QUANTIFYING THE HIGH-FREQUENCY TRADING "ARMS RACE"*

MATTEO AQUILINA
ERIC BUDISH
PETER O'NEILL

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 remarkably large portion of overall trading volume (about 20%). Race participation is concentrated, with the top six firms accounting for over 80% of all race wins and losses. The average race is worth just a small amount (about half a price tick), but because of the large volumes the stakes add up. Our main estimates suggest that races constitute roughly one-third of price impact and the effective spread (key microstructure measures of the cost of liquidity), that latency arbitrage imposes a roughly 0.5 basis point tax on trading, that market designs that eliminate latency

*We thank Andrew Bailey, Markus Baldauf, Fabio Braga, Peter Cramton, Karen Croxson, Sean Foley, Thierry Foucault, Joel Hasbrouck, Terrence Hendershott, Burton Hollifield, Stefan Hunt, Anil Kashyap, Pete Kyle, Robin Lee, Donald MacKenzie, Albert Menkveld, Paul Milgrom, Barry Munson, Brent Neiman, Lubos Pastor, Talis Putnins, Alvin Roth, Edwin Schooling Latter, Makoto Seta, John Shim, and Mao Ye for helpful discussions. We thank the coeditor Andrei Shleifer and four anonymous referees for valuable comments and advice. We are 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 article circulated in January 2020 as a Financial Conduct Authority (FCA) Occasional Paper, which is the FCA's working paper series. 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. The views expressed in this article are those of the authors and not necessarily those of the Financial Stability Board and/or its members. Any errors and omissions are the authors' own. The authors declare that they have no relevant or material financial interests that relate to this research.

The Quarterly Journal of Economics (2022), 493–564. https://doi.org/10.1093/qje/qjab032.
Advance Access publication on September 10, 2021.

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 per year in global equity markets alone.
JEL Codes: D47, G10, G12, G14

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

"Widespread latency arbitrage is a myth." —Bill Harts, CEO of the Modern Markets Initiative, a high-frequency trading (HFT) lobbyist (Michaels 2016)

I. Introduction

As recently as the 1990s and early 2000s, human beings on trading floors, pits, and desks intermediated the large majority of financial market transactions. Now, financial markets across most major asset classes—equities, futures, treasuries, currencies, options, and so on—are almost entirely electronic. This transformation of financial markets from the human era to the modern electronic era on the whole has brought clear, measurable improvements to various measures of the cost of trading and liquidity, much as information technology has brought efficiencies to many other sectors of the economy. But this transformation has also brought considerable controversy, particularly around the importance of speed in modern electronic markets.[1]

At the center of the controversy over speed is a phenomenon called latency arbitrage, also known as "sniping" or "picking off" stale quotes. In plain English, a latency arbitrage is an arbitrage opportunity that is sufficiently mechanical and obvious that capturing it is primarily a contest in speed. For example, if the price of the S&P 500 futures contract changes by a large enough amount in Chicago, there is a race around the world to pick off stale quotes in every asset highly correlated to the S&P 500 index: S&P 500 exchange traded funds (ETFs), other U.S. equity index futures

1. See MacKenzie (2021) for a history of the transformation of financial markets from the human-trading era to the modern electronic era, with detailed documentation of many different aspects of the speed race and numerous additional references. See Hendershott, Jones, and Menkveld (2011), Angel, Harris, and Spatt (2015), and Frazzini, Israel, and Moskowitz (2018) for key studies on the decline of trading costs. Most of the declines appear to be concentrated in the early years of the transformation, specifically the mid- to late 1990s and early to mid-2000s; see especially Figure 20 of Angel, Harris, and Spatt (2015) and Figure IA.8 of Hagströmer (2021). See Jones (2013), Biais and Foucault (2014), O'Hara (2015), and Menkveld (2016) for surveys of the literature on high-frequency trading.

and ETFs, global equity index futures and ETFs, and so on. Many other examples arise from other sets of highly correlated assets: Treasury bonds of slightly different durations, or in the cash market versus the futures market; options and the underlying stock; ETFs and their largest component stocks; currency triangles; commodities at different delivery dates; and many others. Perhaps the simplest example is if the exact same asset trades in many different venues. For example, in the U.S. stock market, there are 16 different exchanges and 50+ alternative trading venues, all trading the same stocks—so if the price of a stock changes by enough on one venue, there is a race to pick off stale quotes on all the others. These races around the world involve microwave links between market centers, transoceanic fiber-optic cables, putting trading algorithms onto hardware as opposed to software, colocation rights and proprietary data feeds from exchanges, real estate adjacent to and even on the rooftops of exchanges, and, perhaps most importantly, high-quality human capital. Just a decade ago, the speed race was commonly measured in milliseconds (thousandths of a second); it is now measured in microseconds (millionths) and even nanoseconds (billionths).[2]

In theoretical terms, Budish, Cramton, and Shim (2015, BCS) define latency arbitrage as arbitrage rents from symmetrically observable public information signals, as distinct from the asymmetrically observable private information signals at the heart of classic models of market microstructure (Kyle 1985; Glosten and Milgrom 1985). 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 in modern electronic markets, even information seen and understood by many market participants essentially simultaneously creates arbitrage rents—because of the underlying market design used by modern financial exchanges. The issue is 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, 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

2. Please see the working paper version of this article, Aquilina, Budish, and O'Neill (2020), for detailed references. The latest salvo, reported April 1, 2021, in what at first seemed like an April Fools joke, is dedicated satellite links between market centers in North America, Europe, and Asia (Osipovich 2021).

