 3 milliseconds. For FTSE 250 stocks, the latency arbitrage tax is as high as 2.49 basis points at 3 milliseconds.

Last we discuss the implied reduction in the cost of liquidity. In our baseline, eliminating latency arbitrage would reduce the cost of liquidity by 20.0% for the average FTSE 100 symbol and by 16.7%

[45] Other sources of speed differential include using code and hardware that is not optimized for speed, not being co-located, and not using microwave connections where possible to do so.

Table 6.1: **Sensitivity Analysis: Different Race Horizons**

| Measure | Baseline | Fixed Race Horizon of Duration T | | | | | | | Info Horizon, Max T | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 50$\mu$s | 100$\mu$s | 200$\mu$s | 500$\mu$s | 1ms | 2ms | 3ms | 100$\mu$s | 1ms |
| **Frequency and Duration of Races** | | | | | | | | | | |
| Races per day | | | | | | | | | | |
| FTSE 100 - per symbol | 537.24 | 296.66 | 388.58 | 521.53 | 793.01 | 869.73 | 921.08 | 946.48 | 387.96 | 542.99 |
| FTSE 250 - per symbol | 70.05 | 41.37 | 52.78 | 69.22 | 112.99 | 127.04 | 134.06 | 138.37 | 52.71 | 70.28 |
| Mean race duration (microseconds) | 78.65 | 16.12 | 30.80 | 72.18 | 194.20 | 304.96 | 450.87 | 572.12 | 30.61 | 84.85 |
| % of races with wrong winner | 4.30 | 8.18 | 6.41 | 4.21 | 1.98 | 1.67 | 1.43 | 1.32 | 6.42 | 4.24 |
| % of volume in races | | | | | | | | | | |
| FTSE 100 | 22.15 | 9.99 | 13.65 | 19.71 | 37.43 | 43.53 | 47.11 | 48.61 | 13.64 | 22.65 |
| FTSE 250 | 16.90 | 8.36 | 11.20 | 15.99 | 33.34 | 38.37 | 41.23 | 42.63 | 11.20 | 17.07 |
| Full Sample | 21.46 | 9.77 | 13.32 | 19.21 | 36.88 | 42.84 | 46.33 | 47.82 | 13.32 | 21.92 |
| Mean number of messages within 500 $\mu$s | 3.46 | 3.51 | 3.51 | 3.51 | 3.39 | 3.01 | 2.83 | 2.76 | 3.51 | 3.44 |
| **Per-Race Profits** | | | | | | | | | | |
| Per-share profits | | | | | | | | | | |
| ticks | 0.55 | 0.54 | 0.53 | 0.51 | 0.53 | 0.55 | 0.56 | 0.57 | 0.53 | 0.56 |
| GBX | 0.17 | 0.16 | 0.16 | 0.16 | 0.16 | 0.16 | 0.17 | 0.16 | 0.16 | 0.17 |
| basis points | 1.66 | 1.68 | 1.63 | 1.57 | 1.61 | 1.64 | 1.65 | 1.65 | 1.63 | 1.67 |
| Per-race profits GBP | | | | | | | | | | |
| displayed depth | 1.85 | 1.58 | 1.59 | 1.60 | 1.84 | 1.94 | 1.97 | 1.97 | 1.60 | 1.90 |
| qty trade/cancel | 1.76 | 1.38 | 1.44 | 1.51 | 1.84 | 1.95 | 2.00 | 2.00 | 1.44 | 1.81 |
| **Aggregate Profits and LA Tax** | | | | | | | | | | |
| Daily Profits | | | | | | | | | | |
| FTSE 100 - per symbol | 1,047 | 490 | 647 | 872 | 1,520 | 1,769 | 1,909 | 1,965 | 647 | 1,089 |
| FTSE 250 - per symbol | 108 | 57 | 73 | 96 | 184 | 211 | 226 | 231 | 73 | 110 |
| Full Sample - aggregate | 132,378 | 63,573 | 83,233 | 111,722 | 198,700 | 230,586 | 248,291 | 255,408 | 83,181 | 137,173 |
| Latency Arbitrage Tax, All Volume (bps) | | | | | | | | | | |
| FTSE 100 | 0.38 | 0.18 | 0.24 | 0.32 | 0.56 | 0.65 | 0.70 | 0.72 | 0.24 | 0.40 |
| FTSE 250 | 0.66 | 0.35 | 0.45 | 0.59 | 1.13 | 1.30 | 1.39 | 1.42 | 0.45 | 0.68 |
| Full Sample | 0.42 | 0.20 | 0.26 | 0.35 | 0.63 | 0.73 | 0.78 | 0.81 | 0.26 | 0.43 |
| Latency Arbitrage Tax, Non-Race Volume (bps) | | | | | | | | | | |
| FTSE 100 | 0.49 | 0.20 | 0.27 | 0.40 | 0.89 | 1.15 | 1.32 | 1.40 | 0.27 | 0.52 |
| FTSE 250 | 0.80 | 0.38 | 0.50 | 0.71 | 1.70 | 2.12 | 2.37 | 2.49 | 0.50 | 0.82 |
| Full Sample | 0.53 | 0.22 | 0.30 | 0.44 | 1.00 | 1.28 | 1.47 | 1.55 | 0.30 | 0.56 |
| **Spread Decomposition** | | | | | | | | | | |
| Price impact in races / All price impact % | 30.58 | 12.84 | 17.89 | 25.69 | 49.79 | 58.71 | 64.34 | 66.82 | 17.88 | 31.76 |
| Price impact in races / Effective spread % | 32.82 | 13.77 | 19.19 | 27.57 | 53.42 | 62.99 | 69.03 | 71.69 | 19.19 | 34.08 |
| Loss avoidance / Effective spread % | 0.19 | 0.07 | 0.13 | 0.26 | 0.53 | 0.94 | 1.31 | 1.48 | 0.13 | 0.20 |
| **Implied Reduction in Cost of Liquidity** | | | | | | | | | | |
| % Reduction in liquidity cost | | | | | | | | | | |
| FTSE 100 - by symbol | 19.95 | 7.98 | 10.97 | 15.91 | 35.73 | 46.95 | 55.24 | 59.20 | 10.97 | 21.00 |
| FTSE 250 - by symbol | 11.93 | 6.17 | 7.96 | 10.79 | 24.36 | 28.46 | 31.70 | 32.90 | 7.95 | 12.17 |
| Full Sample - by date | 16.73 | 6.96 | 9.49 | 13.62 | 30.38 | 39.20 | 45.62 | 48.75 | 9.49 | 17.49 |

**Notes:** For descriptions of the sensitivity scenarios please see the text of Section 6.1. Descriptions of each of the items in this table can be found in the following table notes in Section 5. Races per day: Table 5.1. Mean race duration and % of races with wrong winner: Table 5.2. % of Volume in Races: Table 5.3. Mean number of messages: Table 5.4. Per-race profits: Table 5.6. Aggregate profits: Table 5.8. Latency Arbitrage Tax: Table 5.9. Spread decomposition: Table 5.10. Implied Reduction in Cost of Liquidity: Table 5.12.

for the market overall. Using a 50 microsecond race horizon lowers these figures to 8.0% and 7.0%, respectively. Using a 3 millisecond race horizon increases these figures all the way to 59.2% and 48.8%, respectively. Again, this large change relative to the baseline is driven by both the increase in the numerator (race profits) and decrease in the denominator (non-race effective spread paid).

### 6.2 Sensitivity to Number of Race Participants

Our baseline method requires that there are at least 2 race participants within the information horizon. Table 6.2 presents sensitivity analysis for requiring 3+ participants; the appendix presents the same table for 5+ participants. In both cases, the other race criteria are held the same, specifically we require 1+ aggressors, 1+ successes, and 1+ fails. Given the large effect that the race's time horizon had on the number of races and race profits, we include this sensitivity for multiple race horizons, including the baseline information horizon method and fixed race horizons from 50 microseconds to 3 milliseconds.

Focus first on the 3+ race participants within information horizon column; this column is exactly the same as the baseline but replacing 2+ race participants with 3+. Requiring 3+ race participants reduces the number of races by about 60%; for example, for the FTSE 100 the number of races per symbol per day declines from 537 to 229. However, these races are significantly more profitable, on a per-share basis and particularly on a GBP per-race basis. The net effect is that total race profits are reduced by about 30%. This roughly 30% reduction can be seen in the aggregate race profits measures, the latency arbitrage tax measures, and the liquidity cost reduction measures.

Increasing the race horizon increases the number of races detected, just as in the baseline case with 2+ participants. At a 50 microsecond race horizon there are 87 3+ participant races per day for the average FTSE 100 symbol, up to 482 races per symbol per day at a 500 microsecond race horizon, and up to 686 races at a 3 millisecond race horizon. With this increase in the number of races detected comes a commensurate increase in the various race profits measures and harm-to-liquidity measures.

We note that the 3+ race participants within 500 microseconds sensitivity is on most measures relatively similar to the baseline case of 2+ race participants within the information horizon. The number of races is a bit smaller but they are more profitable on average, with the net effect that the overall profits measures and liquidity-harm measures are about 20-30% higher than in the baseline. The 3+ race participants within 1 millisecond sensitivity yields a latency arbitrage tax (all-volume) of 0.65, versus 0.42 in baseline, and yields an implied harm to the cost of liquidity of 30.7%, versus 16.7% in baseline. In this sense, our baseline specification is meaningfully more conservative than the requirement of 3+ participants within 1 millisecond.

In the appendix we report a similar table for 5+ participants (Table A.10). There are very few (38) races per FTSE 100 symbol per day within the information horizon, versus 537 in the baseline and 229 with 3+. That said, these few races are quite profitable: they are about twice as profitable per share and more than three times as profitable in GBP per race as in the baseline. Increasing the race horizon to 500 microseconds yields 122 races per FTSE 100 symbol per day, and to 1

Table 6.2: **Sensitivity Analysis: Different Number of Race Participants**

| Measure | Baseline | 3+ Race Participants Within: InfoHor | 50$\mu$s | 100$\mu$s | 200$\mu$s | 500$\mu$s | 1ms | 2ms | 3ms |
|---|---|---|---|---|---|---|---|---|---|
| **Frequency and Duration of Races** | | | | | | | | | |
| Races per day | | | | | | | | | |
| FTSE 100 - per symbol | 537.24 | 228.98 | 86.73 | 134.38 | 236.14 | 482.47 | 585.98 | 655.57 | 685.98 |
| FTSE 250 - per symbol | 70.05 | 30.68 | 13.40 | 19.92 | 32.98 | 67.54 | 82.20 | 90.01 | 93.47 |
| Mean race duration (microseconds) | 78.65 | 77.56 | 14.26 | 28.54 | 75.54 | 194.56 | 305.95 | 449.76 | 553.83 |
| % of races with wrong winner | 4.30 | 5.08 | 10.66 | 7.88 | 4.59 | 2.01 | 1.71 | 1.44 | 1.33 |
| % of volume in races | | | | | | | | | |
| FTSE 100 | 22.15 | 12.75 | 3.84 | 6.32 | 11.57 | 27.97 | 35.53 | 39.99 | 41.83 |
| FTSE 250 | 16.90 | 9.33 | 3.38 | 5.24 | 9.35 | 23.68 | 29.59 | 32.95 | 34.39 |
| Full Sample | 21.46 | 12.30 | 3.78 | 6.17 | 11.28 | 27.40 | 34.74 | 39.06 | 40.85 |
| Mean number of messages within 500 $\mu$s | 3.46 | 4.68 | 4.83 | 4.82 | 4.62 | 4.21 | 3.58 | 3.28 | 3.17 |
| **Per-Race Profits** | | | | | | | | | |
| Per-share profits | | | | | | | | | |
| ticks | 0.55 | 0.71 | 0.73 | 0.71 | 0.64 | 0.61 | 0.63 | 0.64 | 0.65 |
| GBX | 0.17 | 0.23 | 0.23 | 0.22 | 0.20 | 0.19 | 0.19 | 0.19 | 0.19 |
| basis points | 1.66 | 2.24 | 2.36 | 2.29 | 2.03 | 1.90 | 1.91 | 1.92 | 1.92 |
| Per-race profits GBP | | | | | | | | | |
| displayed depth | 1.85 | 2.98 | 2.55 | 2.60 | 2.43 | 2.52 | 2.57 | 2.58 | 2.58 |
| qty trade/cancel | 1.76 | 2.87 | 2.29 | 2.40 | 2.33 | 2.55 | 2.62 | 2.65 | 2.64 |
| **Aggregate Profits and LA Tax** | | | | | | | | | |
| Daily Profits | | | | | | | | | |
| FTSE 100 - per symbol | 1,047 | 736 | 238 | 375 | 612 | 1,273 | 1,583 | 1,770 | 1,848 |
| FTSE 250 - per symbol | 108 | 70 | 27 | 41 | 65 | 147 | 181 | 200 | 208 |
| Full Sample - aggregate | 132,378 | 91,506 | 30,701 | 47,980 | 77,738 | 164,760 | 204,272 | 228,064 | 237,757 |
| Latency Arbitrage Tax, All Volume (bps) | | | | | | | | | |
| FTSE 100 | 0.38 | 0.27 | 0.09 | 0.14 | 0.23 | 0.47 | 0.58 | 0.65 | 0.68 |
| FTSE 250 | 0.66 | 0.43 | 0.17 | 0.25 | 0.40 | 0.90 | 1.11 | 1.23 | 1.28 |
| Full Sample | 0.42 | 0.29 | 0.10 | 0.15 | 0.25 | 0.52 | 0.65 | 0.72 | 0.75 |
| Latency Arbitrage Tax, Non-Race Volume (bps) | | | | | | | | | |
| FTSE 100 | 0.49 | 0.35 | 0.10 | 0.16 | 0.28 | 0.75 | 1.03 | 1.23 | 1.32 |
| FTSE 250 | 0.80 | 0.51 | 0.18 | 0.28 | 0.48 | 1.36 | 1.81 | 2.11 | 2.24 |
| Full Sample | 0.53 | 0.37 | 0.11 | 0.18 | 0.31 | 0.83 | 1.14 | 1.35 | 1.45 |
| **Spread Decomposition** | | | | | | | | | |
| Price impact in races / All price impact % | 30.58 | 19.13 | 5.64 | 9.34 | 16.39 | 38.37 | 48.96 | 55.92 | 58.99 |
| Price impact in races / Effective spread % | 32.82 | 20.54 | 6.05 | 10.03 | 17.61 | 41.17 | 52.54 | 60.01 | 63.31 |
| Loss avoidance / Effective spread % | 0.19 | 0.18 | 0.07 | 0.13 | 0.27 | 0.63 | 1.09 | 1.50 | 1.65 |
| **Implied Reduction in Cost of Liquidity** | | | | | | | | | |
| % Reduction in liquidity cost | | | | | | | | | |
| FTSE 100 - by symbol | 19.95 | 12.46 | 3.69 | 5.94 | 10.23 | 26.26 | 36.90 | 45.30 | 49.49 |
| FTSE 250 - by symbol | 11.93 | 7.67 | 3.03 | 4.50 | 7.09 | 19.05 | 24.47 | 28.10 | 28.85 |
| Full Sample - by date | 16.73 | 10.43 | 3.19 | 5.13 | 8.79 | 22.18 | 30.65 | 37.13 | 40.29 |

**Notes:** For descriptions of the sensitivity scenarios please see the text of Section 6.2. Descriptions of each of the items in this table can be found in the following table notes in Section 5. Races per day: Table 5.1. Mean race duration and % of races with wrong winner: Table 5.2. % of Volume in Races: Table 5.3. Mean number of messages: Table 5.4. Per-race profits: Table 5.6. Aggregate profits: Table 5.8. Latency Arbitrage Tax: Table 5.9. Spread decomposition: Table 5.10. Implied Reduction in Cost of Liquidity: Table 5.12.

millisecond yields 202 races per day, again with races that are signficantly more profitable per race than in the baseline. As a consequence, the sensitivity for 5+ participants within 500 microseconds yields overall profits that are about 60% of the baseline, and the sensitivity for 5+ participants within 1 millisecond yields overall profits and harm to liquidity that are just about the same as in the baseline.

The appendix also includes a sensitivity for requiring 2+ unique firms as opposed to our baseline requirement of 2+ unique participants (Table A.11). As mentioned earlier, some firms use different UserIDs for different trading desks. This sensitivity reduces the number of races and various profits measures by about 10%.

### 6.3 Sensitivity to Requiring Cancels or Multiple Takes

Our baseline method requires that of the 2+ messages in a race, at least 1 is aggressive. Thus, a race could have 1+ aggressive messages and 1+ cancel messages, or it could have 2+ aggressive messages and 0 cancel messages. Table 6.3 presents sensitivity analysis for these requirements. In the first set of columns after the baseline, we require 1+ cancel message and 1+ aggressive message, i.e., exclude races with 0 cancels (and hence 2+ aggressive messages). In the second set of columns, we require 2+ aggressive messages, i.e., exclude races with exactly 1 aggressive message (and hence 1+ cancel messages).

Focus first on the 1+ cancel within information horizon column. Requiring a cancel attempt within the race horizon window reduces the number of races significantly, from 537 to 173 per day for the average symbol in the FTSE 100. These races are also less profitable on average. This reduction in profitability is driven by races with exactly 1 aggressive message. If we require 2+ aggressive messages alongside a cancel (see Appendix Table A.12), profits per race are higher than in the baseline, especially in GBP per race where profits are nearly double.

Looking across the different race horizons does not change this picture much. The number of races goes up with the race horizon, as before, but the number of races and overall profitability are meaningfully smaller than without the 1+ cancel requirement, at all horizons. This pattern is consistent with our findings in Section 5 that most message activity in races is take attempts and most races are won by takers.

If we require at least 1 cancel within the information horizon, in addition to our other baseline race requirements, the harm to liquidity and the latency-arbitrage tax are each about 30% of baseline. That said, if we consider races with 1+ cancel within 3 milliseconds the results are closer to baseline, at about 85% of the harm to liquidity and level of latency-arbitrage tax.