they do to a human. Even in the logical extreme in which many market participants observe a new piece of information at *exactly* the same time, and respond with *exactly* the same technology, somebody gets a rent. BCS showed that these arbitrage rents lead to a socially wasteful 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—any firm providing liquidity has to bear the cost of getting sniped. A subtle change to the market design can eliminate the rents—preserving the useful functions of modern 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 £4.2 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 United States 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 noise in timestamps and physical limitations due to the speed at which information travels. Wah (2016) and Ring 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 Ring 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.[3]

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 market design reforms, such as frequent batch auctions, that address it. If the magnitudes are sufficiently large, then Lewis's claim that the market is "rigged for the benefit of insiders" is reasonable if perhaps a bit conspiratorial. Conversely, if the magnitudes are sufficiently small, then the high-frequency trading (HFT) lobby's claim that latency arbitrage is a myth is reasonable if perhaps a bit exaggerated. Notably, while numerous regulators around the world have investigated HFT in some capacity (e.g., the FCA, ESMA, SEC, CFTC, U.S. Treasury, New York attorney general), and in a few specific cases have been required to rule specifically on speed bump proposals designed to address latency arbitrage, there is not a broad consensus on what if any regulatory rules or interventions are appropriate.[4]

This article 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, distinct from widely familiar limit order book data sets such as exchange direct feeds, or consolidated data sets like TAQ (Trades and Quotes) or the SEC's MIDAS data set. 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, and so on—often with ultraprecise timestamps. But what is missing are the messages that do not affect the state of the order book, because they fail.[5]

3. Other papers with empirical evidence that relates to the benefits and costs of HFT include Menkveld (2013), Brogaard et al. (2015, 2018), Foucault, Kozhan, and Tham (2016), Shkilko and Sokolov (2020), Malinova, Park, and Riordan (2018), Weller (2018), Van Kervel and Menkveld (2019), Breckenfelder (2019), and Indriawan, Pascual, and Shkilko (2020).

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 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 (academic, regulatory, or industry) to use exchange message data, defined as the full back-and-forth message traffic between market participants and the exchange. There have been several other studies of HFT with data that goes beyond TAQ or exchange direct feeds, such as

For example, if a market participant seeks to snipe a stale quote but fails—their immediate or cancel (IOC) order is unable to execute immediately 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 Service and Markets Act) all message activity for all stocks in the FTSE 350 index for a nine-week period in fall 2015.[6] The messages are time-stamped with accuracy to the microsecond (one-millionth of a second), and as we 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 these 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, and so on. And, by comparing the price in the race to the prevailing market price a short time later, we can measure the economic stakes, that is, how much was it worth to win.

---

participant identifiers or additional information on order types, but none, to our knowledge, with the complete call-and-response data that are key to our study.

6. The FTSE 350 is an index of the 350 highest capitalization stocks in the United Kingdom. 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).

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 remarkably large proportion of daily trading volume is in races.* For the FTSE 100 index, about 22% of trading volume and 21% of trades are in races. Cochrane (2016) describes that trading volume is "The Great Unsolved Problem of Financial Economics."[7] Our results suggest that latency arbitrage is a meaningful piece of the puzzle. Indeed, in our most inclusive sensitivity scenario, with an up to 3 millisecond race window, races constitute 44% of all FTSE 100 trading volume.
- *Races are worth just small amounts each.* The average race is worth a bit more than half a tick, which on average comes to about £2. Even at the 90th percentile of races, the races are worth just 3 ticks and about £7. There is also a fair amount of noise: about 20% of races have strictly negative profits one second ex post.[8]
- *Race participation is concentrated. The top firms disproportionately snipe.* The top three firms win about 55% of races and lose about 66% of races. For the top six firms, the figures are 82% and 87%. In addition to documenting concentration, we also find that the top six firms are disproportionately aggressive in races, taking about 80% of liquidity in races while providing about 42% of the

7. See also Hong and Stein (2007) who write that "Some of the most interesting empirical patterns in the stock market are linked to volume" and provide numerous additional references.

8. Robert Mercer, former co-CEO of the quantitative trading firm Renaissance Technologies, described quantitative investing as "We're right 50.75 percent of the time ... you can make billions that way" (Zuckerman 2019, 272). Similarly, HFT firm Virtu's CEO Doug Cifu indicated that around 51%–52% of their trades are profitable (Mamudi 2014). Our figures suggest that trading in races is closer to pure arbitrage than 51/49 but still a healthy distance from 100/0.

liquidity that gets taken in races. Market participants outside the top six firms take about 20% of liquidity in races while providing about 58%. Thus, on net, much race activity consists of firms in the top six taking liquidity from market participants outside of the top six. This taking is especially concentrated in a subset of four of the top six firms who account for a disproportionate share of stale-quote sniping relative to liquidity provision.[9]

- *In aggregate, these small races add up to a significant proportion of price impact and the effective spread, key microstructure measures of the cost of liquidity.* We augment the traditional bid-ask spread decomposition suggested by Glosten (1987), which is widely used in the microstructure literature (e.g., Glosten and Harris 1988; Hasbrouck 1991a, 1991b; Stoll 2000; Hendershott, Jones, and Menkveld 2011), to separately incorporate price impact from latency arbitrage races and nonrace trading. Price impact from trading in races is about 31% of all price impact and about 33% of the effective spread. This suggests that latency arbitrage deserves a place alongside traditional adverse selection as one of the primary components of the cost of liquidity.[10]
- *Market designs that eliminate latency arbitrage could meaningfully reduce the market's cost of liquidity.* We

9. In the equilibria studied in BCS, fast trading firms provided all liquidity. Races consisted of some fast trading firms trying to snipe and other fast trading firms trying to cancel. In Online Appendix F, we show that there exists another equilibrium of the BCS model in which both fast and slow firms provide liquidity. If a slow firm provides liquidity and there is a race, they get sniped with probability 1. The key insight is that the same bid-ask spread that leaves fast trading firms indifferent between liquidity provision and stale-quote sniping (either way, earning $\frac{1}{N}$ of the sniping prize, where N is the number of fast firms in the race) is the zero-profit spread for slow trading firms.

10. There are many different strands of literature on the broader importance of liquidity for financial markets. One strand explores the connection between the specific kinds of microstructure measures of liquidity we study and asset pricing—good starting points include Amihud (2002), Pástor and Stambaugh (2003), and Acharya and Pedersen (2005). This literature finds that liquidity is a factor in asset pricing returns, which in turn implies that reforms that improve the market's liquidity reduce required equilibrium returns and hence increase the level of asset prices. Diamond and Dybvig (1983) and a large subsequent literature highlight the role of liquidity in reducing the likelihood of bank runs. Shleifer and Vishny (1992, 1997, 2011), Brunnermeier and Pedersen (2009), Hanson, Kashyap, and Stein (2011), and many others have studied connections between liquidity and various aspects of financial stability.

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 nonrace 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. This implies that market designs that eliminate latency arbitrage, such as frequent batch auctions, would reduce investors' cost of liquidity by 17%. As a complementary analysis, we show that the liquidity provider's realized spread in races is significantly negative, whereas it is modestly positive in nonrace 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.[11]

- *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 volume is robust, with an R^2 of about 0.81, and indeed, the latency arbitrage tax on trading volume is roughly constant in our data. Adding daily volatility to the relationship further improves the fit, albeit only slightly. Using these relationships, we find that the annual sums at stake in latency arbitrage races in the United Kingdom are about £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.[12]

11. Market design research often involves a mix of economic theory, empirical evidence, and institutional detail working together to help bring useful economic ideas from theory to practice. Roth (2002) has called this "the economist as engineer." Other examples from outside of finance include the design of matching markets (Roth 2008), spectrum auctions (Milgrom 2021), kidney exchange mechanisms (Roth, Sönmez, and Ünver 2004), school choice procedures (Pathak 2017), course allocation procedures (Budish et al. 2017), and accelerating COVID-19 vaccination (Castillo et al. 2021). See Kominers, Teytelboym, and Crawford (2017) and Roth (2018) for recent surveys.

12. Over a variety of sensitivity analyses, with race windows ranging from 50 microseconds to 3 milliseconds, our range of estimates is $2.3–$8.4 billion per

I.A 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 because of the large volumes, 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 asset classes that trade on electronic limit order books such as futures, treasuries, currencies, and options.

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 drives a significant fraction of daily trading volume, 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.

I.B Organization of the Article

The remainder of this article is organized as follows. Section II describes the message data in detail. Section III describes our methodology for detecting and measuring latency arbitrage races. Section IV presents the main results. Section V discusses sensitivity analyses and robustness checks. Section VI extrapolates to an annual size of the prize for the UK and global equity markets. Section VII concludes.

II. Message Data

The novel aspect of our data is that it includes all messages sent by participants to the exchange and by the exchange back

year in global equities markets. In 2020, which was a particularly high-volume and high-volatility year due to the COVID-19 pandemic, our point estimate is $7 billion and our range is $3.1–$11.4 billion. We discuss caveats for this extrapolation exercise in detail in Section VI.C.

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 describe in Section II.B, 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.[13]

We obtained these message data from the London Stock Exchange (LSE), 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 August 17 to October 16, 2015, for all stocks in the FTSE 350 index. We drop one day (September 7) 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.

II.A. Overview of a Modern Stock Exchange

The continuous limit order book is at heart a simple protocol.[14] We guess that most undergraduate computer science

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

14. We assume most readers are already familiar with the basics of a limit order book market, but here is a quick refresher. The basic building block is a limit order, which consists of a symbol, price, quantity and direction. 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 cancel it). Trades occur 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. If there are multiple outstanding orders which 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

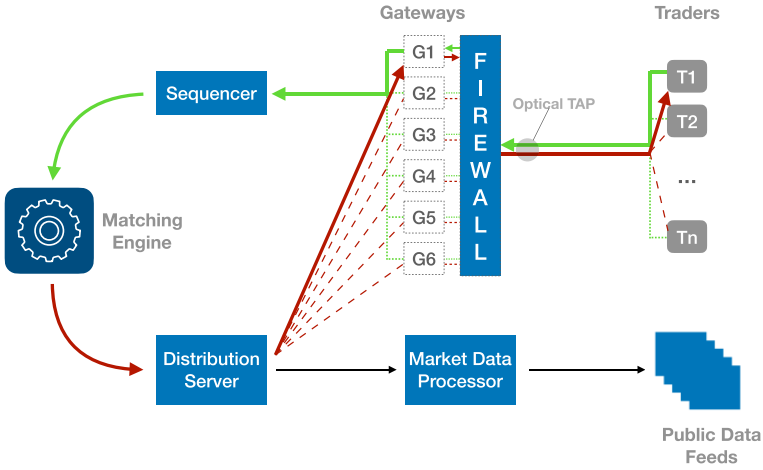

FIGURE I

Exchange Schematic

See the text of Section II.A for a discussion of the overall system architecture depicted in this figure. The solid lines depict a sample message path, for a message sent by participant T1 inbound to the matching engine (light green; color version available online), and then sent by the matching engine outbound back to the participant (dark red). The light green dotted lines depict other sample inbound message paths, and the dark red dashed lines depict other sample outbound message paths. The thin black solid lines at the bottom depict outbound messages sent to public data feeds as distinct from the outbound messages sent to individual market participants used in our study. The circle labeled Optical TAP represents the optical traffic analysis point where our message data are recorded and timestamped. See Section II.B for a discussion of timestamping.