Now focus on the columns that require at least 2 aggressive messages; that is, a race must have 2+ takes, along with 1+ success and 1+ fail, within the race horizon. Relative to the baseline, this excludes races with exactly 1 take and with 1+ cancels, which as we just discussed are relatively unprofitable. The number of races with 2+ takes within the information horizon is 424 for FTSE 100 symbols, versus 537 under the baseline scenario, a reduction of about 20%. These races are more profitable on average than the baseline races, so the net effect on profits and the harm-to-

Table 6.3: **Sensitivity Analysis: Requiring Cancels or Multiple Takes**

| Measure | Baseline | 1+ Cancel Within | | | | 2+ Takes Within | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | InfoHor | 50$\mu$s | 500$\mu$s | 1ms | 3ms | InfoHor | 50$\mu$s | 500$\mu$s | 1ms | 3ms |
| **Frequency and Duration of Races** | | | | | | | | | | | |
| Races per day | | | | | | | | | | | |
| FTSE 100 - per symbol | 537.24 | 172.70 | 71.55 | 242.59 | 303.68 | 380.88 | 423.86 | 241.67 | 695.44 | 774.40 | 851.41 |
| FTSE 250 - per symbol | 70.05 | 14.40 | 6.76 | 23.42 | 31.10 | 40.89 | 60.91 | 36.30 | 103.85 | 117.33 | 127.31 |
| Mean race duration (microseconds) | 78.65 | 92.77 | 19.05 | 206.89 | 373.48 | 768.59 | 74.52 | 15.42 | 194.72 | 300.04 | 547.27 |
| % of races with wrong winner | 4.30 | 3.15 | 7.42 | 2.01 | 1.50 | 0.99 | 4.63 | 8.34 | 1.89 | 1.62 | 1.30 |
| % of volume in races | | | | | | | | | | | |
| FTSE 100 | 22.15 | 8.49 | 2.31 | 12.71 | 17.30 | 22.75 | 17.40 | 8.39 | 33.90 | 40.55 | 46.15 |
| FTSE 250 | 16.90 | 3.31 | 1.08 | 6.17 | 9.21 | 13.02 | 15.20 | 7.65 | 32.02 | 37.04 | 41.19 |
| Full Sample | 21.46 | 7.82 | 2.15 | 11.87 | 16.26 | 21.49 | 17.11 | 8.28 | 33.64 | 40.08 | 45.49 |
| Mean number of messages within 500 $\mu$s | 3.46 | 3.36 | 3.26 | 3.65 | 3.09 | 2.73 | 3.66 | 3.65 | 3.50 | 3.11 | 2.85 |
| **Per-Race Profits** | | | | | | | | | | | |
| Per-share profits | | | | | | | | | | | |
| ticks | 0.55 | 0.37 | 0.24 | 0.37 | 0.39 | 0.40 | 0.62 | 0.62 | 0.57 | 0.59 | 0.62 |
| GBX | 0.17 | 0.11 | 0.07 | 0.11 | 0.12 | 0.12 | 0.19 | 0.19 | 0.17 | 0.18 | 0.18 |
| basis points | 1.66 | 0.99 | 0.70 | 1.03 | 1.11 | 1.14 | 1.92 | 1.93 | 1.75 | 1.79 | 1.82 |
| Per-race profits GBP | | | | | | | | | | | |
| displayed depth | 1.85 | 1.92 | 1.18 | 1.92 | 2.14 | 2.24 | 2.03 | 1.74 | 1.96 | 2.08 | 2.19 |
| qty trade/cancel | 1.76 | 1.82 | 0.95 | 1.83 | 2.07 | 2.19 | 1.92 | 1.54 | 1.97 | 2.11 | 2.24 |
| **Aggregate Profits and LA Tax** | | | | | | | | | | | |
| Daily Profits | | | | | | | | | | | |
| FTSE 100 - per symbol | 1,047 | 361 | 92 | 505 | 705 | 917 | 907 | 441 | 1,418 | 1,690 | 1,968 |
| FTSE 250 - per symbol | 108 | 15 | 5 | 28 | 44 | 64 | 104 | 55 | 181 | 209 | 233 |
| Full Sample - aggregate | 132,378 | 40,205 | 10,502 | 57,933 | 81,993 | 108,273 | 117,054 | 57,996 | 187,719 | 222,151 | 256,194 |
| Latency Arbitrage Tax, All Volume (bps) | | | | | | | | | | | |
| FTSE 100 | 0.38 | 0.13 | 0.03 | 0.19 | 0.26 | 0.34 | 0.33 | 0.16 | 0.52 | 0.62 | 0.72 |
| FTSE 250 | 0.66 | 0.10 | 0.03 | 0.18 | 0.27 | 0.39 | 0.63 | 0.34 | 1.11 | 1.28 | 1.43 |
| Full Sample | 0.42 | 0.13 | 0.03 | 0.19 | 0.26 | 0.35 | 0.37 | 0.18 | 0.59 | 0.70 | 0.81 |
| Latency Arbitrage Tax, Non-Race Volume (bps) | | | | | | | | | | | |
| FTSE 100 | 0.49 | 0.17 | 0.04 | 0.30 | 0.46 | 0.66 | 0.43 | 0.18 | 0.83 | 1.10 | 1.40 |
| FTSE 250 | 0.80 | 0.11 | 0.03 | 0.27 | 0.45 | 0.69 | 0.76 | 0.37 | 1.67 | 2.10 | 2.52 |
| Full Sample | 0.53 | 0.16 | 0.04 | 0.30 | 0.46 | 0.66 | 0.47 | 0.20 | 0.94 | 1.23 | 1.56 |
| **Spread Decomposition** | | | | | | | | | | | |
| Price impact in races / All price impact % | 30.58 | 11.86 | 2.79 | 16.95 | 23.55 | 31.72 | 24.14 | 11.02 | 44.66 | 53.98 | 63.58 |
| Price impact in races / Effective spread % | 32.82 | 12.73 | 2.99 | 18.20 | 25.28 | 34.05 | 25.91 | 11.83 | 47.92 | 57.92 | 68.22 |
| Loss avoidance / Effective spread % | 0.19 | 0.19 | 0.07 | 0.53 | 0.94 | 1.48 | 0.16 | 0.06 | 0.59 | 1.09 | 1.76 |
| **Implied Reduction in Cost of Liquidity** | | | | | | | | | | | |
| % Reduction in liquidity cost | | | | | | | | | | | |
| FTSE 100 - by symbol | 19.95 | 5.41 | 1.23 | 8.17 | 12.44 | 17.83 | 16.24 | 7.17 | 31.12 | 41.37 | 54.89 |
| FTSE 250 - by symbol | 11.93 | 1.57 | 0.57 | 2.85 | 4.43 | 6.60 | 11.32 | 5.94 | 24.11 | 27.57 | 32.13 |
| Full Sample - by date | 16.73 | 4.49 | 1.09 | 6.80 | 10.21 | 14.63 | 13.80 | 6.23 | 26.82 | 35.18 | 45.82 |

**Notes:** For descriptions of the sensitivity scenarios please see the text of Section 6.3. Descriptions of each of the items in this table can be found in the following table notes in Section 5. Races per day: Table 5.1. Mean race duration and % of races with wrong winner: Table 5.2. % of Volume in Races: Table 5.3. Mean number of messages: Table 5.4. Per-race profits: Table 5.6. Aggregate profits: Table 5.8. Latency Arbitrage Tax: Table 5.9. Spread decomposition: Table 5.10. Implied Reduction in Cost of Liquidity: Table 5.12.

liquidity measures is smaller, roughly 10-15%. This magnitude of reduction relative to the baseline requirements persists across the other time horizons.

These overall patterns, as discussed in Section 5 as well, are consistent with equilibria of the BCS model in which many of the fastest traders primarily engage in sniping as opposed to liquidity provision, and significant liquidity is provided by market participants not at the cutting edge of speed.

### 6.4 Sensitivity to Varying the Definitions of Success and Fail

Our baseline method defined success and fail as follows. A take attempt succeeds if it executes at least in part, and otherwise fails. A cancel attempt succeeds if at least some of the order's quantity is successfully canceled, and otherwise fails. As discussed in Section 4.2.3, while the definition of success might sound quite loose — e.g., if there are 10,000 shares in the book, an attempt to take 10,000 shares that "succeeds" in taking just 100 shares is counted as a success — it has some real bite in conjunction with the requirement that a race has a fail, because someone else likely got or canceled the other 9,900 shares, for there then to be yet another participant who then fails to get anything or cancel anything. The exception is if there is a successful take attempt for a small amount (e.g., the order is for just 100 shares) followed by a cancel attempt for a small amount (e.g., 100 shares) where, by coincidence, the cancel fails because it was that user's 100 shares that just got taken. Thus, to deal with this possibility, our first sensitivity imposes that 100% of the depth at the race level is cleared, either through takes or cancels. As can be seen this reduces the number of races by about 13% (from 537 to 467), and reduces our measures of aggregate profits, latency arbitrage tax, and harm to liquidity by about 20%, depending on the measure. For completeness, we also include a sensitivity that requires that 50% of the depth at the race level is cleared.

For our definition of fail, the concern we mentioned in Section 4.2.3 is that we count limit orders that post to the book as a fail. A worry, especially at longer race horizons, is that we are picking up as "latency arbitrage races" cases where the "fail" is in fact simply a participant posting new liquidity at a new price, using a plain vanilla limit order, at a price that happened to be the price of the last successful trade. As a sensitivity, therefore, we only allow failed IOCs and failed cancels to count as fails.[46] That is, we do not allow ordinary limit orders to count as fails, even though some participants may in fact use them in latency arbitrage races, because of the fee advantage described earlier.

In the baseline, the strict fail criterion only reduces the number of races detected by about 8% (from 537 to 494), and race profits by about 5%. At longer horizons, as expected, the strict fails criterion reduces the number of races detected, and overall race profits, by larger amounts—for instance, at 3ms, the reduction in the number of races is about 15% (from 946 to 800) and the reduction in total profits is about 10% (from 255,000 per day to 232,000 per day). This makes sense because at longer horizons we should be more concerned about mistaking limit orders that post to

[46] Note as well that this sensitivity has the interpretation of only allowing as fails the "error messages"—failed IOCs and failed cancel attempts—that are unique to our message data relative to ordinary limit-order book data.

the book as failed race attempts. For this reason, when we consider what the sensitivity analyses suggest about upper bounds on race profits in the next section, when we use longer race horizons we will always do so in conjunction with the strict fail requirement.

### 6.5 Discussion of Sensitivity Analyses

Based on what we have learned from the various sensitivity analyses, Table 6.5 highlights several specific scenarios that we feel give a sense of the overall range of estimates for race profits and the effect on liquidity.

As Low scenarios, since we learned that race profits are especially sensitive to the choice of race horizon (Table 6.1) and to stricter requirements on the level of participation (Table 6.2), we highlight: 2+ within 50 microseconds, 2+ within 100 microseconds, 3+ within 100 microseconds, and 3+ within the information horizon.

As High scenarios, we highlight: 2+ within 1 millisecond, 2+ with 3 milliseconds, 3+ within 1 millisecond, and 3+ within 3 milliseconds. For each of these scenarios we also add the strict fails requirement, given the importance of this requirement at longer time horizons (as discussed around Table 6.4).

Over this set of scenarios, the latency arbitrage tax ranges from 0.15 to 0.74 basis points on the all-volume measure, and from 0.18 to 1.31 basis points on the non-race volume measure. The overall percentage harm to liquidity ranges from 5.1% to 41.6%.

We acknowledge that this exercise is somewhat subjective. At the lower end, we know conceptually that if we reduce the race horizon sufficiently and/or increase the participation requirements sufficiently we can find a lower bound that is essentially zero (e.g., 5+ within 50 microseconds yields very low numbers, see Appendix Table A.10). Similarly, at the high end, one could be more inclusive than seems reasonable (e.g., not imposing the strict fails requirement, or looking at horizons even longer than 3 milliseconds). Still, we think this exercise provides a useful sense for the range of magnitudes we find using our method. This range will inform our analysis in Section 7.

Table 6.4: **Sensitivity Analysis: Definitions of Success and Fail**

| | | Strict Success | | Strict Fail | | | | |
|---|---|---|---|---|---|---|---|---|
| **Measure** | Baseline | 100% | $\geq 50\%$ | InfoHor | 50$\mu$s | 500$\mu$s | 1ms | 3ms |
| **Frequency and Duration of Races** | | | | | | | | |
| Races per day | | | | | | | | |
| FTSE 100 - per symbol | 537.24 | 466.72 | 504.52 | 494.26 | 266.32 | 719.71 | 768.02 | 799.91 |
| FTSE 250 - per symbol | 70.05 | 62.22 | 66.38 | 65.95 | 38.08 | 105.09 | 115.94 | 123.01 |
| Mean race duration (microseconds) | 78.65 | 69.87 | 75.16 | 81.74 | 15.87 | 195.83 | 294.52 | 509.37 |
| % of races with wrong winner | 4.30 | 4.47 | 4.27 | 4.50 | 8.79 | 2.07 | 1.79 | 1.47 |
| % of volume in races | | | | | | | | |
| FTSE 100 | 22.15 | 18.82 | 21.49 | 20.64 | 8.83 | 35.33 | 40.54 | 44.27 |
| FTSE 250 | 16.90 | 15.07 | 16.52 | 16.26 | 7.87 | 32.16 | 36.79 | 40.17 |
| Full Sample | 21.46 | 18.32 | 20.84 | 20.06 | 8.70 | 34.89 | 40.03 | 43.72 |
| Mean number of messages within 500 $\mu$s | 3.46 | 3.45 | 3.48 | 3.51 | 3.54 | 3.48 | 3.15 | 2.94 |
| **Per-Race Profits** | | | | | | | | |
| Per-share profits | | | | | | | | |
| ticks | 0.55 | 0.55 | 0.55 | 0.54 | 0.52 | 0.52 | 0.54 | 0.55 |
| GBX | 0.17 | 0.17 | 0.17 | 0.17 | 0.16 | 0.16 | 0.16 | 0.16 |
| basis points | 1.66 | 1.68 | 1.66 | 1.64 | 1.64 | 1.59 | 1.64 | 1.64 |
| Per-race profits GBP | | | | | | | | |
| displayed depth | 1.85 | 1.66 | 1.81 | 1.89 | 1.57 | 1.91 | 2.04 | 2.09 |
| qty trade/cancel | 1.76 | 1.74 | 1.82 | 1.78 | 1.38 | 1.92 | 2.07 | 2.16 |
| **Aggregate Profits and LA Tax** | | | | | | | | |
| Daily Profits | | | | | | | | |
| FTSE 100 - per symbol | 1,047 | 800 | 953 | 985 | 437 | 1,437 | 1,650 | 1,783 |
| FTSE 250 - per symbol | 108 | 93 | 103 | 103 | 53 | 174 | 200 | 213 |
| Full Sample - aggregate | 132,378 | 103,745 | 121,493 | 124,904 | 57,048 | 187,989 | 215,794 | 232,457 |
| Latency Arbitrage Tax, All Volume (bps) | | | | | | | | |
| FTSE 100 | 0.38 | 0.30 | 0.35 | 0.36 | 0.16 | 0.53 | 0.61 | 0.65 |
| FTSE 250 | 0.66 | 0.57 | 0.63 | 0.63 | 0.32 | 1.07 | 1.23 | 1.31 |
| Full Sample | 0.42 | 0.33 | 0.39 | 0.40 | 0.18 | 0.60 | 0.68 | 0.74 |
| Latency Arbitrage Tax, Non-Race Volume (bps) | | | | | | | | |
| FTSE 100 | 0.49 | 0.38 | 0.45 | 0.46 | 0.18 | 0.82 | 1.02 | 1.18 |
| FTSE 250 | 0.80 | 0.69 | 0.76 | 0.76 | 0.35 | 1.59 | 1.95 | 2.20 |
| Full Sample | 0.53 | 0.42 | 0.49 | 0.50 | 0.20 | 0.92 | 1.14 | 1.31 |
| **Spread Decomposition** | | | | | | | | |
| Price impact in races / All price impact % | 30.58 | 26.10 | 29.78 | 28.92 | 11.70 | 47.38 | 55.27 | 61.61 |
| Price impact in races / Effective spread % | 32.82 | 28.02 | 31.97 | 31.04 | 12.56 | 50.84 | 59.31 | 66.11 |
| Loss avoidance / Effective spread % | 0.19 | 0.14 | 0.17 | 0.18 | 0.06 | 0.62 | 1.07 | 1.51 |
| **Implied Reduction in Cost of Liquidity** | | | | | | | | |
| % Reduction in liquidity cost | | | | | | | | |
| FTSE 100 - by symbol | 19.95 | 14.37 | 17.95 | 18.66 | 7.24 | 33.19 | 42.45 | 50.59 |
| FTSE 250 - by symbol | 11.93 | 10.12 | 11.28 | 11.42 | 5.64 | 23.12 | 27.56 | 29.90 |
| Full Sample - by date | 16.73 | 12.74 | 15.33 | 15.63 | 6.25 | 28.12 | 35.37 | 41.64 |

**Notes:** For descriptions of the sensitivity scenarios please see the text of Section 6.4. Descriptions of each of the items in this table can be found in the following table notes in Section 5. Races per day: Table 5.1. Mean race duration and % of races with wrong winner: Table 5.2. % of Volume in Races: Table 5.3. Mean number of messages: Table 5.4. Per-race profits: Table 5.6. Aggregate profits: Table 5.8. Latency Arbitrage Tax: Table 5.9. Spread decomposition: Table 5.10. Implied Reduction in Cost of Liquidity: Table 5.12.