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 to process large and time-varying quantities of messages with extremely low latency and essentially zero system downtime.

In this subsection, we provide a stylized description of a modern electronic exchange, illustrated in Figure I. We do this both because it is a necessary input for understanding our data, and because we expect it will be useful to academic researchers and regulators who seek a better understanding of the detailed plumbing of modern financial markets.

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.

The core of a modern exchange, 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 a modern exchange, and market participants do not even interact with the matching engine directly, in either direction. Rather, market participants send messages to the exchange via what are known as gateways, which verify the integrity of messages, perform risk checks, and translate messages from the participant interface language into a language optimized for the matching engine.[15] Gateways in turn pass messages on to a sequencer, which in essence translates input from many parallel gateways into, for each symbol, a single sequence of messages that is passed on to the matching engine. The matching engine, once it does its work, 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.

A fuller description of these components is in the working paper version of this article (Aquilina, Budish, and O'Neill 2020). Here, we briefly emphasize the overall rationale for this system

15. There is intrinsically a small amount of randomness in this piece of the systems architecture, because how long a particular gateway takes to process a particular message is stochastic. This randomness will manifest in our data below in Figure II. We did not find any evidence in our data of firms attempting to exploit this randomness, for example, by sending the same message to multiple gateways via multiple accounts. Our best guess as to why is this behavior would be easy for the LSE to detect.

architecture. Given the limit order book market design, the matching engine must 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.[16] Therefore, it is critical for latency to take as much of the work as possible "off the shoulders" of the matching engine and put it on to other components of the system.

II.B. Where and How Messages Are Recorded and Timestamped

As described, participants send messages to the exchange and receive messages from the exchange via gateways. Between the participants' own systems and the exchange gateways 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); refer to Figure I. 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 this point 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 or 0.0000001 second) precision, at this point of capture, by a hardware clock. Importantly, all messages are timestamped by a single clock. Therefore, although 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. Based on our discussions with the LSE, we are comfortable treating our data as accurate to the microsecond.

Please note that the optical TAP timestamps we observe in our data are not seen by market participants.

16. 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).

II.C. Translating Message Data into Market Events

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 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 the quantity/price traded and the quantity/price that remains and is posted to the book.

An important piece of our code is to classify sets of such messages into what we call market events—for instance, a "new order—executed in full" event or a "resting order—passive execution" event. We first describe the contents of inbound and outbound messages and then describe how we classify messages into market events. For more complete details, please see the Online Appendix.

1. *Inbound Messages.* Each inbound message contains the following kinds of information:

- 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. In addition, if the message is a cancel or modification of an existing order, then the message contains identifying information for the existing order.
- 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.
- Price/Quantity/Side Information. Last, if a message is a new order or a modification of an existing order, it will indicate the price, quantity, and direction (buy/sell).

2. *Outbound Messages.* Each outbound message contains the following kinds of information:

- Identifiers. These fields typically contain 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.

- Message Outcome Information. Outbound messages contain several fields that provide information on the outcome of an inbound message just submitted. One field reports on what type of activity the matching engine just executed: for instance, a post to the book, a trade, or a failed immediate-or-cancel request. A second field indicates the current status of the order: the main status options are new, filled, partially filled, canceled, and expired. A third field specifically allows us to see if a cancel request failed; failed cancels require a special treatment because the order the user tried to cancel no longer exists in the matching engine's state.
- Trade Execution Reports. If a new order results in a trade, outbound messages will be sent to both parties in the trade with trade execution reports detailing the price, quantity, and side. If an order matches with multiple counterparties or at multiple prices, there will be a separate pair of outbound messages 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.

3. *Event Classification.* Combinations of inbound and outbound messages indicate market events, as listed in Table I. To perform this classification, our code loops through all messages sequentially, and at each inbound message loops ahead to find all related outbound messages (using the information from both the participant-supplied and matching-engine-supplied identifiers), to classify events as listed in the table. For complete details of this key piece of code, please see the Online Appendix.

For all events other than passive fills, we define the time of the event based on the time of the inbound message timestamp; this timestamp is what will be relevant for race detection. For passive fills, we define the time of the event based on the time of the outbound message; this information is not related to race detection per se but will help us maintain the order book as discussed next.

4. *Maintaining the Order Book.* Observe that neither inbound nor outbound messages contain the state of the limit

TABLE I

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 II.C.1 for a description of the contents of inbound messages, Section II.C.2 for a description of the contents of outbound messages, and Section II.C.3 for a description of event classification.

order book—that is, 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, although convenient, would mean larger and hence slower messages. We thus have to build and maintain the state of the limit order book ourselves.

We maintain the state of the limit order book, for each symbol-date, on outbound messages. We use outbound messages rather than inbounds because outbound messages report what the matching engine actually did. Whenever we compute race statistics that rely on the order book, we use the state of the order book as of the first inbound message in the race. There are a few technical details related to maintaining the order book with message data that we discuss in more detail in the Online Appendix, along with a discussion of robustness checks.

III. Defining and Measuring Latency Arbitrage Races

The theory in BCS,[17] and a modest extension we include as Online Appendix F.1, suggest that the empirical signature of a latency arbitrage race in response to public information, as distinct from Kyle-Glosten-Milgrom-style informed trading based on private information, is:

i. Multiple market participants acting on the same symbol, price, and side;
ii. Either a mix of take attempts and cancel attempts (equilibrium emphasized in BCS), or all take attempts (if the liquidity provider is slow, see Online Appendix F.1);
iii. Some succeed, some fail;
iv. All at the "same time."

Of these, characteristics i–iii are relatively straightforward to define and implement. 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.

Characteristic iv is conceptually more difficult. In a theory model there is such a thing as the "same time" but in data no two things happen at exactly the same time, if time is measured finely enough. We structure the analysis so that the baseline method

17. Please see Section 4.1 of the working paper version of this article for a brief review of the relevant theory.

is conservative and then consider a wide range of sensitivity analyses.

We emphasize that throughout, when we describe either actions or timestamps, we refer to the *inbound* messages and timestamps, enhanced with the event classification information described in Section II.C 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.

III.A. *Characteristic i: Multiple Market Participants Act on the Same Symbol, Price, and Side*

The "same symbol, price, and side" aspect is straightforward. Every limit order message (including IOCs) includes the symbol, price, and side of the order. We interpret a limit or IOC order to buy at p as relevant to any potential 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.[18]

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 in the same firm as potentially distinct competitors in a latency-sensitive trading opportunity.

In sensitivity analyses, we also consider larger minimum requirements for the number of participants in the race, especially 3+, and requiring that the FirmIDs are unique, not just UserIDs.

III.B. *Characteristic ii: Either a Mix of Take and Cancel Attempts, or All Take Attempts*

For our baseline, we require that a race consist of either a mix of take and cancel attempts (i.e., 1+ aggressors and 1+ cancelers) or all take attempts (i.e., 2+ aggressors and 0 cancelers).

18. 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.

In sensitivity analyses, we consider requiring both an aggressor and a canceler (that is, excluding races with 2+ aggressors and 0 cancelers), and requiring 2+ aggressors.

III.C. Characteristic iii: Some Succeed, Some Fail

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

1. *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.

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 $\geqslant p$ or an order to sell at $\leqslant p$) but obtains zero quantity at p. Although 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 postrace order book. Second, at the LSE, there was a small (£0.01 per message) fee advantage to using plain-vanilla limit orders instead of IOC orders. 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.[19]

In sensitivity analysis, we consider an alternative in which only failed IOCs and failed cancel attempts count as fails and plain-vanilla limit orders cannot count as fails. This sensitivity reflects the possibility that a limit order that obtains zero

19. At the time of our data and as of this writing, the LSE assessed an "order management charge" of £0.01 for nonpersistent orders such as IOCs, whereas there was no order management charge for plain-vanilla limit orders (London Stock Exchange Group 2015). 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 per message, so the synthetic IOC would be nearly twice as expensive as an IOC (London Stock Exchange Group 2015). Our understanding is that triggering this surcharge is rare.

quantity at p and instead posts to the book may represent postrace liquidity provision reflecting the postrace value, as opposed to a failed attempt to snipe. We emphasize this alternative, which we refer to as strict fail, especially in sensitivity analyses with longer time horizons where we are more concerned about the postrace liquidity provision issue.

2. *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 $\geqslant p$ or an order to sell at $\leqslant p$) and obtains positive quantity at a price p or better (i.e., it buys positive quantity at a price $\leqslant p$ or sells positive quantity at a price of $\geqslant p$). 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—that is, 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.

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 because 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.

In sensitivity analysis, we consider requiring proof that 100% of depth at the race price is successfully cleared in the race. This can be satisfied in three ways: observing a failed IOC at the race price p; observing a limit order at the race price p that posts to the book at least in part; or observing quantity traded plus quantity canceled of 100% of the displayed depth at the start of the race.

III.D. Characteristic iv: All at the "Same Time"

Of the four characteristics, this last one is conceptually the hardest. In a theory model there can be a precise distinction between simultaneous and nonsimultaneous actions, but in 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.[20]

We consider two different approaches to this issue.

1. *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:

$$\begin{aligned}&\textit{Information Horizon}\\&= \textit{Actual Observed Latency} : \textit{M1 In} \rightarrow \textit{M1 Out}\\&+ \textit{Minimum Observed Reaction Time} : \textit{M1 Out} \rightarrow \textit{M2 In},\end{aligned}$$

where *M1* refers to the first message in a race; *M2* refers to the second message in the race; *Actual Observed Latency M1 In* $\rightarrow$ *M1 Out* 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 Out* $\rightarrow$ *M2 In* 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 timestamp to the response's inbound message timestamp.

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

20. 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 MacKenzie (2019).

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,[21] and the median *Actual Observed Latency* is about 150 microseconds (90th percentile: about 300 microseconds). We provide further details in Online Appendix A. 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 these parameters to be potentially different for different exchanges or different periods in time.

2. *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, 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 conduct this analysis for T ranging from 50 microseconds to 3 milliseconds. A summary of the results are presented in Section V.A with full details in Online Appendix C.1. What T's would be of interest we would expect to evolve over time as technology evolves.

21. This reaction time of 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 composed of about 17 microseconds from the first component and about 12 microseconds from the second component, as we describe in Online Appendix A.

3. *A Note on Code Structure and Overlapping Races.* If 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 (in the baseline and sensitivities, respectively). That is, we do not allow for "overlapping" races at a single price level.

Relatedly, in the event of a latency arbitrage race that occurs across multiple levels of the book (e.g., in the event of a large change in public information about the value of an asset), 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 multilevel 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'.

IV. Main Results

This section presents all of our main results under the baseline specification as described in Section III. In Section V we discuss various alternative specifications and sensitivity analyses. Section IV.A presents results on race frequency, duration, and trading volume. Section IV.B presents results on race participation patterns. Section IV.C presents results on profits per race. Section IV.D presents results on aggregate profits and the latency arbitrage tax. Section IV.E presents two spread decompositions that explore what proportion of the cost of liquidity is the latency arbitrage component versus the traditional adverse selection component.