Table 6.5: **Sensitivity Analysis: Selected Low and High Scenarios**

| | | Low Scenarios | | | | High Scenarios | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Measure** | **Baseline** | 2+, 50$\mu$s | 2+, 100$\mu$s | 3+, 100$\mu$s | 3+, IH | 2+, 1ms | 3+, 1ms | 2+, 3ms | 3+, 3ms |
| **Frequency and Duration of Races** | | | | | | | | | |
| Races per day | | | | | | | | | |
| FTSE 100 - per symbol | 537.24 | 296.66 | 388.58 | 134.38 | 228.98 | 768.02 | 544.87 | 799.91 | 609.01 |
| % of volume in races | | | | | | | | | |
| Full Sample | 21.46 | 9.77 | 13.32 | 6.17 | 12.30 | 40.03 | 33.20 | 43.72 | 38.14 |
| **Per-Race Profits** | | | | | | | | | |
| Per-share profits | | | | | | | | | |
| ticks | 0.55 | 0.54 | 0.53 | 0.71 | 0.71 | 0.54 | 0.62 | 0.55 | 0.64 |
| GBX | 0.17 | 0.16 | 0.16 | 0.22 | 0.23 | 0.16 | 0.19 | 0.16 | 0.19 |
| basis points | 1.66 | 1.68 | 1.63 | 2.29 | 2.24 | 1.64 | 1.92 | 1.64 | 1.92 |
| Per-race profits GBP | | | | | | | | | |
| displayed depth | 1.85 | 1.58 | 1.59 | 2.60 | 2.98 | 2.04 | 2.63 | 2.09 | 2.67 |
| qty trade/cancel | 1.76 | 1.38 | 1.44 | 2.40 | 2.87 | 2.07 | 2.69 | 2.16 | 2.76 |
| **Aggregate Profits and LA Tax** | | | | | | | | | |
| Daily Profits | | | | | | | | | |
| Full Sample - aggregate | 132,378 | 63,573 | 83,233 | 47,980 | 91,506 | 215,794 | 195,552 | 232,457 | 221,526 |
| Latency Arbitrage Tax, All Volume (bps) | | | | | | | | | |
| Full Sample | 0.42 | 0.20 | 0.26 | 0.15 | 0.29 | 0.68 | 0.62 | 0.74 | 0.70 |
| Latency Arbitrage Tax, Non-Race Volume (bps) | | | | | | | | | |
| Full Sample | 0.53 | 0.22 | 0.30 | 0.18 | 0.37 | 1.14 | 1.04 | 1.31 | 1.25 |
| **Spread Decomposition** | | | | | | | | | |
| Price impact in races / All price impact % | 30.58 | 12.84 | 17.89 | 9.34 | 19.13 | 55.27 | 47.01 | 61.61 | 55.54 |
| Price impact in races / Effective spread % | 32.82 | 13.77 | 19.19 | 10.03 | 20.54 | 59.31 | 50.45 | 66.11 | 59.61 |
| **Implied Reduction in Cost of Liquidity** | | | | | | | | | |
| % Reduction in liquidity cost | | | | | | | | | |
| FTSE 100 - by symbol | 19.95 | 7.98 | 10.97 | 5.94 | 12.46 | 42.45 | 34.63 | 50.59 | 44.21 |
| FTSE 250 - by symbol | 11.93 | 6.17 | 7.96 | 4.50 | 7.67 | 27.56 | 23.95 | 29.90 | 28.12 |
| Full Sample - by date | 16.73 | 6.96 | 9.49 | 5.13 | 10.43 | 35.37 | 28.83 | 41.64 | 36.20 |

**Notes:** For descriptions of the sensitivity scenarios please see the text of Section 6.5. Descriptions of each of the items in this table can be found in the following table notes in Section 5. Races per day: Table 5.1. % of Volume in Races: Table 5.3. Mean number of messages: Table 5.4. Per-race profits: Table 5.6. Aggregate profits: Table 5.8. Latency Arbitrage Tax: Table 5.9. Spread decomposition: Table 5.10. Implied Reduction in Cost of Liquidity: Table 5.12.

# 7 Total Sums at Stake

## 7.1 Extrapolation Models

Figure 5.5 in Section 5.4 showed visually that daily latency arbitrage profits are highly correlated to market volume and volatility, as expected given the theory. Table 7.1 presents these same relationships in regression form.

Columns (1-2) regress daily in-sample latency arbitrage profits on daily LSE regular-hours trading volume in GBP (10,000s). The coefficient of 0.421 in (2) is directly interpretable as the all-volume latency arbitrage tax in basis points. Including a constant term changes the coefficient only slightly, to 0.432. This single variable has an $R^2$ of 0.81.

Columns (3-4) regress daily in-sample latency arbitrage profits on daily realized 1-minute volatility.[47] To make the results interpretable in units of latency arbitrage tax, realized volatility in percentage points is multiplied by the sample-average of daily trading volume.[48] Here, including the constant term does provide a meaningfully better fit, which can also be seen visually in the scatterplot in Figure 5.5, Panel B. The coefficient of 0.023 in (3) means that every additional percentage point of realized volatility adds 0.023 basis points to that day's latency arbitrage tax. This variable has lower explanatory power than volume, but still high, with an $R^2$ of 0.661.

Columns (5-6) present results for a two-variable model in which daily latency arbitrage profits are regressed on both trading volume and realized volatility. Again, to make the results interpretable, realized volatility is multiplied by average daily trading volume.[49] Both variables are significant, and the two-variable model has higher explanatory power than the single-variable model, but the difference is modest, with an $R^2$ of 0.83 versus 0.81. The reason for this is that volume and volatility are highly correlated to each other, with an in-sample correlation of 0.82 in our data. The coefficients can be interpreted as follows. On a day with average 1-minute volatility (about 13% in our sample), the latency-arbitrage tax is 0.3354+13*0.0066=0.42 basis points, the overall sample average. On a particularly high realized volatility day, say 25%, the latency arbitrage tax would be 0.50 basis points. On a relatively calm day, say 10% realized volatility, the latency arbitrage tax would be 0.40 basis points.

Before we turn to out-of-sample extrapolation, we emphasize that the standard errors on these coefficients are much smaller than the variation in the latency-arbitrage tax we found in Section 6 when we considered different specifications for race detection. Therefore, we will emphasize two

[47] In the appendix we report regression results for 5-minute volatility and for a measure of volatility emphasized in BCS called distance traveled. 5-minute volatility has lower explanatory power than 1-minute volatility. Distance traveled actually has greater explanatory power than 1-minute volatility, but we emphasize the latter because it is more easily measurable across markets and over time, and more widely utilized in practice and in the literature.

[48] That is, we regress $\text{LatencyArbProfits}_t = \alpha + \beta(\sigma_t \cdot \text{AvgDailyVolume})$ where $\sigma_t$ is in percentage points and AvgDailyVolume is in GBP 10,000s.

[49] That is, we regress $\text{LatencyArbProfits}_t = \alpha + \beta\text{Volume}_t + \gamma(\sigma_t \cdot \text{AvgDailyVolume})$. We also considered the specification $\text{LATax}_t = \alpha + \beta \cdot \frac{\text{Volume}_t - \text{AvgDailyVolume}}{\text{AvgDailyVolume}} + \gamma\sigma_t$, that is, the latency arbitrage tax in basis points is the LHS variable. In this specification, the coefficient on volatility is roughly the same as in Column 6, at 0.0061, and the coefficient on volume is -0.0008 and statistically insignificant. These coefficients imply that on a day where trading volume is 10 percentage points higher than the average, holding volatility fixed, the latency arbitrage tax is -0.008 basis points lower than average.

Table 7.1: **Extrapolation Models**

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Latency Arbitrage Profits (GBP) | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Volume (10,000 GBP) | 0.4319*** | 0.4213*** | | | 0.3405*** | 0.3354*** |
| | (0.0326) | (0.0082) | | | (0.0544) | (0.0415) |
| Volatility (1 min) * Average Volume | | | 0.0228*** | 0.0313*** | 0.0065** | 0.0066** |
| | | | (0.0025) | (0.0009) | (0.0032) | (0.0031) |
| Constant | −3,562 | | 39,226*** | | −1,532 | |
| | (10,611) | | (11,032) | | (10,263) | |
| Observations | 43 | 43 | 43 | 43 | 43 | 43 |
| $R^2$ | 0.811 | 0.810 | 0.661 | 0.567 | 0.829 | 0.829 |

*p<0.1; **p<0.05; ***p<0.01

**Notes:** The dependent variable in all regressions is daily race profits in GBP, for the full sample, as described in Table 5.8. Volume is daily regular-hours LSE trading volume in GBP, as first described in Table 5.3, in units of 10,000 GBP so that the coefficient is interpretable as a latency-arbitrage tax in basis points. Volatility is realized 1-minute volatility for the FTSE 350 index in percentage points, using TRTH data, as described in Figure 5.5. Volatility in percentage points is multiplied by average daily volume in 10,000 GBP so that the coefficient has the interpretation of the effect of a 1 percentage point change in volatility on the latency arbitrage tax in basis points. Regressions are ordinary least squares. $R^2$ in the regressions without constant terms is computed according to the formula $1 - \text{Var}(\hat{e})/\text{Var}(y)$. P-values are computed using the student-t distribution.

kinds of out-of-sample results: (i) results based on the volume and volatility model presented in Column (6); and (ii) results based on the volume-only model in column (2), which is economically equivalent to a constant latency arbitrage tax model, using both the baseline latency arbitrage tax and the range of latency arbitrage taxes across the various sensitivity analyses discussed in Section 6.5.

## 7.2 Out-of-Sample Extrapolation: UK Equity Markets

Table 7.2 presents our estimates of the annual sums at stake in latency arbitrage races in the UK for the five year period 2014-2018. In Column (1) we present the estimate based on the volume and volatility regression model, i.e., column (6) of Table 7.2. For volume data we use LSE reports of their daily trading volume and monthly regular-hours market share to estimate total daily regular-hours trading volume. For volatility data, we compute daily one-minute realized volatility of the FTSE 350 index using Thomson Reuters data. In Column (2) we present the estimate based on the volume-only model, i.e., based on the latency-arbitrage tax of 0.42 basis points. In Columns (3)-(4) we present the range of estimates implied by the sensitivity analyses discussed in Section 6.5; these are based on latency-arbitrage taxes of 0.15 basis points in the lowest of the Low scenarios and 0.74 basis points in the highest of the High scenarios.

The volume-and-volatility model implies annual latency arbitrage profits in UK equity markets ranging between GBP 51.0 Million to GBP 63.3 Million per year. The volume-only model yields slightly higher estimates. At the low end of our sensitivity analyses the annual profits are about

Table 7.2: **Annual Latency Arbitrage Profits in UK Equity Markets (GBP Millions)**

| Year | (1) Volume-Volatility | (2) Volume-Only | (3) Low Scenario | (4) High Scenario |
|---|---|---|---|---|
| 2014 | 52.0 | 56.7 | 20.5 | 99.1 |
| 2015 | 58.9 | 61.6 | 22.3 | 107.7 |
| 2016 | 63.3 | 63.8 | 23.0 | 111.4 |
| 2017 | 51.0 | 57.5 | 20.8 | 100.4 |
| 2018 | 55.8 | 60.6 | 21.9 | 105.9 |

**Note:** We compute UK regular-hours trading volume by dividing LSE's monthly reported regular-hours trading volume by LSE's monthly reported regular-hours market share. We compute UK 1-minute realized volatility using TRTH data for the FTSE 350 index, computing the realized volatility on each day and then computing the root mean square. Model (1) uses the coefficients from Regression (6) in Table 7.1. Model (2) uses the coefficient from Regression (2) in Table 7.1. Model (3) and Model (4) use the min and max latency-arbitrage taxes found in Table 6.5, of 0.15 bps and 0.74 bps, respectively.

GBP 20 million and at the high end the annual profits are about GBP 100 million.

## 7.3 Out-of-Sample Extrapolation: Global Equity Markets

This section presents estimates of the annual sums at stake in latency arbitrage races in global equities markets. The goal is to get a sense of magnitudes for what our results using the LSE message data imply about the overall global size of the latency-arbitrage prize. Please note that this extrapolation does not attempt to account for differences in equity market structure across countries that may affect the level of latency arbitrage (e.g., the level of fragmentation, role of ETFs, geography), nor does it include other asset classes besides equities. As we will further emphasize in the conclusion, we hope that other researchers in the future will use message data from other countries and additional asset classes to produce better numbers.

We use volume data from the World Federation of Exchanges (2018). The advantage of WFE data is that it covers nearly all exchange groups around the world, but a caveat is that there may be some inconsistencies in how exchange groups report their data to the WFE. We consulted with the WFE to obtain their advice regarding how best to utilize their data. Unfortunately, exchange groups appear to be inconsistent about whether they include volume from opening and closing auctions, which ideally we would exclude. In the other direction, this data does not include electronic off-exchange trading volume (i.e., dark pools) that is vulnerable to latency arbitrage, and which is a significant share of equities trading volume in many countries. We compute volatility based on the one-minute realized volatility of regional equity market indices using Thomson Reuters data. As in Table 7.2 above, Table 7.3 Column (1) presents estimates based on the volume and volatility regression model, Column (2) presents estimates based on the volume-only model, and Columns (3)-(4) present the range implied by the sensitivity analyses.

Our main estimate of a latency arbitrage tax of 0.42 basis points implies annual latency arbitrage profits of $4.8 billion for global equities markets. The volume-and-volatility model yields a slightly lower estimate since volatility was lower in 2018 than in our sample period. At the low end of our

Table 7.3: **Annual Latency Arbitrage Profits in Global Equity Markets in 2018 (USD Millions)**

| Exchange Group | (1) Volume-Volatility | (2) Volume-Only | (3) Low Scenario | (4) High Scenario |
|---|---|---|---|---|
| NYSE Group | 1,006 | 1,023 | 370 | 1,787 |
| BATS Global Markets - U.S. | 895 | 910 | 329 | 1,590 |
| Nasdaq - U.S. | 847 | 862 | 311 | 1,505 |
| Shenzhen Stock Exchange | 327 | 336 | 122 | 588 |
| Japan Exchange Group | 281 | 286 | 103 | 500 |
| Shanghai Stock Exchange | 260 | 268 | 97 | 468 |
| Korea Exchange | 118 | 120 | 43 | 209 |
| London Stock Exchange Group** | 109 | 119 | 43 | 207 |
| BATS Chi-X Europe | 110 | 119 | 43 | 207 |
| Hong Kong Exchanges and Clearing | 102 | 104 | 38 | 182 |
| Euronext | 89 | 96 | 35 | 168 |
| Deutsche Börse Group | 78 | 85 | 31 | 148 |
| TMX Group | 56 | 61 | 22 | 107 |
| National Stock Exchange of India | 47 | 49 | 18 | 86 |
| SIX Swiss Exchange | 40 | 43 | 16 | 76 |
| Global Total (WFE Data Universe) | 4,674 | 4,799 | 1,734 | 8,383 |

**London Stock Exchange Group includes London Stock Exchange as well as Borsa Italiana

**Note:** Trading volume is from the World Federation of Exchanges (2018). Per guidance from the WFE, we sum the volume of listed symbols and exchange traded funds traded on electronic order books ("EOB Value of Share Trading" and "ETFs EOB Turnover"). Please note that there may be inconsistencies across exchanges in how they report data to WFE. The data is comprehensive and helps give a sense of the overall global magnitudes but for any particular exchange better volume data may be available. Volatility is computed using TRTH data for the following indices. NYSE, BATS and Nasdaq: S&P 500. Shenzhen and Shanghai: Shanghai composite. Japan: Nikkei225. Korea: KOSPI. LSE Group: FTSE 350. BATS Chi-X, Euronext, Deutsche Börse, Swiss: EuroStoxx600. Hong Kong: Hang Seng. India: SENSEX. Canada TMX Group: TSX Composite. The row denoted Global Total (WFE Data Universe) includes all exchange groups in the WFE data. All estimates reported in the table are computed analogously to Table 7.2 with the exception of the global total in Column (1): since we do not have volatility indices for all exchange around the world, we compute this as (Sum of Volume-and-Volatility Model Profits for Top 15 Exchange Groups) / (Sum of Volume-Only Model Profits for Top 15 Exchange Groups) * (Global Total Profits Based on Volume-Only Model).

sensitivity analyses the annual latency arbitrage profits for global equity markets are about $1.7 billion, and at the high end the annual profits are about $8.4 billion.

# 8 Conclusion

We conclude by summarizing the paper's contributions to the academic literature and discussing our hopes for future work.

The paper's first contribution is methodological: utilizing exchange message data to measure latency arbitrage. The central insight of the method is simple: an important part of the activity that theory implies should occur in a latency-arbitrage race will not actually manifest in traditional limit order book data—the *losers* of the race. To see the full picture of a latency-arbitrage race requires seeing the full message traffic to and from the exchange, including the exchange error messages sent to losers of the race (specifically, failed IOCs and failed cancels). Armed with this simple insight and the correct data, it was conceptually straightforward, albeit human-time and computer-time

intensive, to develop and implement the empirical method described in Section 4.[50]

The paper's second—and we think main—contribution is the set of empirical facts we document about latency arbitrage in Section 5. We show that races are very frequent and very fast, with an average of 537 races per day for FTSE 100 stocks, lasting an average of just 81 microseconds, and with a mode of just 5-10 microseconds, or less than 1/10000th of the time it takes to blink your eye. Over 20% of trading volume takes place in races. A small number of firms win the large majority of races, disproportionately as takers of liquidity. Most races are for very small amounts of money, averaging just over half a tick and just under GBP 2. But, because of the large volume, these small amounts add up. The "latency arbitrage tax," defined as latency arbitrage profits divided by trading volume, is 0.42 basis points based on all trading volume, and 0.53 basis points based on all non-race volume. This amounts to about GBP 60 million annually in the UK. Extrapolating from our UK data, our estimates imply that latency arbitrage is worth on the order of $5 billion annually in global equity markets.

A third contribution, narrower and more technical in nature but we hope useful to the microstructure literature, is the development of two new approaches to quantifying latency arbitrage as a proportion of the overall cost of liquidity. These new methods, used in conjunction with the results described above, show that latency arbitrage accounts for 33% of the effective spread, 31% of all price impact, and that eliminating latency arbitrage would reduce the cost of liquidity for investors by 17%.

One natural direction for future research is to utilize this paper's method for detecting latency-arbitrage races to then try to better understand their sources. One could imagine, for instance, trying to quantify what proportion of latency arbitrage races involve public signals from the same symbol traded on a different venue, what proportion involve a change in a correlated market index, what proportion involve signals from different asset classes or geographies, etc.

Our main hope for future research, however, is simply that other researchers and regulatory authorities replicate our analysis for markets beyond UK equities. Of particular interest would be markets like US equities that are more fragmented than the UK; and assets such as ETFs, futures and currencies that have lots of mechanical arbitrage relationships with other highly-correlated assets. The "hard" part of such a study is obtaining the message data. Once one has the message data, applying the method we have developed in this paper is relatively straightforward.[51] To our knowledge, most regulators do not currently capture message data from exchanges, and exchanges

[50]The final run of our code, including all sensitivity analyses, required about 24 days of computer time on a 128-core AWS server (about 60 hours for data preparation and the baseline analysis, plus an additional 35 hours per sensitivity analysis). From initial receipt of data to first completed draft, the paper required about 3 years of work. The main reason the project has been time intensive, despite its conceptual simplicity, is that message data had never been used before for research (neither academic research nor, we think, industry research) and it took a lot of false starts and iterations to fully understand. This work presumably is evident from Sections 2 and 3. We expect that future research using message data will be a lot more efficient than our study. First, our study can be used as a blueprint. Second, some code re-optimization we are including in the code that will be disseminated publicly reduces the computational run time by about 75%.