IV.A. Frequency and Duration of Latency Arbitrage Races

1. *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 every seven 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 II, 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 II, Panel B.

TABLE II

RACES PER DAY

Panel A: Number of races per day across symbols

Description	Mean	Std. dev.	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	Std. dev.	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 III for a detailed description of the baseline race detection criteria and Section II 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.

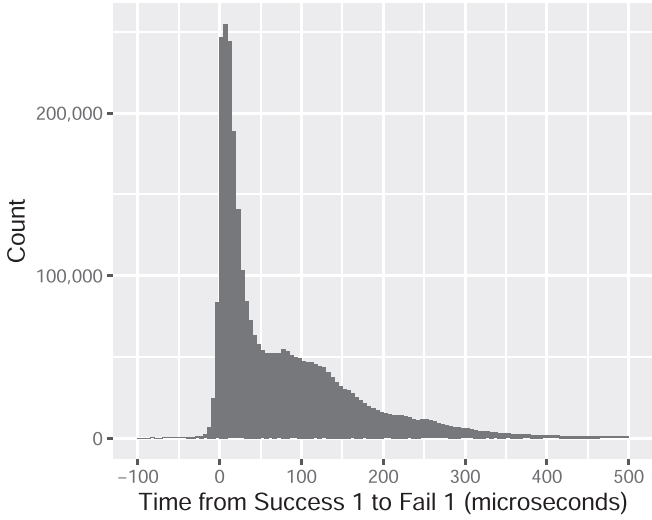

FIGURE II

Duration of Races

For each race detected by our baseline method, 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 III.C). Denote these messages S1 and F1, respectively. The figure plots the distribution of F1's timestamp minus S1's timestamp in microseconds, that is, by how long the first successful message in the race beat the first failed message. The histogram has a bin size of 5 microseconds.

2. *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. Figure II depicts the distribution of race durations. The mode of the distribution is 5–10 microseconds, and the median is 46 microseconds. There is then steady mass in the distribution up until about 150 microseconds, the 90th percentile is about 200 microseconds, and there is a tail up to our truncation point of 500 microseconds. Online Appendix Table B.1 provides additional details on the distribution.

3. *Sometimes the "Wrong" Message Wins.* Interestingly, in Figure II, 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 II.B that our timestamps are obtained at the outer wall of the exchange's system. It is therefore possible that if two race messages arrive to different gateways at nearly

the same time, 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.[22]

4. *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. 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).[23] 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. See Table III.

IV.B. Race Participation

1. *Number of Participants.* Table IV, Panel A provides data on the number of participants in races. Because 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

22. Please also see a recent essay by MacKenzie (2019) on various aspects of randomness in HFT races.

23. We compute daily trading volume in our data by obtaining all outbound messages during regular hours that are aggressive fills—that is, that report a trade execution to a just-received inbound message that aggressed against a previous resting order. In the event classification table (Table I), these are the events called "new order aggressively executed in full" and "new order aggressively executed in part." We count just the aggressive side of the trade to prevent double counting.

TABLE III
VOLUME AND TRADES IN RACES

Description	Mean	Std. dev.	Min	Pct10	Pct25	Median	Pct75	Pct90	Max
Panel A: Percentage of volume (value weighted) in races across dates									
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									
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 data set, we obtain all outbound messages in regular-hours trading that are aggressive fills (see note 23 for more detail). We then obtain the inbound message associated with each such outbound aggressive fill and check whether the inbound is part of a race (as defined in Section III). For Panel A, for each date, we 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 a race. For Panel B, for each date, we 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 a race.

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.

2. *Number of Takes and Cancels.* Table IV, Panels B and C 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 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 Online Appendix F.1 for theoretical discussion of these equilibria). We return to this pattern shortly.

Of these 3.07 take attempts, the large majority, 2.81, are 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 Online Appendix Table B.5 for this and additional participation data.

TABLE IV

NUMBER OF PARTICIPANTS AND MESSAGES IN RACES

Description	Mean	Std. dev.	Min	Pct01	Pct10	Pct25	Median	Pct75	Pct90	Pct99	Max
Panel A: Number of participants											
Participants within 50 μs	1.77	0.86	1	1	1	1	2	2	3	5	12
Participants within 100 μs	2.08	0.97	1	1	1	1	2	2	3	5	13
Participants within 200 μs	2.56	1.13	1	1	2	2	2	3	4	6	16
Participants within 500 μs	3.27	1.56	2	2	2	2	3	4	5	9	23
Participants within 1,000 μs	3.64	1.94	2	2	2	2	3	4	6	11	26
Panel B: Number of take messages											
Takes within 50 μs	1.66	0.97	0	0	1	1	1	2	3	5	14
Takes within 100 μs	1.93	1.08	0	0	1	1	2	2	3	5	15
Takes within 200 μs	2.37	1.30	0	1	1	1	2	3	4	7	17
Takes within 500 μs	3.07	1.78	1	1	1	2	3	4	5	9	29
Takes within 1,000 μs	3.45	2.19	1	1	1	2	3	4	6	11	40
Panel C: Number of cancel messages											
Cancels within 50 μs	0.17	0.41	0	0	0	0	0	0	1	1	8
Cancels within 100 μs	0.22	0.47	0	0	0	0	0	0	1	2	8
Cancels within 200 μs	0.30	0.56	0	0	0	0	0	1	1	2	12
Cancels within 500 μs	0.40	0.70	0	0	0	0	0	1	1	3	14
Cancels within 1,000 μs	0.44	0.78	0	0	0	0	0	1	1	3	21

Notes. For each race detected by our baseline method, we obtain the timestamp of the first inbound message and the price and side of the race. We then use the message data to obtain all 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 or cancels at exactly the race price and side. 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.