[51]To this end, our codebase and a user guide will be made publicly available upon publication of this paper. Regulators and researchers interested in obtaining this codebase and user guide prior to publication should contact the authors.

seem to preserve message data somewhat inconsistently. We hope this will change. Limit order book data has historically been viewed as the official record of what happened in the market, but we argue that the message data, and especially the "error messages" that indicate that a particular participant has failed in their request, are key to understanding speed-sensitive trading.

## References


**Aquilina, Matteo, Sean Foley, Peter O'Neill, and Thomas Ruf.** 2016. "Asymmetries in Dark Pool Reference Prices." FCA Occasional Paper 21.

**Baker, Nick, and Bryan Gruley.** 2019. "The Gazillion-Dollar Standoff over Two High-Frequency Trading Towers." *Bloomberg Businessweek*. March 8. Retrieved from `https://www.bloomberg.com/news/features/2019-03-08/the-gazillion-dollar-standoff-over-two-high-frequency-trading-towers`.

**Baldauf, Markus, and Joshua Mollner.** 2020. "High-Frequency Trading and Market Performance." *The Journal of Finance*, 75(3): 1495–1526.

**Baron, Matthew, Jonathan Brogaard, Björn Hagströmer, and Andrei Kirilenko.** 2019. "Risk and Return in High-Frequency Trading." *Journal of Financial and Quantitative Analysis*, 54(3): 993–1024.

**Battalio, Robert, Shane A. Corwin, and Robert Jennings.** 2016. "Can Brokers Have It All? On the Relation Between Make-Take Fees and Limit Order Execution Quality." *The Journal of Finance*, 71(5): 2193–2238.

**Benos, Evangelos, and Satchit Sagade.** 2016. "Price Discovery and the Cross-Section of High-Frequency Trading." *Journal of Financial Markets*, 30: 54–77.

**Biais, Bruno, and Thierry Foucault.** 2014. "HFT and Market Quality." *Bankers, Markets & Investors*, 128(1): 5–19.

**Biais, Bruno, Thierry Foucault, and Sophie Moinas.** 2015. "Equilibrium Fast Trading." *Journal of Financial Economics*, 116(2): 292–313.

**Breckenfelder, Johannes.** 2019. "Competition Among High-Frequency Traders, and Market Quality." *SSRN*. Available from SSRN: `https://ssrn.com/abstract=3402867`.

**Brogaard, Jonathan, Allen Carrion, Thibaut Moyaert, Ryan Riordan, Andriy Shkilko, and Konstantin Sokolov.** 2018. "High Frequency Trading and Extreme Price Movements." *Journal of Financial Economics*, 128(2): 253–265.

**Brogaard, Jonathan, Björn Hagströmer, Lars Nordén, and Ryan Riordan.** 2015. "Trading Fast and Slow: Colocation and Liquidity." *The Review of Financial Studies*, 28(12): 3407–3443.

**Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan.** 2014. "High-Frequency Trading and Price Discovery." *The Review of Financial Studies*, 27(8): 2267–2306.

**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(4): 1547–1621.

**Budish, Eric, Robin S. Lee, and John J. Shim.** 2019. "A Theory of Stock Exchange Competition and Innovation: Will the Market Fix the Market?" National Bureau of Economic Research. NBER Working Paper No. 25855.

**Carrion, Allen.** 2013. "Very Fast Money: High-Frequency Trading on the NASDAQ." *Journal of Financial Markets*, 16(4): 680–711.

**Cboe EDGA.** 2019. "Notice of Filing of a Proposed Rule Change to Introduce a Liquidity Provider Protection on EDGA." Release No 34-86168; File No. SR-CboeEDGA-2019-012. Retrieved from `https://www.sec.gov/rules/sro/cboeedga/2019/34-86168.pdf`.

**Chicago Stock Exchange.** 2016. "Notice of Filing of Proposed Rule Change to Adopt the CHX Liquidity Taking Access Delay." Release No. 34-78860; File No. SR-CHX-2016-16. Retrieved from `https://www.sec.gov/rules/sro/chx/2016/34-78860.pdf`.

**CME Group, Inc.** 2019. "Hibernia Networks." Available from `https://www.cmegroup.com/partner-services/hibernia-networks.html`.

**Commodity Futures Trading Commission.** 2015. "Concept Release on Risk Controls and System Safeguards for Automated Trading Environments." 78 FR 56541. Retrieved from `https://www.federalregister.gov/documents/2013/09/12/2013-22185/concept-release-on-risk-controls-and-system-safeguards-for-automated-trading-environments`.

**Conrad, Jennifer, and Sunil Wahal.** 2019. "The Term Structure of Liquidity Provision." *Journal of Financial Economics*.

**Deutsche Börse Group.** 2018. "Insights into Trading System Dynamics." Retrieved August 8, 2018 from `http://web.archive.org/web/20180806172740/https://www.eurexchange.com/blob/238346/6d353d9701d70b82cd1a6281b3bf2595/data/presentation_insights-into-trading-system-dynamics_en.pdf`.

**Dewhurst, David Rushing, Colin M. Van Oort, IV Ring, H. John, Tyler J. Gray, Christopher M. Danforth, and Brian F. Tivnan.** 2019. "Scaling of Inefficiencies in the US Equity Markets: Evidence from Three Market Indices and More than 2900 Securities." *arXiv Preprint arXiv:1902.04691*.

**Ding, Shengwei, John Hanna, and Terrence Hendershott.** 2014. "How Slow is the NBBO? A Comparison with Direct Exchange Feeds." *Financial Review*, 49(2): 313–332.

**Du, Songzi, and Haoxiang Zhu.** 2017. "What is the Optimal Trading Frequency in Financial Markets?" *The Review of Economic Studies*, 84(4): 1606–1651.

**European Securities Market Authority.** 2014. "High-Frequency Trading Activity in EU Equity Markets." Retrieved from `https://www.esma.europa.eu/system/files_force/library/2015/11/esma20141_-_hft_activity_in_eu_equity_markets.pdf`.

**Financial Conduct Authority.** 2018. "Algorithmic Trading Compliance in Wholesale Markets." Retrieved from https://www.fca.org.uk/news/press-releases/fca-publishes-report-supervision-algorithmic-trading.

**Foucault, Thierry, Roman Kozhan, and Wing Wah Tham.** 2016. "Toxic Arbitrage." *The Review of Financial Studies*, 30(4): 1053–1094.

**Glosten, Lawrence R.** 1987. "Components of the Bid-Ask Spread and the Statistical Properties of Transaction Prices." *The Journal of Finance*, 42(5): 1293–1307.

**Glosten, Lawrence R., and Lawrence E. Harris.** 1988. "Estimating the Components of the Bid/Ask Spread." *Journal of Financial Economics*, 21(1): 123–142.

**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(1): 71–100.

**Hasbrouck, Joel.** 1991*a*. "Measuring the Information Content of Stock Trades." *The Journal of Finance*, 46(1): 179–207.

**Hasbrouck, Joel.** 1991*b*. "The Summary Informativeness of Stock Trades: An Econometric Analysis." *The Review of Financial Studies*, 4(3): 571–595.

**Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld.** 2011. "Does Algorithmic Trading Improve Liquidity?" *The Journal of Finance*, 66(1): 1–33.

**Hoffmann, Peter.** 2014. "A Dynamic Limit Order Market with Fast and Slow Traders." *Journal of Financial Economics*, 113(1): 156–169.

**ICE Futures.** 2019. "Re: Amendments to Rule 4.26 Order Execution (New Passive Order Protection Functionality) Submission Pursuant to Section 5c(c)(1) of the Act and Regulation 40.6(a)." Retrieved from https://www.cftc.gov/sites/default/files/2019-02/ICEFuturessPassiveOrder020119.pdf.

**Investors' Exchange.** 2015. "Form 1 Application for Registration as a National Securities Exchange Pursuant to Section 6 of the Securities Exchange Act of 1934." Release No. 34-75925; File No. 10-222. Retrieved from https://www.sec.gov/rules/other/2015/investors-exchange-form-1.htm.

**Investors' Exchange.** 2019. "The Cost of Exchange Services." Retrieved from https://iextrading.com/docs/The\%20Cost\%20of\%20Exchange\%20Services.pdf.

**Joint Staff Report.** 2015. "Joint Staff Report: The U.S. Treasury Market on October 15, 2014." *U.S. Department of the Treasury, Board of Governors of the Federal Reserve System, Federal*

*Reserve Bank of New York, U.S. Securities and Exchange Commission, U.S. Commodity Futures Trading Commission*, Retrieved from https://www.treasury.gov/press-center/press-releases/Documents/Joint_Staff_Report_Treasury_10-15-2015.pdf.

**Jones, Charles M.** 2013. "What do We Know About High-Frequency Trading?" *Columbia Business School Research Paper*, , (13-11).

**Jump Trading International Limited.** 2018. "Report and Financial Statements for the Year Ended 31 December 2017." Retrieved from https://beta.companieshouse.gov.uk/company/05976015/filing-history/MzIxNTgxODgyN2FkaXF6a2N4/document?format=pdf&download=0.

**Korajczyk, Robert A, and Dermot Murphy.** 2019. "High-Frequency Market Making to Large Institutional Trades." *The Review of Financial Studies*, 32(3): 1034–1067.

**Kyle, Albert S.** 1985. "Continuous Auctions and Insider Trading." *Econometrica*, 1315–1335.

**Laughlin, Gregory, Anthony Aguirre, and Joseph Grundfest.** 2014. "Information Transmission Between Financial Markets in Chicago and New York." *Financial Review*, 49(2): 283–312.

**Laumonier, Alexandre.** 2014. "HFT in My Backyard - I." September 22. Blog Post. Retrieved from https://sniperinmahwah.wordpress.com/2014/09/22/hft-in-my-backyard-part-i/.

**Laumonier, Alexandre.** 2019. *4.* Zones Sensibles.

**Lewis, Michael.** 2014. *Flash Boys: A Wall Street Revolt.* New York, NY:W. W. Norton & Company.

**Li, Sida, Xin Wang, and Mao Ye.** 2020. "Who Provides Liquidity and When?" *Journal of Financial Economics, Forthcoming*. Available from SSRN: https://ssrn.com/abstract=2902984orhttp://dx.doi.org/10.2139/ssrn.2902984.

**Lockwood, John W, Adwait Gupte, Nishit Mehta, Michaela Blott, Tom English, and Kees Vissers.** 2012. "A Low-Latency Library in FPGA Hardware for High-Frequency Trading (HFT)." 9–16, IEEE.

**London Metals Exchange.** 2019. "Technical Change to LMEselect FIX Message Processing for the LMEprecious Market to Introduce a Fixed Minimum Delay." Retrieved from https://www.lme.com/-/media/Files/News/Notices/2019/05/19-165-Technical-change-to-LMEselect-FIX-message-processing-for-the-LMEprecious-market-to-introduce-fixed-minimum-delay.pdf.

**London Stock Exchange Group.** 2015*a*. "MIT1001 - Connectivity Guide, Issue 2.3." London Stock Exchange Group, Retrieved July 3, 2015 from https://web.archive.org/web/20150703141903/http://www.londonstockexchange.com/products-and-services/millennium-exchange/millennium-exchange-migration/londonstockexchangeconnectivityguidev6.pdf.

**London Stock Exchange Group.** 2015*b*. "MIT201 - Guide to the Trading System, Issue 12.4." London Stock Exchange Group, Retrieved July 3, 2015 from https://web.archive.org/web/20150703141903/https://www.londonstockexchange.com/products-and-services/trading-services/guide-to-new-trading-system.pdf.

**London Stock Exchange Group.** 2015*c*. "MIT202 - FIX Trading Gateway (FIX5.0), Issue 11.3." London Stock Exchange Group, Retrieved July 3, 2015 from https://web.archive.org/web/20150703141903/https://www.londonstockexchange.com/products-and-services/millennium-exchange/millennium-exchange-migration/mit202issuev11-1new.pdf.

**London Stock Exchange Group.** 2015*d*. "MIT203 - Native Trading Gateway, Issue 11.4." London Stock Exchange Group, Retrieved July 3, 2015 from https://web.archive.org/web/20150703141903/http://www.londonstockexchange.com/products-and-services/millennium-exchange/millennium-exchange-migration/mit203issuev11-1.pdf.

**London Stock Exchange Group.** 2015*e*. "MIT801 - Reject Codes, Issue 10." London Stock Exchange Group, Retrieved July 3, 2015 from https://web.archive.org/web/20150703141903/http://www.londonstockexchange.com/products-and-services/millennium-exchange/millennium-exchange-migration/mit801-rejectcodes091213.xls.

**London Stock Exchange Group.** 2015*f*. "Trading Services Price List (On-Exchange and OTC)." London Stock Exchange Group, Retrieved August 1, 2015 from https://web.archive.org/web/20170308142624/http://www.lseg.com/sites/default/files/content/documents/Trading%20Services%20Price%20List%2020150801.pdf.

**MacKenzie, Donald.** 2019. "How Fragile is Competition in High-Frequency Trading." March 26. Retrieved from https://tabbforum.com/opinions/how-fragile-is-competition-in-high-frequency-trading/.

**Malinova, Katya, Andreas Park, and Ryan Riordan.** 2018. "Do Retail Traders Suffer from High Frequency Traders?" *SSRN*. Available from SSRN: https://ssrn.com/abstract=2183806.

**Menkveld, Albert J.** 2013. "High Frequency Trading and the New Market Makers." *Journal of financial Markets*, 16(4): 712–740.

**Menkveld, Albert J.** 2016. "The Economics of High-Frequency Trading: Taking Stock." *Annual Review of Financial Economics*, 8: 1–24.

**Michaels, Dave.** 2016. "Chicago Stock Exchange Targets Rapid-Fire Traders With Speed Bump, Echoing IEX." *The Wall Street Journal*. August 30. Retrieved from https://www.wsj.com/articles/chicago-stock-exchange-targets-rapid-fire-traders-with-speed-bump-echoing-iex-1472591832.

**Mulholland, Rory.** 2015. "Flashboys Return: the Transatlantic War for Milliseconds." *The Irish Times*. October 3. Retrieved from https://www.irishtimes.com/news/environment/flashboys-return-the-transatlantic-war-for-milliseconds-1.2376441.

**Narang, Manoj.** 2014. "A Much-Needed HFT Primer for 'Flash Boys' Author Michael Lewis." *Institutional Investor*. April 7. Retrieved from https://www.institutionalinvestor.com/article/b14zbj2trgncsl/a-much-needed-hft-primer-for-flash-boys-author-michael-lewis.

**New York Attorney General's Office.** 2014. "Remarks on High-Frequency Trading & Insider Trading 2.0." New York Law School Panel on "Insider Trading 2.0 - A New Initiative to Crack Down on Predatory Practices". Retrieved from https://ag.ny.gov/pdfs/HFT_and_market_structure.pdf.

**NYSE Group.** 2018. "Technology FAQ and Best Practices: Equities." Retrieved Feb 22, 2019 from https://www.nyse.com/publicdocs/nyse/markets/nyse/NYSE_Group_Equities_Technology_FAQ.pdf.

**O'Hara, Maureen.** 2015. "High Frequency Market Microstructure." *Journal of Financial Economics*, 116(2): 257–270.

**Osipovich, Alexander.** 2018. "High-Speed Traders Profit From Return of Loophole at CME." *The Wall Street Journal*. Feb 12. Retrieved from https://www.wsj.com/articles/glitch-exploited-by-high-speed-traders-is-back-at-cme-1518431401.

**Osipovich, Alexander.** 2020. "Ultrafast Trading Costs Stock Investors Nearly $5 Billion a Year, Study Says." *The Wall Street Journal*. Jan 27. Retrieved from https://www.wsj.com/articles/ultrafast-trading-costs-stock-investors-nearly-5-billion-a-year-study-says-11580126036.

**Pagnotta, Emiliano S., and Thomas Philippon.** 2018. "Competing on Speed." *Econometrica*, 86(3): 1067–1115.

**Patterson, Scott, Jenny Strasburg, and Liam Pleven.** 2013. "High-Speed Traders Exploit Loophole." *The Wall Street Journal*. May 1. Retrieved from https://www.wsj.com/articles/SB10001424127887323798104578455032466082920.

**Securities and Exchange Commission.** 2010. "Concept Release on Equity Market Structure." Release No. 34-61358; File No. S7-02-10. 75 FR 3594, 3606. January 21. Retrieved from https://www.sec.gov/rules/concept/2010/34-61358.pdf.

**Shkilko, Andriy, and Konstantin Sokolov.** 2016. "Every Cloud Has a Silver Lining: Fast Trading, Microwave Connectivity and Trading Costs." *Journal of Finance, Forthcoming*. Available from SSRN: https://ssrn.com/abstract=2848562.

**Stoll, Hans R.** 1989. "Inferring the Components of the Bid-Ask Spread: Theory and Empirical Tests." *the Journal of Finance*, 44(1): 115–134.

**Tabb, Larry.** 2014. "'Flash Boys: Not So Fast' - A Review." *Tabb Forum*. December 17. Retrieved from https://tabbforum.com/opinions/flash-boys-not-so-fast-a-review/.

**Van Kervel, Vincent, and Albert J. Menkveld.** 2019. "High-Frequency Trading Around Large Institutional Orders." *The Journal of Finance*, 74(3): 1091–1137.

**Virtu Financial, Inc.** 2019*a*. "Fiscal Year 2018 10-K." Retrieved from http://ir.virtu.com/financials-and-filings/sec-filings/sec-filings-details/default.aspx?FilingId=13270405.

**Virtu Financial, Inc.** 2019*b*. "Re: NYSE Mahwah Roof." Retrieved from https://www.sec.gov/comments/4-729/4729-5880550-188760.pdf.

**Wah, Elaine.** 2016. "How Prevalent and Profitable are Latency Arbitrage Opportunities on US Stock Exchanges?" *SSRN*. Available from SSRN: https://ssrn.com/abstract=2729109.

**Weller, Brian M.** 2018. "Does Algorithmic Trading Reduce Information Acquisition?" *The Review of Financial Studies*, 31(6): 2184–2226.

**World Federation of Exchanges.** 2018. "World Federation of Exchanges Database." Available from https://www.world-exchanges.org/our-work/statistics.

# A Additional Results (Not for Publication)

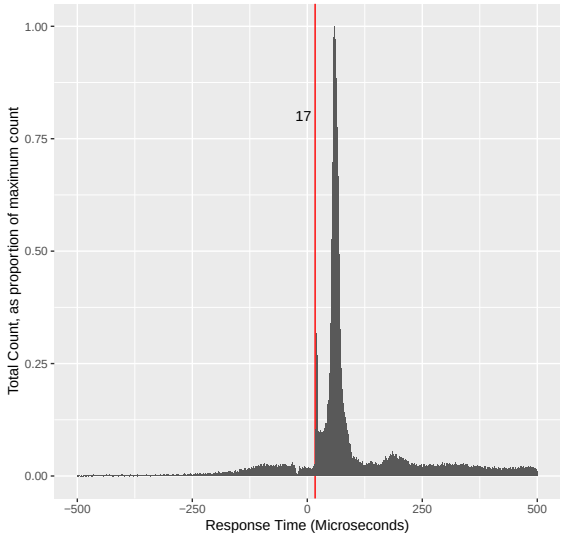
This online appendix contains additional results that are primarily alternate specifications of tables or figures reported in the main text. The results are presented in sequential order based on the location of the corresponding table or figure in the main text.

### Additional Results Related to Computing the Information Horizon (Section 4.3)

Figure 4.1 in the main text reports the distribution of time between observed M1-M2 message pairs where M1 is an outbound message reporting a new limit order that has been added to the book, and M2 is an inbound message that is aggressive at the price level associated with M1. We use the spike in this distribution, at 29 microseconds, as an input into our computation of the information horizon.

The following figure reports an analogous analysis but with M1-M2 message pairs where M1 is an outbound message reporting that an existing limit order has been partially filled, and M2 is an inbound message that cancels the remainder of the limit order. The difference versus Figure 4.1 in the text is that in Figure 4.1 the response message M2 is sent by a different participant from M1, whereas in this figure, the participant who received M1 outbound then is the same participant who send M2 inbound. Thus, the difference in response times between this figure and Figure 4.1 reflects the difference in speed between reactions to a publicly disseminated book update, versus reactions to a privately received trade update. The former is more appropriate for computing the information horizon, but the latter may also be of interest and is reported here:

Figure A.1: **Distribution of Time between M1 Outbound partial fill → M2 Inbound Successful Cancel**




**Notes:** Over all regular-hour messages from four high-volume symbols, BP, GLEN, HSBA, VOD, we obtain all cases where some outbound message is a partial fill and a subsequent outbound message is a successful cancel. We then obtain the inbound cancel request message associated with the outbound cancel success message, and compute the difference in the message timestamp between the partial fill outbound message (M1) and the cancel request inbound message (M2). Note that this difference can be negative if M2's inbound is sent by the participant before M1's outbound is sent by the outbound gateway. The distribution depicted is a microsecond-binned histogram truncated at -500 microseconds and +500 microseconds. As described in the text of Section 4.3, we compute the start of the spike by computing the mean and standard deviation of the distribution in the period -100 microseconds to 0 microseconds, and then finding the first microsecond after 0 that is at least 5 standard deviations above this pre-0 mean.

## Symbol-Date Version of Table 5.1

Table 5.1 in the main text reports the number of races per day at the symbol level averaged across all dates (Panel A), and at the date level summed across all symbols (Panel B). The following table presents the number of races at the symbol-date level, i.e., without aggregating across either symbols or dates.

Table A.1: **Number of Races Per Day Across Symbol-Dates**

| Description | Mean | sd | Min | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 537.24 | 542.96 | 29 | 73 | 152 | 215 | 346 | 629 | 1,194 | 2,635 | 7,014 |
| FTSE 250 | 70.05 | 103.26 | 0 | 0 | 0 | 2 | 35 | 97 | 182 | 477 | 1,392 |
| Full Sample | 206.03 | 372.02 | 0 | 0 | 0 | 11 | 81 | 231 | 513 | 1,919 | 7,014 |

**Notes:** Please see Section 4.2 for a detailed description of the baseline race-detection criteria and Section 3 for details of the message data including how we classify inbound messages and how we maintain the order book. This appendix table reports the distribution of the number of races detected at the symbol-date level. Table 5.1 in the main text reports the distribution at the symbol level and date level.

## Total Time in Races

In the text of Section 5.1 we report the distribution of the number of races per day (Table 5.1) and the distribution of the duration of races (Table 5.2). In this appendix table we report the distribution of the total time in races per day. This is reported in seconds per day at the symbol-date level.

Table A.2: **Total Time in Races Across Symbol-Dates**

| Description | Mean | sd | Min | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 0.044 | 0.047 | 0.002 | 0.006 | 0.012 | 0.017 | 0.028 | 0.052 | 0.096 | 0.235 | 0.739 |
| FTSE 250 | 0.005 | 0.008 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.007 | 0.013 | 0.036 | 0.093 |
| Full Sample | 0.016 | 0.032 | 0.000 | 0.000 | 0.000 | 0.001 | 0.006 | 0.018 | 0.042 | 0.153 | 0.739 |

**Notes:** For each race detected by our baseline method (see Section 4.2 for detailed description) we compute the difference in message timestamps between the first inbound message in the race that is a success and the first inbound message in the race that is a fail (success and fail are defined in Section 4.2.3). Denote these messages S1 and F1, respectively. The duration of a race is defined as the difference between F1's timestamp minus S1's timestamp, that is, by how long did the first successful message in the race beat the first failed message. For each symbol-date in our dataset, we sum all race durations and report the distribution. For example, the table indicates that in the mean FTSE 100 symbol-date, the sum of the duration of all races is 0.044 seconds.

## Symbol-level Version of Table 5.3

Table 5.3 in the main text reports the percentage of volume and trades in races at the date level, i.e., averaged across all symbols in the FTSE 100, FTSE 250, and full sample respectively. In this appendix table we report the percentage of volume and trades in races at the symbol level averaged across all dates.

Table A.3: **Volume and Trades in Races**

Panel A: Percentage of volume (value-weighted) in races across symbols

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 23.48 | 4.90 | 13.08 | 17.84 | 20.07 | 23.30 | 26.34 | 30.62 | 33.75 |
| FTSE 250 | 11.33 | 8.48 | 0.00 | 0.61 | 1.99 | 12.69 | 18.48 | 22.07 | 27.30 |
| Full Sample | 14.86 | 9.40 | 0.00 | 1.11 | 5.79 | 17.20 | 22.02 | 25.78 | 33.06 |

Panel B: Percentage of number of trades in races across symbols

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 22.19 | 4.56 | 12.54 | 16.69 | 19.58 | 21.79 | 24.78 | 28.44 | 32.09 |
| FTSE 250 | 11.31 | 8.37 | 0.00 | 0.55 | 2.00 | 13.21 | 18.32 | 21.63 | 27.31 |
| Full Sample | 14.48 | 8.95 | 0.00 | 0.87 | 6.05 | 16.70 | 21.36 | 24.67 | 31.16 |

**Notes:** Please see the notes for Table 5.3 in the main text. Table 5.3 reports the distribution of percentage of volume and trades in races at the date level. This appendix table reports the same distribution but at the symbol level.

## Additional Data on Messages Per Race

Table 5.4 in the main text reports the number of participants, takes, and cancels in the $T$ microseconds after the start of a race for values of $T$ between 50us and 1ms. In this appendix table we break out the take messages into two types: immediate-or-cancels (IOCs) and limit orders. Recall that in many of the sensitivity analyses discussed in Section 6 we only allow for IOC take messages to count towards the 1+ fails requirement for race detection.

This appendix table also reports the total number of messages and total number of firms in races. The number of firms can be lower than the number of participants in case there are multiple active trading desks within the same firm in a race, and the number of participants can in turn be lower than the number of messages in case some participants send multiple messages in a race.

Table A.4: **Number of IOC / Limit Takes and Number of Messages / Firms in Races**

Panel A: Number of take (IOC) messages

| Description | Mean | sd | Min | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| IOC takes within 50us | 1.56 | 0.99 | 0 | 0 | 1 | 1 | 1 | 2 | 3 | 5 | 14 |
| IOC takes within 100us | 1.80 | 1.10 | 0 | 0 | 1 | 1 | 2 | 2 | 3 | 5 | 15 |
| IOC takes within 200us | 2.20 | 1.32 | 0 | 0 | 1 | 1 | 2 | 3 | 4 | 6 | 17 |
| IOC takes within 500us | 2.81 | 1.73 | 0 | 0 | 1 | 2 | 2 | 4 | 5 | 8 | 29 |
| IOC takes within 1000us | 3.07 | 2.00 | 0 | 0 | 1 | 2 | 3 | 4 | 6 | 10 | 40 |

Panel B: Number of take (limit) messages

| Description | Mean | sd | Min | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Limit takes within 50us | 0.10 | 0.32 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 5 |
| Limit takes within 100us | 0.13 | 0.39 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 6 |
| Limit takes within 200us | 0.17 | 0.45 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 7 |
| Limit takes within 500us | 0.25 | 0.60 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 11 |
| Limit takes within 1000us | 0.37 | 0.82 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 4 | 17 |

Panel C: Number of messages

| Description | Mean | sd | Min | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Messages within 50us | 1.83 | 0.93 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 5 | 14 |
| Messages within 100us | 2.15 | 1.05 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 6 | 15 |
| Messages within 200us | 2.67 | 1.23 | 1 | 1 | 2 | 2 | 2 | 3 | 4 | 7 | 17 |
| Messages within 500us | 3.46 | 1.72 | 2 | 2 | 2 | 2 | 3 | 4 | 6 | 9 | 29 |
| Messages within 1000us | 3.90 | 2.19 | 2 | 2 | 2 | 2 | 3 | 5 | 7 | 12 | 41 |

Panel D: Number of firms

| Description | Mean | sd | Min | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Firms within 50us | 1.55 | 0.69 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 4 | 7 |
| Firms within 100us | 1.77 | 0.76 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 4 | 8 |
| Firms within 200us | 2.12 | 0.82 | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 8 |
| Firms within 500us | 2.60 | 1.01 | 1 | 1 | 2 | 2 | 2 | 3 | 4 | 6 | 10 |
| Firms within 1000us | 2.82 | 1.19 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 6 | 12 |

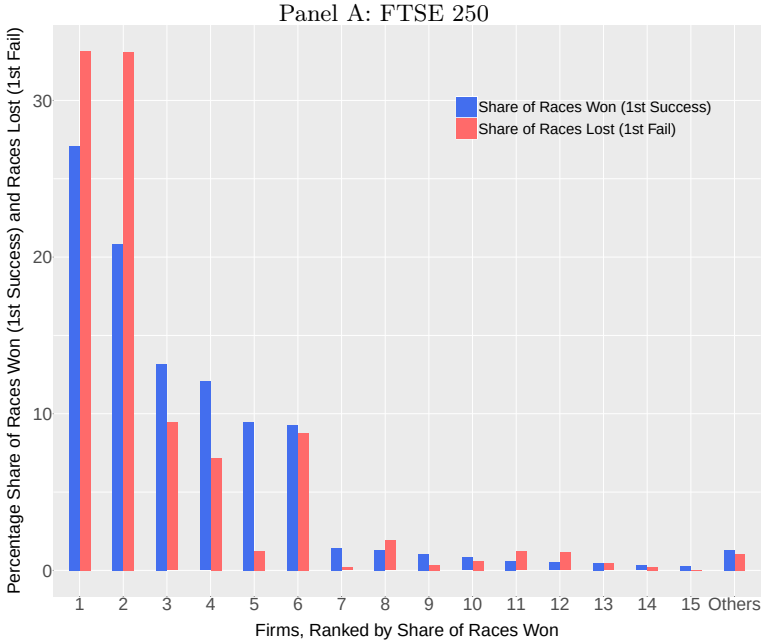
**Notes:** Please see the notes for Table 5.4 and the description in the text above this table.

## Additional Versions of Percentage of 1st Successful and Failed Messages by Firm

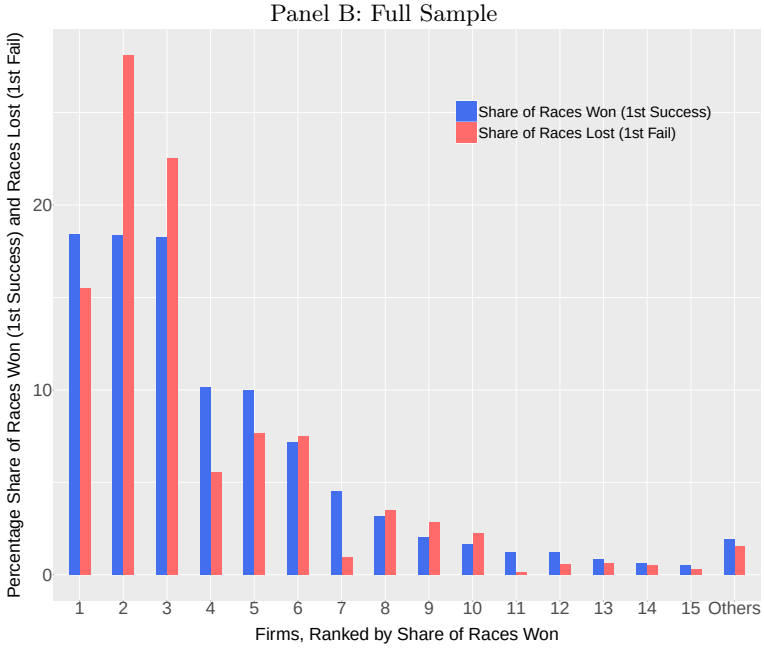
Figure 5.2 in the main text reports the percentage of 1st successful and 1st failed messages in races, by firm, over all races in the FTSE 100. The following two appendix figures report the same figure for the FTSE 250 and full sample.

Figure A.2: **Percentage of 1st Successful and Failed Messages by Firm**

Panel A: FTSE 250




Panel B: Full Sample




**Notes:** Please see the notes for Figure 5.2 and the description in the text above this figure.

## Details for Expected Number of Races by Chance Analysis

In Section 5.2 of the main text, in the subsection "Expected Number of Races by Chance," we discussed the number of times per day we would see $N$ messages on the same side of the order book within $T$ microseconds, by chance, if orders arrive randomly according to a Poisson process. Poisson processes are memoryless meaning that the arrival of a message at one point in time does not make it any more or less likely for other messages to arrive in the interval of time thereafter. We concluded that clusters of messages within short time horizons would be very rare if messages arrive Poisson.

This appendix table provides the support for that discussion. We determine the Poisson arrival rate for each symbol-date based on the total number of potentially-race-relevant messages (i.e., marketable takes or cancels at the best bid or offer) for that symbol-date. We then report the expected number of instances per day in which one would see $N$ participants within $T$ microseconds, given these Poisson arrival rates.

Table A.5: **Number of Instances Per Day With $N$ Participants Within $T$ Microseconds, if Messages Arrive Poisson Randomly**

FTSE 100

| N | T | Mean | sd | Pct01 | Pct25 | Median | Pct75 | Pct99 |
|---|---|---|---|---|---|---|---|---|
| 2 | 50 | 0.35 | 0.80 | 0.01 | 0.04 | 0.09 | 0.32 | 3.28 |
| 2 | 100 | 0.71 | 1.60 | 0.02 | 0.08 | 0.18 | 0.64 | 6.56 |
| 2 | 200 | 1.42 | 3.20 | 0.03 | 0.15 | 0.37 | 1.29 | 13.13 |
| 2 | 500 | 3.55 | 7.99 | 0.08 | 0.38 | 0.91 | 3.22 | 32.81 |
| 2 | 1000 | 7.09 | 15.96 | 0.15 | 0.77 | 1.83 | 6.44 | 65.57 |
| 3 | 1000 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 |
| Actual Number of Races | | | | | | | | |
| Baseline analysis | | 537.24 | 542.96 | 73 | 215 | 346 | 629 | 2,635 |
| Sensitivity: 3+ within Info Horizon | | 228.98 | 206.88 | 28 | 100 | 161 | 278 | 1,002 |

FTSE 250

| N | T | Mean | sd | Pct01 | Pct25 | Median | Pct75 | Pct99 |
|---|---|---|---|---|---|---|---|---|
| 2 | 50 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 |
| 2 | 100 | 0.01 | 0.02 | 0.00 | 0.00 | 0.00 | 0.01 | 0.09 |
| 2 | 200 | 0.02 | 0.04 | 0.00 | 0.00 | 0.01 | 0.02 | 0.17 |
| 2 | 500 | 0.04 | 0.10 | 0.00 | 0.00 | 0.02 | 0.04 | 0.43 |
| 2 | 1000 | 0.08 | 0.20 | 0.00 | 0.01 | 0.03 | 0.08 | 0.86 |
| 3 | 1000 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Actual Number of Races | | | | | | | | |
| Baseline analysis | | 70.05 | 103.26 | 0 | 2 | 35 | 97 | 477 |
| Sensitivity: 3+ within Info Horizon | | 30.68 | 49.17 | 0 | 0 | 12 | 43 | 223 |

**Notes:** This table details the distribution of the expected number of races that would occur by chance in a symbol-date given a Poisson arrival process for messages. For each symbol-date the arrival rate of potentially-race-relevant messages (see text of Section 5.2 for description) is calculated and the expected number of occurrences of N such messages within T microseconds, on the same side of the order book, is computed if messages arrive at this rate via a Poisson arrival process. We also report the actual number of races, both overall and conditioning on their being at least 3+ participants within the 1 millisecond following the start of the race. The analysis is reported separately for FTSE 100 and FTSE 250.

## Distribution of the Bid-Ask Spread by Symbol and Date

Table 5.10 in the main text presents a decomposition of the bid-ask spread into price impact in races, price impact not in races, loss avoidance, and the realized spread. For context on this analysis, this appendix table presents the distribution of the bid-ask spread across symbol (averaged over all dates) and dates (averaged over all symbols). Spreads are presented based on both the time-weighted displayed spread (Panel A) and the quantity-weighted traded spread (Panel B); this latter quantity-weighted spread corresponds to the term effective spread utilized in the literature and in the text of Section 5.5. For each analysis, we present results in both ticks (sub-panel A) and basis points (sub-panel B); this latter measurement corresponds to the spread decomposition reported in the text. All spreads are reported as the "half-spread", i.e., half the distance between the bid and the offer, which corresponds to the difference between the tradable or traded price and the midpoint price. The half-spread is the standard measure in the literature.