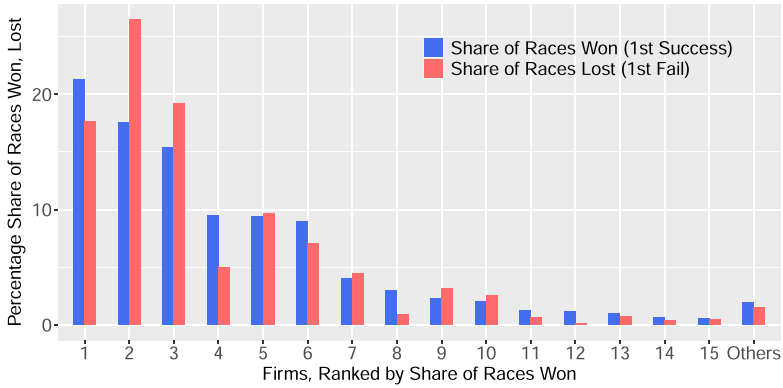

FIGURE III

Percentage of First Successful and First Failed Messages by Firm

For each race detected by our baseline method, 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 Figure II). 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 figure sorts FirmIDs based on the proportion of races won. The "Others" bar sums all FirmIDs outside of the top 15.

3. *Pattern of Winners and Losers.* Figure III displays data on the pattern of winners and losers across races. 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 three 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 three firms each win about another 9% of races each, and then there are another four firms that win 2%–4% of races each.

It is notable that there is clear concentration of winners, with the top three firms winning 54% of races, and the top six firms winning 82% of races. Yet these same firms who win a lot of races also lose a lot of races. The top three winning firms lose 63% of races, and the top six 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 small number of firms competitive in this race.[24]

4. *Pattern of Takes, Cancels, and Liquidity Provision.* Figure IV, 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 IV, which showed that most of the message activity in races is take attempts as opposed to cancel attempts.

Figure IV, Panel B provides data on the pattern of successful takes, successful cancels, and liquidity provision across firms. The top six firms, as defined by the proportion of races won (Figure III), account for about 80% each of race wins, liquidity taken in races, and liquidity successfully canceled in races. In contrast, these six 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 six firms.

Within these top six firms there are two distinct patterns of race participation. Two of the top six firms 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 two firms engage in meaningful quantities of stale-quote sniping and liquidity provision; their ratio of liquidity taken in races to liquidity provided in races is about 2:3. The remaining four of the top six 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 four 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 six" and "takers in top six," respectively.[25]

Market participants outside of the top six firms account for about 20% each of race wins, liquidity taken in races, and

24. Around this time, the CEO of a prominent high-frequency trading firm described to one of the authors of this study that, in the United States, there were seven firms in what he called the "lead lap" of the speed race.

25. 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 have an aggressiveness ratio of 82%, which means that 82% of their overall trading volume is aggressive, with the remaining

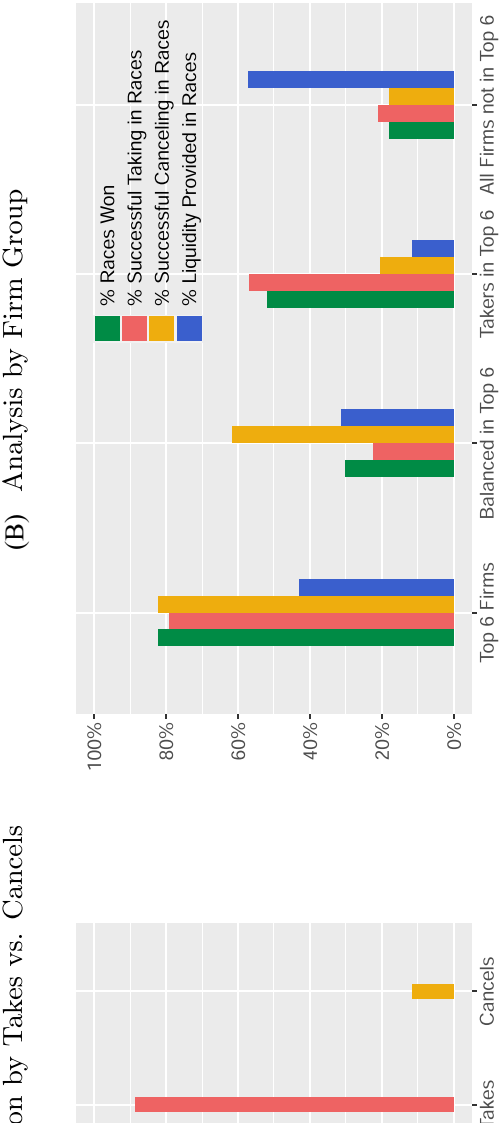

FIGURE IV

Pattern of Takes, Cancels, and Liquidity Provision

Panel A: For each FTSE 100 race detected by our baseline method, 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 III 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 and using the FirmID associated with the aggressive order in each trade. The third bar, % Successful Canceling in Races, is computed by taking all successful cancels in FTSE 100 races and using the FirmID associated with the cancel attempt. The fourth bar, % Liquidity Provided in Races, is computed by taking all trading volume in all FTSE 100 races and using the FirmID associated with the passive side of each trade, that is, the resting order that was taken by the aggressive order used in the % Successful Taking bar.