Table A.6: **Spread by Date**

Panel A: Time-Weighted Average Half-Spread

Sub-Panel A: Ticks

| Description | Mean | sd | Min | Pct10 | Pct25 | Median | Pct75 | Pct90 | Max |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 0.97 | 0.06 | 0.86 | 0.92 | 0.93 | 0.96 | 1.00 | 1.04 | 1.20 |
| FTSE 250 | 3.40 | 0.35 | 2.83 | 2.99 | 3.19 | 3.34 | 3.61 | 3.81 | 4.38 |
| Full Sample | 2.70 | 0.26 | 2.29 | 2.39 | 2.53 | 2.63 | 2.86 | 2.98 | 3.45 |

Sub-Panel B: Basis Points

| Description | Mean | sd | Min | Pct10 | Pct25 | Median | Pct75 | Pct90 | Max |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 3.77 | 0.20 | 3.42 | 3.54 | 3.66 | 3.76 | 3.82 | 3.97 | 4.39 |
| FTSE 250 | 15.76 | 1.48 | 13.11 | 13.97 | 14.81 | 15.62 | 16.66 | 17.67 | 19.62 |
| Full Sample | 12.27 | 1.09 | 10.35 | 10.92 | 11.55 | 12.22 | 12.93 | 13.63 | 15.19 |

Panel B: Quantity-Weighted Average Half-Spread ("Effective Spread")

Sub-Panel A: Ticks

| Description | Mean | sd | Min | Pct10 | Pct25 | Median | Pct75 | Pct90 | Max |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 0.85 | 0.17 | 0.70 | 0.74 | 0.76 | 0.80 | 0.86 | 1.00 | 1.71 |
| FTSE 250 | 1.44 | 0.13 | 1.15 | 1.31 | 1.37 | 1.44 | 1.47 | 1.53 | 1.82 |
| Full Sample | 0.93 | 0.15 | 0.77 | 0.83 | 0.85 | 0.88 | 0.95 | 1.06 | 1.66 |

Sub-Panel B: Basis Points

| Description | Mean | sd | Min | Pct10 | Pct25 | Median | Pct75 | Pct90 | Max |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 2.65 | 0.29 | 2.28 | 2.45 | 2.52 | 2.59 | 2.72 | 2.80 | 4.28 |
| FTSE 250 | 6.76 | 0.58 | 5.72 | 6.24 | 6.44 | 6.66 | 6.95 | 7.19 | 8.97 |
| Full Sample | 3.17 | 0.27 | 2.74 | 2.92 | 3.06 | 3.12 | 3.22 | 3.38 | 4.52 |

**Notes:** Please see the description in the text above this table for a description of the spread variables. This table reports distributions of the spread at the date level, averaging over symbols.

Table A.7: **Spread by Symbol**

Panel A: Time-Weighted Average Half-Spread

Sub-Panel A: Ticks

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 0.97 | 0.32 | 0.56 | 0.64 | 0.83 | 0.92 | 1.02 | 1.32 | 2.14 |
| FTSE 250 | 3.40 | 3.00 | 0.83 | 1.09 | 1.53 | 2.57 | 3.94 | 6.52 | 16.73 |
| Full Sample | 2.70 | 2.76 | 0.58 | 0.85 | 1.01 | 1.79 | 3.25 | 5.67 | 12.86 |

Sub-Panel B: Basis Points

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 3.77 | 1.56 | 1.09 | 1.70 | 2.56 | 3.77 | 4.85 | 5.49 | 7.59 |
| FTSE 250 | 15.76 | 13.67 | 3.38 | 6.36 | 7.74 | 11.32 | 17.92 | 29.90 | 59.41 |
| Full Sample | 12.27 | 12.76 | 1.21 | 3.09 | 4.95 | 8.10 | 15.04 | 27.07 | 56.01 |

Panel B: Quantity-Weighted Average Half-Spread ("Effective Spread")

Sub-Panel A: Ticks

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 0.80 | 0.27 | 0.52 | 0.55 | 0.64 | 0.73 | 0.89 | 1.17 | 1.71 |
| FTSE 250 | 2.09 | 1.42 | 0.60 | 0.84 | 1.13 | 1.75 | 2.58 | 3.80 | 6.62 |
| Full Sample | 1.72 | 1.34 | 0.54 | 0.66 | 0.81 | 1.32 | 2.17 | 3.21 | 6.38 |

Sub-Panel B: Basis Points

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| FTSE 100 | 3.27 | 1.22 | 1.22 | 1.75 | 2.28 | 3.18 | 4.13 | 4.91 | 5.79 |
| FTSE 250 | 11.61 | 9.53 | 2.66 | 4.90 | 5.99 | 8.22 | 13.67 | 22.96 | 47.35 |
| Full Sample | 9.18 | 8.90 | 1.29 | 2.59 | 4.21 | 6.26 | 10.38 | 18.47 | 40.07 |

**Notes:** Please see the description in the text above this table for a description of the spread variables. This table reports distributions of the spread at the symbol level, averaging over dates.

## Spread Decomposition with Different Time Horizons

Table 5.10 in the main text reports results of our spread decomposition (Section 5.5, Approach #1) using a 10 second mark-to-market time horizon for calculating price impact and loss avoidance. In this appendix we report the same decomposition but using 100 millisecond and 1 second time horizons instead. Notably, the realized spread appears to decline with the time horizon, from 100 millisecond to 1 second to 10 seconds, both in and out of races. While the overall sample realized spread is slightly negative at 10 seconds, it is slightly positive at 100 millisecond and 1 second. This pattern is consistent with price impact being smaller at shorter time horizons as discussed in Conrad and Wahal (2019).

Table A.8: **Spread Decomposition - 100ms**

Panel A: FTSE 100 by Symbol

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| Effective spread paid - overall (bps) | 3.27 | 1.22 | 1.22 | 1.75 | 2.28 | 3.18 | 4.13 | 4.91 | 5.79 |
| Effective spread paid - in races (bps) | 3.18 | 1.22 | 0.99 | 1.70 | 2.21 | 3.17 | 4.05 | 4.89 | 5.98 |
| Effective spread paid - not in races (bps) | 3.29 | 1.22 | 1.25 | 1.78 | 2.30 | 3.17 | 4.15 | 4.96 | 5.71 |
| Price impact - overall (bps) | 3.18 | 1.25 | 1.16 | 1.71 | 2.18 | 3.06 | 3.96 | 5.06 | 5.82 |
| Price impact - in races (bps) | 4.52 | 1.75 | 1.61 | 2.52 | 3.07 | 4.26 | 5.76 | 7.23 | 7.89 |
| Price impact - not in races (bps) | 2.75 | 1.03 | 1.03 | 1.47 | 1.92 | 2.72 | 3.36 | 4.25 | 4.94 |
| Loss avoidance (bps) | 0.00 | 0.01 | -0.01 | -0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.02 |
| Realized spread - overall (bps) | 0.09 | 0.27 | -0.43 | -0.20 | -0.03 | 0.06 | 0.18 | 0.37 | 1.06 |
| Realized spread - in races (bps) | -1.33 | 0.62 | -2.80 | -2.32 | -1.68 | -1.11 | -0.88 | -0.71 | -0.53 |
| Realized spread - not in races (bps) | 0.55 | 0.30 | 0.08 | 0.22 | 0.29 | 0.50 | 0.74 | 0.92 | 1.41 |
| PI in races / PI total (%) | 33.26 | 6.28 | 21.27 | 25.97 | 29.36 | 31.77 | 37.35 | 43.12 | 46.06 |
| PI in races / Effective spread (%) | 32.49 | 7.56 | 18.81 | 23.89 | 28.30 | 30.96 | 36.37 | 43.84 | 49.45 |

Panel B: FTSE 250 by Symbol

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| Effective spread paid - overall (bps) | 8.06 | 3.81 | 2.65 | 4.63 | 5.59 | 7.14 | 9.84 | 13.10 | 19.10 |
| Effective spread paid - in races (bps) | 6.74 | 3.03 | 2.42 | 4.32 | 4.97 | 6.08 | 7.63 | 9.96 | 15.62 |
| Effective spread paid - not in races (bps) | 8.22 | 3.87 | 2.72 | 4.70 | 5.72 | 7.31 | 9.94 | 13.34 | 19.55 |
| Price impact - overall (bps) | 5.99 | 2.47 | 2.24 | 3.58 | 4.34 | 5.44 | 7.09 | 9.23 | 14.30 |
| Price impact - in races (bps) | 9.38 | 4.87 | 3.50 | 5.39 | 6.51 | 8.23 | 11.07 | 13.93 | 26.88 |
| Price impact - not in races (bps) | 5.53 | 2.45 | 2.02 | 3.26 | 3.86 | 4.89 | 6.55 | 8.94 | 13.50 |
| Loss avoidance (bps) | -0.00 | 0.02 | -0.05 | -0.02 | -0.01 | -0.00 | 0.00 | 0.01 | 0.06 |
| Realized spread - overall (bps) | 2.07 | 1.69 | -0.04 | 0.45 | 1.17 | 1.82 | 2.57 | 3.51 | 6.97 |
| Realized spread - in races (bps) | -2.64 | 2.75 | -12.96 | -5.92 | -3.14 | -1.97 | -1.06 | -0.47 | 0.99 |
| Realized spread - not in races (bps) | 2.69 | 1.70 | 0.42 | 1.22 | 1.74 | 2.44 | 3.18 | 4.28 | 7.07 |
| PI in races / PI total (%) | 21.82 | 9.31 | 2.14 | 7.49 | 15.08 | 23.34 | 28.22 | 32.29 | 39.41 |
| PI in races / Effective spread (%) | 17.14 | 8.59 | 1.30 | 4.59 | 10.97 | 17.30 | 22.54 | 27.63 | 37.15 |

Panel C: Full Sample by Date

| Description | Mean | sd | Min | Pct10 | Pct25 | Median | Pct75 | Pct90 | Max |
|---|---|---|---|---|---|---|---|---|---|
| Effective spread paid - overall (bps) | 3.17 | 0.27 | 2.74 | 2.92 | 3.06 | 3.12 | 3.22 | 3.38 | 4.52 |
| Effective spread paid - in races (bps) | 2.99 | 0.13 | 2.64 | 2.84 | 2.90 | 2.99 | 3.06 | 3.16 | 3.28 |
| Effective spread paid - not in races (bps) | 3.22 | 0.32 | 2.77 | 2.95 | 3.10 | 3.17 | 3.29 | 3.44 | 4.90 |
| Price impact - overall (bps) | 2.88 | 0.16 | 2.54 | 2.71 | 2.79 | 2.90 | 2.95 | 3.13 | 3.18 |
| Price impact - in races (bps) | 4.22 | 0.17 | 3.81 | 4.00 | 4.13 | 4.22 | 4.35 | 4.45 | 4.60 |
| Price impact - not in races (bps) | 2.52 | 0.15 | 2.19 | 2.33 | 2.43 | 2.52 | 2.58 | 2.72 | 2.84 |
| Loss avoidance (bps) | 0.00 | 0.00 | -0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 |
| Realized spread - overall (bps) | 0.29 | 0.23 | 0.11 | 0.17 | 0.20 | 0.24 | 0.30 | 0.39 | 1.66 |
| Realized spread - in races (bps) | -1.24 | 0.08 | -1.48 | -1.33 | -1.28 | -1.23 | -1.19 | -1.13 | -1.06 |
| Realized spread - not in races (bps) | 0.70 | 0.26 | 0.51 | 0.57 | 0.61 | 0.65 | 0.73 | 0.81 | 2.26 |
| PI in races / PI total (%) | 31.43 | 2.31 | 24.08 | 28.54 | 30.40 | 31.69 | 32.47 | 34.07 | 36.64 |
| PI in races / Effective spread (%) | 28.77 | 3.12 | 15.24 | 26.47 | 27.76 | 29.26 | 30.37 | 31.92 | 34.52 |

**Notes:** Please see the notes for Table 5.10 in the main text. This table is the same except that price impact and loss avoidance are calculated based on mark-to-market at 100 milliseconds instead of 10 seconds.

Table A.9: **Spread Decomposition - 1s**

Panel A: FTSE 100 by Symbol

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| Effective spread paid - overall (bps) | 3.27 | 1.22 | 1.22 | 1.75 | 2.28 | 3.18 | 4.13 | 4.91 | 5.79 |
| Effective spread paid - in races (bps) | 3.18 | 1.22 | 0.99 | 1.70 | 2.21 | 3.17 | 4.05 | 4.89 | 5.98 |
| Effective spread paid - not in races (bps) | 3.29 | 1.22 | 1.25 | 1.78 | 2.30 | 3.17 | 4.15 | 4.96 | 5.71 |
| Price impact - overall (bps) | 3.39 | 1.29 | 1.27 | 1.85 | 2.34 | 3.34 | 4.15 | 5.20 | 6.30 |
| Price impact - in races (bps) | 4.81 | 1.78 | 1.83 | 2.78 | 3.33 | 4.63 | 6.04 | 7.44 | 8.33 |
| Price impact - not in races (bps) | 2.93 | 1.07 | 1.13 | 1.60 | 2.06 | 2.98 | 3.51 | 4.44 | 5.39 |
| Loss avoidance (bps) | 0.00 | 0.01 | -0.00 | -0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.02 |
| Realized spread - overall (bps) | -0.12 | 0.25 | -0.56 | -0.38 | -0.25 | -0.15 | -0.00 | 0.14 | 0.76 |
| Realized spread - in races (bps) | -1.63 | 0.62 | -3.24 | -2.54 | -1.98 | -1.48 | -1.15 | -0.91 | -0.76 |
| Realized spread - not in races (bps) | 0.36 | 0.28 | -0.09 | 0.06 | 0.16 | 0.32 | 0.55 | 0.72 | 1.13 |
| PI in races / PI total (%) | 33.29 | 6.26 | 20.88 | 25.73 | 29.49 | 32.11 | 37.49 | 42.69 | 46.16 |
| PI in races / Effective spread (%) | 34.74 | 7.42 | 19.79 | 26.20 | 30.94 | 34.06 | 39.08 | 44.93 | 49.85 |

Panel B: FTSE 250 by Symbol

| Description | Mean | sd | Pct01 | Pct10 | Pct25 | Median | Pct75 | Pct90 | Pct99 |
|---|---|---|---|---|---|---|---|---|---|
| Effective spread paid - overall (bps) | 8.06 | 3.81 | 2.65 | 4.63 | 5.59 | 7.14 | 9.84 | 13.10 | 19.11 |
| Effective spread paid - in races (bps) | 6.74 | 3.03 | 2.42 | 4.32 | 4.97 | 6.08 | 7.63 | 9.96 | 15.62 |
| Effective spread paid - not in races (bps) | 8.22 | 3.87 | 2.72 | 4.70 | 5.72 | 7.31 | 9.94 | 13.34 | 19.55 |
| Price impact - overall (bps) | 6.71 | 2.83 | 2.43 | 4.14 | 4.95 | 5.98 | 7.79 | 10.34 | 17.10 |
| Price impact - in races (bps) | 10.44 | 5.46 | 3.75 | 6.14 | 7.33 | 9.10 | 12.28 | 15.39 | 29.90 |
| Price impact - not in races (bps) | 6.20 | 2.82 | 2.18 | 3.63 | 4.41 | 5.41 | 7.23 | 9.85 | 16.38 |
| Loss avoidance (bps) | -0.00 | 0.01 | -0.04 | -0.01 | -0.00 | -0.00 | 0.00 | 0.01 | 0.07 |
| Realized spread - overall (bps) | 1.35 | 1.44 | -0.46 | 0.06 | 0.57 | 1.11 | 1.73 | 2.66 | 5.68 |
| Realized spread - in races (bps) | -3.70 | 3.14 | -16.39 | -6.99 | -4.13 | -2.65 | -1.99 | -1.44 | -0.69 |
| Realized spread - not in races (bps) | 2.02 | 1.44 | 0.22 | 0.81 | 1.25 | 1.80 | 2.43 | 3.38 | 5.89 |
| PI in races / PI total (%) | 21.79 | 9.41 | 2.10 | 6.72 | 15.03 | 23.58 | 28.40 | 32.31 | 39.77 |
| PI in races / Effective spread (%) | 19.03 | 9.41 | 1.61 | 5.19 | 12.08 | 19.61 | 25.39 | 30.01 | 41.32 |

Panel C: Full Sample by Date

| Description | Mean | sd | Min | Pct10 | Pct25 | Median | Pct75 | Pct90 | Max |
|---|---|---|---|---|---|---|---|---|---|
| Effective spread paid - overall (bps) | 3.17 | 0.27 | 2.74 | 2.92 | 3.06 | 3.12 | 3.22 | 3.38 | 4.52 |
| Effective spread paid - in races (bps) | 2.99 | 0.13 | 2.64 | 2.84 | 2.90 | 2.99 | 3.06 | 3.16 | 3.28 |
| Effective spread paid - not in races (bps) | 3.22 | 0.32 | 2.77 | 2.95 | 3.10 | 3.17 | 3.29 | 3.44 | 4.90 |
| Price impact - overall (bps) | 3.10 | 0.17 | 2.72 | 2.90 | 3.00 | 3.11 | 3.21 | 3.36 | 3.44 |
| Price impact - in races (bps) | 4.51 | 0.20 | 4.08 | 4.26 | 4.39 | 4.51 | 4.66 | 4.75 | 4.98 |
| Price impact - not in races (bps) | 2.71 | 0.17 | 2.35 | 2.54 | 2.61 | 2.71 | 2.78 | 2.99 | 3.06 |
| Loss avoidance (bps) | 0.00 | 0.00 | -0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 |
| Realized spread - overall (bps) | 0.07 | 0.22 | -0.11 | -0.06 | -0.02 | 0.02 | 0.12 | 0.19 | 1.31 |
| Realized spread - in races (bps) | -1.52 | 0.11 | -1.86 | -1.65 | -1.58 | -1.52 | -1.45 | -1.40 | -1.32 |
| Realized spread - not in races (bps) | 0.50 | 0.24 | 0.29 | 0.36 | 0.41 | 0.46 | 0.55 | 0.62 | 1.89 |
| PI in races / PI total (%) | 31.24 | 2.41 | 23.10 | 28.32 | 30.29 | 31.69 | 32.37 | 33.99 | 36.59 |
| PI in races / Effective spread (%) | 30.71 | 3.37 | 16.41 | 28.06 | 29.47 | 31.27 | 32.89 | 34.03 | 36.64 |

**Notes:** Please see the notes for Table 5.10 in the main text. This table is the same except that price impact and loss avoidance are calculated based on mark-to-market at 1 second instead of 10 seconds.

Sensitivity Analysis: 5+ Race Participants

Table 6.2 in the text of Section 6.2 reports a sensitivity analysis for requiring 3+ participants in a race. This appendix table is analogous except that it requires 5+ participants in a race instead of 3+.

Table A.10: **Sensitivity Analysis: 5+ Race Participants**

| Measure | Baseline | 5+ Race Participants Within InfoHor | 50$\mu$s | 100$\mu$s | 200$\mu$s | 500$\mu$s | 1ms | 2ms | 3ms |
|---|---|---|---|---|---|---|---|---|---|
| **Frequency and Duration of Races** | | | | | | | | | |
| Races per day | | | | | | | | | |
| FTSE 100 - per symbol | 537.24 | 37.83 | 5.96 | 13.58 | 35.27 | 121.76 | 202.00 | 268.66 | 297.78 |
| FTSE 250 - per symbol | 70.05 | 4.91 | 0.88 | 2.03 | 5.21 | 16.36 | 26.67 | 33.77 | 36.62 |
| Mean race duration (microseconds) | 78.65 | 73.23 | 11.14 | 23.94 | 61.66 | 170.05 | 304.84 | 469.80 | 582.24 |
| % of races with wrong winner | 4.30 | 5.62 | 14.93 | 9.48 | 4.98 | 2.32 | 1.84 | 1.45 | 1.30 |
| % of volume in races | | | | | | | | | |
| FTSE 100 | 22.15 | 3.39 | 0.38 | 0.99 | 2.70 | 10.36 | 17.94 | 23.31 | 25.51 |
| FTSE 250 | 16.90 | 2.23 | 0.33 | 0.78 | 2.12 | 7.65 | 12.81 | 16.43 | 17.87 |
| Full Sample | 21.46 | 3.24 | 0.37 | 0.97 | 2.62 | 10.01 | 17.27 | 22.41 | 24.52 |
| Mean number of messages within 500 $\mu$s | 3.46 | 7.04 | 7.37 | 7.37 | 7.06 | 6.23 | 4.79 | 4.11 | 3.90 |
| **Per-Race Profits** | | | | | | | | | |
| Per-share profits | | | | | | | | | |
| ticks | 0.55 | 1.01 | 1.02 | 0.98 | 0.92 | 0.84 | 0.83 | 0.86 | 0.87 |
| GBX | 0.17 | 0.34 | 0.29 | 0.31 | 0.30 | 0.28 | 0.27 | 0.27 | 0.27 |
| basis points | 1.66 | 3.39 | 3.25 | 3.30 | 3.13 | 2.79 | 2.69 | 2.66 | 2.64 |
| Per-race profits GBP | | | | | | | | | |
| displayed depth | 1.85 | 6.30 | 4.52 | 5.14 | 5.01 | 4.89 | 4.82 | 4.63 | 4.58 |
| qty trade/cancel | 1.76 | 6.29 | 4.28 | 4.91 | 4.96 | 5.06 | 5.03 | 4.84 | 4.80 |
| **Aggregate Profits and LA Tax** | | | | | | | | | |
| Daily Profits | | | | | | | | | |
| FTSE 100 - per symbol | 1,047 | 262 | 29 | 77 | 195 | 637 | 1,037 | 1,310 | 1,433 |
| FTSE 250 - per symbol | 108 | 21 | 3 | 7 | 19 | 63 | 102 | 129 | 139 |
| Full Sample - aggregate | 132,378 | 31,663 | 3,699 | 9,609 | 24,265 | 79,717 | 129,773 | 163,927 | 178,855 |
| Latency Arbitrage Tax, All Volume (bps) | | | | | | | | | |
| FTSE 100 | 0.38 | 0.10 | 0.01 | 0.03 | 0.07 | 0.24 | 0.38 | 0.48 | 0.53 |
| FTSE 250 | 0.66 | 0.13 | 0.02 | 0.04 | 0.12 | 0.38 | 0.62 | 0.79 | 0.85 |
| Full Sample | 0.42 | 0.10 | 0.01 | 0.03 | 0.08 | 0.25 | 0.41 | 0.52 | 0.57 |
| Latency Arbitrage Tax, Non-Race Volume (bps) | | | | | | | | | |
| FTSE 100 | 0.49 | 0.12 | 0.01 | 0.03 | 0.09 | 0.38 | 0.68 | 0.91 | 1.03 |
| FTSE 250 | 0.80 | 0.16 | 0.02 | 0.05 | 0.14 | 0.58 | 1.02 | 1.36 | 1.50 |
| Full Sample | 0.53 | 0.13 | 0.01 | 0.04 | 0.10 | 0.40 | 0.73 | 0.97 | 1.09 |
| **Spread Decomposition** | | | | | | | | | |
| Price impact in races / All price impact % | 30.58 | 6.01 | 0.63 | 1.72 | 4.61 | 16.26 | 27.53 | 35.79 | 39.63 |
| Price impact in races / Effective spread % | 32.82 | 6.46 | 0.68 | 1.85 | 4.96 | 17.46 | 29.56 | 38.43 | 42.55 |
| Loss avoidance / Effective spread % | 0.19 | 0.06 | 0.01 | 0.03 | 0.10 | 0.42 | 0.92 | 1.38 | 1.61 |
| **Implied Reduction in Cost of Liquidity** | | | | | | | | | |
| % Reduction in liquidity cost | | | | | | | | | |
| FTSE 100 - by symbol | 19.95 | 3.88 | 0.50 | 1.15 | 2.91 | 10.46 | 18.79 | 25.71 | 29.26 |
| FTSE 250 - by symbol | 11.93 | 2.76 | 0.77 | 1.18 | 2.40 | 7.29 | 12.09 | 15.75 | 17.14 |
| Full Sample - by date | 16.73 | 3.31 | 0.38 | 0.99 | 2.55 | 8.94 | 15.82 | 21.39 | 24.19 |

**Notes:** Please see the notes and surrounding text for Table 6.2. This table is identical except it conditions on 5+ participants in a race instead of 3+ participants.

### Sensitivity Analysis: 2+ Unique Firms

Our baseline method requires that a race contains at least 2 unique participants as determined by their UserID in our data. As discussed in the text, some firms use different UserIDs for different trading desks. Typically, this will be the case when the trading desks are operated sufficiently separately that if they happen to trade with each other the firm would not be in violation of wash-trade requirements. This economic separation is the reason why our baseline uses UserIDs as the measurement of the number of participants. The following appendix table provides results if the requirement is changed from 2+ unique participants to 2+ unique firms. The format is analogous to Table 6.2 in the main text, and the results can also be compared to Table 6.1 in the main text.

Table A.11: **Sensitivity Analysis: 2+ Participating Firms**

| Measure | Baseline | 2+ Participating Firms Within: InfoHor | 50$\mu$s | 100$\mu$s | 200$\mu$s | 500$\mu$s | 1ms | 2ms | 3ms |
|---|---|---|---|---|---|---|---|---|---|
| **Frequency and Duration of Races** | | | | | | | | | |
| Races per day | | | | | | | | | |
| FTSE 100 - per symbol | 537.24 | 479.32 | 247.25 | 332.99 | 462.39 | 736.14 | 818.92 | 871.31 | 891.49 |
| FTSE 250 - per symbol | 70.05 | 60.44 | 32.74 | 43.39 | 59.97 | 102.41 | 116.32 | 122.08 | 124.89 |
| Mean race duration (microseconds) | 78.65 | 81.59 | 16.08 | 31.24 | 74.03 | 196.91 | 306.40 | 447.04 | 552.85 |
| % of races with wrong winner | 4.30 | 4.67 | 9.46 | 7.22 | 4.57 | 2.05 | 1.73 | 1.48 | 1.38 |
| % of volume in races | | | | | | | | | |
| FTSE 100 | 22.15 | 20.08 | 8.15 | 11.51 | 17.62 | 35.79 | 42.20 | 45.86 | 47.26 |
| FTSE 250 | 16.90 | 15.19 | 6.89 | 9.54 | 14.37 | 31.57 | 36.83 | 39.52 | 40.63 |
| Full Sample | 21.46 | 19.44 | 7.98 | 11.25 | 17.19 | 35.23 | 41.49 | 45.02 | 46.39 |
| Mean number of messages within 500 $\mu$s | 3.46 | 3.52 | 3.52 | 3.54 | 3.58 | 3.47 | 3.08 | 2.90 | 2.83 |
| **Per-Race Profits** | | | | | | | | | |
| Per-share profits | | | | | | | | | |
| ticks | 0.55 | 0.54 | 0.51 | 0.50 | 0.50 | 0.53 | 0.56 | 0.57 | 0.58 |
| GBX | 0.17 | 0.17 | 0.16 | 0.16 | 0.16 | 0.16 | 0.17 | 0.17 | 0.17 |
| basis points | 1.66 | 1.65 | 1.62 | 1.58 | 1.56 | 1.62 | 1.67 | 1.69 | 1.70 |
| Per-race profits GBP | | | | | | | | | |
| displayed depth | 1.85 | 1.93 | 1.58 | 1.60 | 1.65 | 1.94 | 2.04 | 2.08 | 2.09 |
| qty trade/cancel | 1.76 | 1.83 | 1.40 | 1.45 | 1.56 | 1.94 | 2.05 | 2.10 | 2.12 |
| **Aggregate Profits and LA Tax** | | | | | | | | | |
| Daily Profits | | | | | | | | | |
| FTSE 100 - per symbol | 1,047 | 971 | 404 | 553 | 793 | 1,482 | 1,744 | 1,889 | 1,945 |
| FTSE 250 - per symbol | 108 | 98 | 46 | 62 | 87 | 176 | 205 | 221 | 226 |
| Full Sample - aggregate | 132,378 | 122,218 | 52,221 | 70,992 | 101,416 | 192,912 | 226,603 | 245,049 | 252,001 |
| Latency Arbitrage Tax, All Volume (bps) | | | | | | | | | |
| FTSE 100 | 0.38 | 0.36 | 0.15 | 0.20 | 0.29 | 0.54 | 0.64 | 0.69 | 0.71 |
| FTSE 250 | 0.66 | 0.60 | 0.29 | 0.38 | 0.53 | 1.08 | 1.26 | 1.35 | 1.39 |
| Full Sample | 0.42 | 0.39 | 0.17 | 0.23 | 0.32 | 0.61 | 0.72 | 0.77 | 0.80 |
| Latency Arbitrage Tax, Non-Race Volume (bps) | | | | | | | | | |
| FTSE 100 | 0.49 | 0.46 | 0.17 | 0.24 | 0.36 | 0.87 | 1.13 | 1.31 | 1.39 |
| FTSE 250 | 0.80 | 0.72 | 0.31 | 0.43 | 0.64 | 1.63 | 2.06 | 2.32 | 2.43 |
| Full Sample | 0.53 | 0.49 | 0.18 | 0.26 | 0.40 | 0.97 | 1.26 | 1.45 | 1.53 |
| **Spread Decomposition** | | | | | | | | | |
| Price impact in races / All price impact % | 30.58 | 28.12 | 10.66 | 15.27 | 23.26 | 48.00 | 57.26 | 62.91 | 65.21 |
| Price impact in races / Effective spread % | 32.82 | 30.18 | 11.43 | 16.38 | 24.97 | 51.50 | 61.44 | 67.50 | 69.97 |
| Loss avoidance / Effective spread % | 0.19 | 0.19 | 0.07 | 0.13 | 0.26 | 0.53 | 0.94 | 1.32 | 1.48 |
| **Implied Reduction in Cost of Liquidity** | | | | | | | | | |
| % Reduction in liquidity cost | | | | | | | | | |
| FTSE 100 - by symbol | 19.95 | 18.15 | 6.55 | 9.30 | 14.28 | 34.09 | 45.35 | 53.50 | 57.24 |
| FTSE 250 - by symbol | 11.93 | 10.49 | 4.97 | 6.68 | 9.52 | 22.75 | 28.20 | 30.26 | 31.44 |
| Full Sample - by date | 16.73 | 15.15 | 5.66 | 7.99 | 12.18 | 28.90 | 37.77 | 44.04 | 46.86 |

**Notes:** Please see the description in the text above this table for a description. The table is identical to Table 6.1 in the main text except it conditions on 2+ unique firms in a race whereas the baseline conditions on 2+ unique participants.

### Sensitivity Analysis: 1+ Cancels and 2+ Takes

Table 6.3 in Section 6.3 of the main text presents sensitivity analysis for requiring 1+ cancel in a race and, separately, for requiring 2+ takes in a race. The former rules out races with 0 cancels (and hence 2+ takes, at least one of which succeeds and one of which fails); the latter rules out races with 1+ cancels and exactly 1 take. The following appendix table presents sensitivity analysis for requiring both criteria simultaneously. This rules out races with either 0 cancels, or with 1+ cancels and exactly 1 take.

Table A.12: **Sensitivity Analysis: 1+ Cancels and 2+ Takes**

| Measure | Baseline | 1+ Cancel and 2+ Takes Within InfoHor | 50$\mu$s | 500$\mu$s | 1ms | 3ms |
|---|---|---|---|---|---|---|
| **Frequency and Duration of Races** | | | | | | |
| Races per day | | | | | | |
| FTSE 100 - per symbol | 537.24 | 59.32 | 16.56 | 145.02 | 208.36 | 285.82 |
| FTSE 250 - per symbol | 70.05 | 5.27 | 1.69 | 14.29 | 21.39 | 29.83 |
| Mean race duration (microseconds) | 78.65 | 86.51 | 17.41 | 218.12 | 384.76 | 754.95 |
| % of races with wrong winner | 4.30 | 3.60 | 7.57 | 1.59 | 1.22 | 0.79 |
| % of volume in races | | | | | | |
| FTSE 100 | 22.15 | 3.74 | 0.72 | 9.18 | 14.32 | 20.28 |
| FTSE 250 | 16.90 | 1.60 | 0.38 | 4.85 | 7.88 | 11.57 |
| Full Sample | 21.46 | 3.46 | 0.67 | 8.62 | 13.49 | 19.16 |
| Mean number of messages within 500 $\mu$s | 3.46 | 4.72 | 4.68 | 4.42 | 3.53 | 3.00 |
| **Per-Race Profits** | | | | | | |
| Per-share profits | | | | | | |
| ticks | 0.55 | 0.59 | 0.44 | 0.47 | 0.50 | 0.52 |
| GBX | 0.17 | 0.20 | 0.15 | 0.15 | 0.15 | 0.16 |
| basis points | 1.66 | 1.80 | 1.44 | 1.42 | 1.46 | 1.51 |
| Per-race profits GBP | | | | | | |
| displayed depth | 1.85 | 3.42 | 2.36 | 2.59 | 2.80 | 3.02 |
| qty trade/cancel | 1.76 | 3.23 | 2.01 | 2.54 | 2.77 | 3.02 |
| **Aggregate Profits and LA Tax** | | | | | | |
| Daily Profits | | | | | | |
| FTSE 100 - per symbol | 1,047 | 220 | 43 | 402 | 626 | 920 |
| FTSE 250 - per symbol | 108 | 11 | 3 | 26 | 42 | 65 |
| Full Sample - aggregate | 132,378 | 24,881 | 4,925 | 46,952 | 73,558 | 109,059 |
| Latency Arbitrage Tax, All Volume (bps) | | | | | | |
| FTSE 100 | 0.38 | 0.08 | 0.02 | 0.15 | 0.23 | 0.34 |
| FTSE 250 | 0.66 | 0.07 | 0.02 | 0.16 | 0.26 | 0.41 |
| Full Sample | 0.42 | 0.08 | 0.02 | 0.15 | 0.23 | 0.35 |
| Latency Arbitrage Tax, Non-Race Volume (bps) | | | | | | |
| FTSE 100 | 0.49 | 0.10 | 0.02 | 0.24 | 0.41 | 0.66 |
| FTSE 250 | 0.80 | 0.08 | 0.02 | 0.24 | 0.43 | 0.71 |
| Full Sample | 0.53 | 0.10 | 0.02 | 0.24 | 0.41 | 0.67 |
| **Spread Decomposition** | | | | | | |
| Price impact in races / All price impact % | 30.58 | 5.42 | 0.98 | 11.82 | 18.82 | 28.48 |
| Price impact in races / Effective spread % | 32.82 | 5.82 | 1.05 | 12.70 | 20.21 | 30.58 |
| Loss avoidance / Effective spread % | 0.19 | 0.16 | 0.06 | 0.59 | 1.09 | 1.76 |
| **Implied Reduction in Cost of Liquidity** | | | | | | |
| % Reduction in liquidity cost | | | | | | |
| FTSE 100 - by symbol | 19.95 | 3.10 | 0.59 | 6.18 | 10.35 | 16.88 |
| FTSE 250 - by symbol | 11.93 | 1.28 | 0.41 | 2.86 | 4.49 | 7.02 |
| Full Sample - by date | 16.73 | 2.61 | 0.50 | 5.22 | 8.68 | 14.05 |

**Notes:** Please see the description in the text above this table and in Section 6.3 for a description. The table is similar to Table 6.3 in the main text except that it conditions on both 1+ cancels and 2+ takes.

## Additional Extrapolation Models

Table 7.1 in the main text presents regressions of daily latency arbitrage profits on volume and 1-minute realized volatility. These regressions were used for the purpose of out-of-sample extrapolation in Section 7. The following appendix table presents analogous regressions using additional volatility variables, as was discussed in the main text. Columns (1)-(4) are analogous to Columns (3)-(6) in Table 7.1, but using 5-minute realized volatility instead of 1-minute realized volatility. Columns (5)-(8) are analogous to the same columns in Table 7.1, but using midpoint distance traveled (Budish, Cramton and Shim, 2015) as the volatility measure. As discussed in the main text, the fit is worse with 5-minute realized volatility than with 1-minute realized volatility, and is slightly better with midpoint distance traveled. We nevertheless utilize 1-minute realized volatility in the main text since it is more easily interpreted, and its measurement does not depend on the number of significant digits of the trading index (or the tick size if using a futures contract price for the index) in the way that distance traveled does.

Table A.13: **Extrapolation Models (Appendix)**

| | *Dependent variable:* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Latency Arbitrage Profits (GBP) | | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Volume (10,000 GBP) | | | 0.4237*** | 0.4123*** | | | 0.2561*** | 0.2833*** |
| | | | (0.0583) | (0.0320) | | | (0.0790) | (0.0578) |
| Volatility (5 min) * Average Volume | 0.0147*** | 0.0276*** | 0.0004 | 0.0006 | | | | |
| | (0.0020) | (0.0013) | (0.0024) | (0.0022) | | | | |
| Volatility (Midpoint Distance Travelled) * Average Volume | | | | | 0.0072*** | 0.0090*** | 0.0032** | 0.0030** |
| | | | | | (0.0006) | (0.0002) | (0.0013) | (0.0012) |
| Constant | 68,085*** | | −2,768 | | 28,891*** | | 5,464 | |
| | (9,796) | | (11,717) | | (8,771) | | (10,709) | |
| Observations | 43 | 43 | 43 | 43 | 43 | 43 | 43 | 43 |
| $R^2$ | 0.561 | 0.134 | 0.811 | 0.811 | 0.791 | 0.742 | 0.835 | 0.834 |

*p<0.1; **p<0.05; ***p<0.01

**Notes:** Please see the description in the text above this table and the notes for Table 7.1 in the main text. 5-minutes volatility is computed as realized 5-minute volatility for the FTSE 350 index in percentage points, using TRTH data. The distance travelled for each day is calculated as the sum of the absolute value of the change in midpoint on each update of the FTSE350. The FTSE350 is disseminated 4 times a second, or every 250 milliseconds.

# B Theory Appendix

This theory appendix covers three topics. First, discussion of equilibrium in the case where the firm providing liquidity is slow. Second, the analysis behind the bid-ask spread decomposition (5.3). Third, the algebra in support of equation (5.6) and its empirical counterpart (5.7), which express the proportional reduction of the cost of liquidity if latency arbitrage were eliminated.

## B.1 Equilibrium with Slow Liquidity Providers

In the equilibria of the continuous limit order book market studied in Budish, Cramton and Shim (2015), fast trading firms both engage in stale-quote sniping and provide all of the market's liquidity. There is a fringe of slow trading firms but they play no role in these equilibria (see especially Section VI.D and Proposition 3). The slow firms only play a role in equilibrium in Budish, Cramton and Shim (2015) under the frequent batch auctions market design.

In the BCS equilibria of the continuous market, fast trading firms are indifferent between liquidity provision and stale-quote sniping at the equilibrium bid-ask spread $s^{CLOB}$, characterized by

$$\lambda_{invest}\frac{s^{CLOB}}{2} = \lambda_{public}L(\frac{s^{CLOB}}{2}), \tag{B.1}$$

where $\lambda_{invest}$ denotes the arrival rate of investors (i.e., liquidity traders), $\lambda_{public}$ denotes the arrival rate of new public information, and $L(\frac{s^{CLOB}}{2}) \equiv \Pr(J \geq \frac{s^{CLOB}}{2})\mathbb{E}(J - \frac{s^{CLOB}}{2}|J \geq \frac{s^{CLOB}}{2})$ denotes the expected loss to a liquidity provider if there is a jump larger than their half-spread and they get sniped ($J$ is the random variable describing the absolute value of jump sizes). In the event of a jump larger than the half-spread, stale-quote snipers are successful $\frac{1}{N}$ of the time, where $N$ is the number of fast trading firms, and hence earn expected profits of $\frac{1}{N}\lambda_{public}L(\frac{s^{CLOB}}{2})$. A fast trading firm that provides liquidity earns revenues of $\lambda_{invest}\frac{s^{CLOB}}{2}$ from providing liquidity to investors, but, if there is a public jump, they get sniped with probability $\frac{N-1}{N}$, hence incurring costs of $\frac{N-1}{N}\lambda_{public}L(\frac{s^{CLOB}}{2})$. At the equilibrium spread, the revenue benefits of liquidity provision less these sniping costs net to the same $\frac{1}{N}\lambda_{public}L(\frac{s^{CLOB}}{2})$ earned by snipers. This net profit can be interpreted as the fast liquidity provider earning the opportunity cost of not sniping.

Under slightly different modeling formalities, introduced in Budish, Lee and Shim (2019), there also exist equilibria in which slow trading firms provide liquidity, at exactly the same bid-ask spread $\frac{s^{CLOB}}{2}$ characterized by (B.1), and the $N$ fast trading firms all engage in stale-quote sniping. The economic intuition for why this can also be an equilibrium is as follows. First, at this bid-ask spread, slow trading firms earn zero profits from liquidity provision, so slow trading firms are indifferent between liquidity provision here, and doing nothing as before. Second, with all $N$ fast trading firms now engaged in sniping, and the bid-ask spread the same as before, the fast trading firms all earn the same profits of $\frac{1}{N}\lambda_{public}L(\frac{s^{CLOB}}{2})$ as before. And, as before, at this bid-ask spread the fast trading firms are indifferent between providing liquidity or being one of $N-1$ snipers, so they do not strictly prefer to change from sniping to liquidity provision.

Formally, the configuration of play in which a slow trading firm provides liquidity at the spread

characterized by (B.1) (or its slight generalization to include adverse selection as well, presented as equation (5.2) in the main text) is an Order Book Equilibrium as defined in Budish, Lee and Shim (2019). The argument that this play constitutes an Order Book Equilibrium is as follows:

- If the slow TF deviates by widening their spread to $s' > s^{CLOB}$: another TF (whether slow or fast) can profitably undercut the deviation by providing liquidity at a better spread. Order Book Equilibrium requires that any deviation be robust to another TF providing better liquidity in response, so this potential deviation does not violate Order Book Equilibrium.

- If the slow TF deviates by narrowing their spread to $s' < s^{CLOB}$: they earn strictly negative profits as opposed to zero profits, so this is not a profitable deviation.

- If a fast TF undercuts the slow TF's spread to $s' < s^{CLOB}$: this is a profitable unilateral deviation for a fast TF for $s'$ close enough to $s^{CLOB}$, because the fast TF gets to both earn positive expected profits from liquidity provision, of just less than $\frac{1}{N}\lambda_{public}L(\frac{s^{CLOB}}{2})$, and potentially snipe the slow TF (the "have your cake and eat it too" deviation). However, the deviation is not robust to the slow TF canceling in response. Order Book Equilibrium requires that deviations are robust to other firms' responses with either cancels or price improvements ("no robust deviations").[52]

- If any other slow TF undercuts to $s' < s^{CLOB}$: this is not a profitable unilateral deviation for slow TFs, because $s^{CLOB}$ is the bid-ask spread at which slow TFs earn zero expected profits from liquidity provision. (The reason why providing liquidity at $s'$ close enough to $s^{CLOB}$ is profitable for a fast TF but not a slow TF is that fast TFs get sniped with probability $\frac{N-1}{N}$, whereas slow TFs get sniped with probability 1.)

Thus there exist order book equilibria in which fast TFs provide all liquidity as well as order book equilibria in which slow TFs provide all liquidity. It follows that there also exist order book equilibria in which, proportion $\rho_{fast} \in (0, 1)$ of the time, a fast TF provides liquidity at $s^{CLOB}$, while the remaining $1-\rho_{fast}$ of the time a slow TF provides liquidity at $s^{CLOB}$. Either way, the spread is the same, the profits of all fast TFs are the same ($\frac{1}{N}\lambda_{public}L(\frac{s^{CLOB}}{2})$), and the profits of all slow TFs are zero.

## B.2 Support for Bid-Ask Spread Decomposition (5.3)

Equation (5.3) in the main text provides a novel bid-ask spread decomposition that includes Price Impact both in and out of races, as well as a Loss Avoidance term for the case where a liquidity

[52]This case is the key technical difference between the modeling approach in Budish, Lee and Shim (2019) versus that in BCS. In the continuous-time game form considered in BCS a fast TF undercutting a slow TF in this way is a profitable deviation for the fast trading firm, because, in the small amount of time before a slow trading firm is able to respond to this deviation, the deviating fast trading firm both earns potential revenues from liquidity provision and earns potential profits from sniping the slow trading firm. In contrast, the Order Book Equilibrium concept introduced in Budish, Lee and Shim (2019) requires that the order book is at a resting point, where, if any one trading firm can profitably deviate from this resting point the deviation is no longer profitable after other trading firms respond with either price improvements or cancelations.

provider successfully cancels in a race. In this section we provide formal support for this decomposition.

Begin with the bid-ask spread characterization presented in the main text as (5.2),

$$\lambda_{invest}\frac{s^{CLOB}}{2} = (\lambda_{public} + \lambda_{private}) \cdot L(\frac{s^{CLOB}}{2}),$$

where $\lambda_{public}$ and $\lambda_{private}$ denote the arrival rate of public and private information, respectively, and $L(\frac{s^{CLOB}}{2})$ denotes the expected loss to a liquidity provider conditional on getting sniped or adversely selected. For simplicity, we assume that the jump size $J$ is identically distributed for public and private information, and that all jumps are of size of at least the equilibrium half-spread $\frac{s^{CLOB}}{2}$, so all jumps generate attempts to trade. These assumptions can be relaxed but at considerable notational burden.[53] With these assumptions, we have $L(\frac{s^{CLOB}}{2}) = E(J) - \frac{s^{CLOB}}{2}$.[54]

As discussed in the previous subsection, there exist equilibria in which only fast TFs provide liquidity, only slow TFs provide liquidity, and in which both fast and slow TFs provide liquidity. The former case was emphasized in BCS but the latter case appears to fit the data better. Let $\rho_{fast} \in [0, 1]$ denote the proportion of liquidity provided by fast TFs in equilibrium with the remaining $1 - \rho_{fast}$ provided by slow TFs. We can now formally define the terms utilized in equation (5.3).

- *EffectiveSpread* is equal to $[\lambda_{invest} + \lambda_{public}(1 - \frac{\rho_{fast}}{N}) + \lambda_{private}] \cdot \frac{s^{CLOB}}{2}$. Trade occurs whenever an investor arrives (at rate $\lambda_{invest}$), whenever an informed trader arrives ($\lambda_{private}$), and whenever there is public news ($\lambda_{public}$) and the race is won by a sniper: which occurs with probability $\frac{N-1}{N}$ if the TF providing liquidity is fast, where $N$ is the number of fast traders, and probability 1 if the TF providing liquidity is slow, hence total probability of $\rho_{fast}\frac{N-1}{N} + (1 - \rho_{fast}) = 1 - \frac{\rho_{fast}}{N}$.

- $PriceImpact_{Race}$ is equal to $\lambda_{public}(1 - \frac{\rho_{fast}}{N}) \cdot E(J)$: the $\lambda_{public}(1 - \frac{\rho_{fast}}{N})$ probability that a sniper wins a race, times the size of the jump $E(J)$, which will be the change in the midpoint. Using $L(\frac{s^{CLOB}}{2}) = E(J) - \frac{s^{CLOB}}{2}$ this can be rewritten as $\lambda_{public}(1 - \frac{\rho_{fast}}{N})E(J) = \lambda_{public}(1 - \frac{\rho_{fast}}{N})(\frac{s^{CLOB}}{2} + L(\frac{s^{CLOB}}{2}))$.

- $PriceImpact_{NonRace}$, by similar logic, is equal to $\lambda_{private}E(J)$: the $\lambda_{private}$ probability that there is an informed trader times the size of the jump $E(J)$, which will be the change in the midpoint. This can be rewritten as $\lambda_{private}E(J) = \lambda_{private}(\frac{s^{CLOB}}{2} + L(\frac{s^{CLOB}}{2}))$.

[53] Formally, if $J_{private}$ and $J_{public}$ are, respectively, the jump distributions for private and public information, with cumulative distribution functions $F_{private}(x)$ and $F_{public}(x)$, respectively, then the conditional distributions of interest are $J^*_{private}$ and $J^*_{public}$ with cdf's $F^*_{private}(x) = \frac{F_{private}(x) - F^-_{private}(\frac{s^{CLOB}}{2})}{1 - F^-_{private}(\frac{s^{CLOB}}{2})}$ and $F^*_{public}(x) = \frac{F_{public}(x) - F^-_{public}(\frac{s^{CLOB}}{2})}{1 - F^-_{public}(\frac{s^{CLOB}}{2})}$, respectively, for $x \geq \frac{s^{CLOB}}{2}$ and $F^*_{private}(x) = F^*_{public}(x) = 0$ for $x < \frac{s^{CLOB}}{2}$.

[54] In the generalization described in the previous footnote the appropriate formulas to use are $L_{private}(\frac{s^{CLOB}}{2}) \equiv E(J^*_{private}) - \frac{s^{CLOB}}{2}$ and $L_{public}(\frac{s^{CLOB}}{2}) \equiv E(J^*_{public}) - \frac{s^{CLOB}}{2}$. In the mathematics that follows it is then convenient to define $\lambda^*{}_{public} = \lambda_{public}(1 - F^-_{public}(\frac{s^{CLOB}}{2}))$ and $\lambda^*_{private} = \lambda_{private}(1 - F^-_{private}(\frac{s^{CLOB}}{2}))$ as the arrival rates of jumps that are larger than the equilibrium spread.

- *LossAvoidance* is equal to $\lambda_{public}\frac{\rho_{fast}}{N}L(\frac{s^{CLOB}}{2})$: the $\lambda_{public}\frac{\rho_{fast}}{N}$ probability that a fast liquidity provider wins a race with a cancel, times the size of the avoided loss $L(\frac{s^{CLOB}}{2})$.

Now take the equilibrium bid-ask spread as characterized in equation (5.2),

$$\lambda_{invest}\frac{s^{CLOB}}{2} = (\lambda_{public} + \lambda_{private}) \cdot L(\frac{s^{CLOB}}{2}),$$

and add $(\lambda_{public}(1 - \frac{\rho_{fast}}{N}) + \lambda_{private}) \cdot \frac{s^{CLOB}}{2}$ to both sides of the equation. This yields

$$\begin{aligned}&\left(\lambda_{invest} + \lambda_{public}(1 - \frac{\rho_{fast}}{N}) + \lambda_{private}\right) \cdot \frac{s^{CLOB}}{2} \\ &\quad = \left(\lambda_{public}(1 - \frac{\rho_{fast}}{N}) + \lambda_{private}\right) \cdot \left(\frac{s^{CLOB}}{2} + L(\frac{s^{CLOB}}{2})\right) + \lambda_{public}\frac{\rho_{fast}}{N}L(\frac{s^{CLOB}}{2}).\end{aligned}$$

If we substitute in terms as defined above, this in turn yields

$$\textit{EffectiveSpread} = \textit{PriceImpact}_{Race} + \textit{PriceImpact}_{NonRace} + \textit{LossAvoidance}.$$

We follow the spread decomposition literature and include *RealizedSpread* as the residual in this equation for the purpose of bringing it to data, yielding equation (5.3) in the text:

$$\textit{EffectiveSpread} = \textit{PriceImpact}_{Race} + \textit{PriceImpact}_{NonRace} + \textit{LossAvoidance} + \textit{RealizedSpread}.$$

## B.3 Support for the Proportional Reduction in Cost of Liquidity Equations (5.6)-(5.7)

We start with equation (5.4) in the main text, which defines this proportional reduction theoretically:

$$\frac{\frac{s^{CLOB}}{2} - \frac{s^{FBA}}{2}}{\frac{s^{CLOB}}{2}}$$

where $s^{CLOB}$ denotes the equilibrium bid-ask spread in the continuous limit order book market, and $s^{FBA}$ denotes the equilibrium bid-ask spread in the frequent batch auctions market, which eliminates sniping. Next, multiply both the numerator and denominator by $(\lambda_{invest} + \lambda_{private})$:

$$\frac{(\lambda_{invest} + \lambda_{private})(\frac{s^{CLOB}}{2} - \frac{s^{FBA}}{2})}{(\lambda_{invest} + \lambda_{private})\frac{s^{CLOB}}{2}}$$

Next, use the bid-ask spread characterization (5.2) in the main text to solve out for $\lambda_{invest}\frac{s^{CLOB}}{2}$

in the numerator:

$$\frac{(\lambda_{public}+\lambda_{private})\cdot L(\frac{s^{CLOB}}{2})+\lambda_{private}\frac{s^{CLOB}}{2}-(\lambda_{invest}+\lambda_{private})(\frac{s^{FBA}}{2})}{(\lambda_{invest}+\lambda_{private})\frac{s^{CLOB}}{2}}$$

Analogously, use equation (5.1) of Budish, Lee and Shim (2019) to solve out for $\lambda_{invest}\frac{s^{FBA}}{2}$ in the numerator:

$$\frac{(\lambda_{public}+\lambda_{private})\cdot L(\frac{s^{CLOB}}{2})+\lambda_{private}\frac{s^{CLOB}}{2}-\lambda_{private}L(\frac{s^{FBA}}{2})-\lambda_{private}(\frac{s^{FBA}}{2})}{(\lambda_{invest}+\lambda_{private})\frac{s^{CLOB}}{2}}$$

Next, regroup terms to place $\lambda_{public}\cdot L(\frac{s^{CLOB}}{2})$ on the left of the numerator, and then utilize $L(\frac{s}{2})=E(J)-\frac{s}{2}$ for $\lambda_{private}L(\frac{s^{CLOB}}{2})$ and $\lambda_{private}L(\frac{s^{FBA}}{2})$:

$$\frac{\lambda_{public}\cdot L(\frac{s^{CLOB}}{2})+\lambda_{private}(E(J)-\frac{s^{CLOB}}{2})+\lambda_{private}\frac{s^{CLOB}}{2}-\lambda_{private}(E(J)-\frac{s^{FBA}}{2})-\lambda_{private}(\frac{s^{FBA}}{2})}{(\lambda_{invest}+\lambda_{private})\frac{s^{CLOB}}{2}}$$

Observe that most of the terms in the numerator cancel. Specifically, we have $\lambda_{private}(E(J)-\frac{s^{CLOB}}{2})+\lambda_{private}\frac{s^{CLOB}}{2}-\lambda_{private}(E(J)-\frac{s^{FBA}}{2})-\lambda_{private}(\frac{s^{FBA}}{2})=0$. This leaves us with:

$$\frac{\lambda_{public}\cdot L(\frac{s^{CLOB}}{2})}{(\lambda_{invest}+\lambda_{private})\frac{s^{CLOB}}{2}}$$

as claimed in the text as equation (5.6). Equation (5.6)'s empirical implementation, equation (5.7), then follows immediately as described in the main text.