TABLE V

LIQUIDITY TAKER-PROVIDER MATRIX: % OF RACE VOLUME BY TAKER-PROVIDER COMBINATION

		Provider		
		Takers in top six	Balanced in top six	Non-top six
Taker	Takers in top six	5.7	17.2	34.3
	Balanced in top six	2.5	6.4	13.3
	Non-top six	3.2	7.4	10.1

Notes. For each race detected by our baseline method, 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.

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 three times as much liquidity in races as they take.

Thus, on net, much race activity consists of firms in the top six taking liquidity from market participants outside of the top six. 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 six firm taking from a market participant outside the top six (34.3% of all race volume). This pattern seems more consistent with the "rigged" as opposed to "myth" point of view, as discussed in the introduction.

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 six firm sniping a balanced in top six firm (17.2%). See Table V for a matrix of race trading volume organized by such taker-provider combinations.

5. *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 Poisson randomly.[26] We say that a message

18% passive. Baron et al. (2019) report that the 90th percentile of firms in their sample have an aggressiveness ratio of 88%.

26. An influential paper by Engle and Russell (1998) provides more sophisticated econometric techniques to deal with the fact that messages arrive at time-varying rates. In the introduction (p. 1128) they write: "Even more intriguing is the case of transactions that are generally infrequent but that may suddenly exhibit very high activity. This may be due to some observable event such as a news

is potentially race relevant if the message is either a marketable limit order (including marketable IOCs) or 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 messages 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 $T = 200$ microseconds, about the mean information horizon in our data set, is 1.42. The number of times we would expect to see $N = 2$ such messages within $T = 500$ microseconds, the upper bound on the information horizon, is 3.55. For the mean FTSE 250 symbol-date, the figures are 0.02 and 0.04. The number of times we would expect to see $N = 3$ or more such messages arrive by chance in such a time window, for either FTSE 100 or FTSE 250, is 0.00. See Table VI.

Accounting for the fact that the rate of message arrivals is higher near the open and close of the UK trading day, and during the window that coincides with the U.S. open, increases these numbers only modestly. Even if we assume that the entire trading day is as busy as the symbol-date's busiest half-hour segment, the average number of times we would observe two messages that could possibly be racing, within 500 microseconds, is just 13.26 for FTSE 100 symbols and 0.21 for FTSE 250 symbols.

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).

---

release or to an unobservable event which may best be thought of as a stochastic process." Our article relates to Engle and Russell (1998) in that it relates to *why* messages sometimes arrive in bursts, and in particular in clusters in amounts of time that would have been difficult to fathom at the time of Engle and Russell (1998). For this reason, we use Poisson as a particularly simple benchmark to give a sense of how many races one might expect to observe by chance, and then use as a sensitivity an increased Poisson arrival rate, reflecting the kinds of higher arrival rates Engle and Russell (1998) had in mind at, for instance, the market's open and close.

TABLE VI

EXPECTED NUMBER OF POTENTIAL RACE EVENTS BY CHANCE

	FTSE 100		FTSE 250	
	Average rate	Busiest 30 mins	Average rate	Busiest 30 mins
Expected occurrences by chance				
2+ within 50 μs	0.35	1.33	0.00	0.02
2+ within 100 μs	0.71	2.65	0.01	0.04
2+ within 200 μs	1.42	5.31	0.02	0.09
2+ within 500 μs	3.55	13.26	0.04	0.21
2+ within 1,000 μs	7.09	26.49	0.08	0.43
3+ within 1,000 μs	0.00	0.03	0.00	0.00
Actual number of races				
Baseline analysis	537.24		70.05	
3+ within info horizon	228.98		30.68	

Notes. For each symbol-date, we calculate the arrival rate of potentially race-relevant messages (see text for description) and use this to compute the expected number of occurrences of N such messages within T microseconds, on the same side of the order book, if messages arrive at this rate via a Poisson arrival process. For each symbol-date, we also perform this calculation using the arrival rate of potentially race-relevant messages during the busiest 30 minutes of the day, assuming the entire day has this level of activity. We also report the actual number of races, both for the baseline and for the sensitivity in which we condition on there being at least 3+ participants in the information horizon.

The bottom line is that the number of races we would observe by chance is de minimis. For additional details, see Online Appendix B.2.

IV.C. *Race Profits*

1. *Profits per Race.* Table VII 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.[27] Our main results use the midpoint

27. 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 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. In any event, LSE's trading fees are small relative to average race profits: 0.15 basis points for aggressive orders from high-volume participants, and 0 for orders that are passively executed.

TABLE VII

DETAIL ON RACE PROFITS (PER SHARE AND PER RACE) MARKED TO MARKET AT 10 S

Description	Mean	Std. dev.	Pct01	Pct10	Pct25	Median	Pct75	Pct90	Pct99
Panel A: FTSE 100									
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									
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

Notes. For each race detected by our baseline method, 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.

10 seconds out, and we also report figures for horizons ranging from 1 millisecond to 100 seconds shortly.[28]

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 per race, measured either using all of the displayed depth at the start of the race (£1.95) or all of the quantity traded or canceled during the race (£1.84). For the FTSE 250, the figures are 0.77 ticks, 3.09 basis points, and £1.55 per race based on displayed depth, and £1.48 per race based on quantity traded or canceled. For the full sample, the figures are 0.55 ticks, 1.66 basis points, £1.85, and £1.76.

Of course, there is significant variation in profitability across races. This reflects both that some races are more profitable ex ante than others, that is, 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 VIII presents statistics on average profits per race for different mark-to-market time horizons. As can be seen, average profits per race increase with the time horizon, eventually flattening out at around 10 seconds for the FTSE 100 and around 60 seconds for the FTSE 250. Our finding that it takes nonzero 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 (2020) on what they call the "term structure of liquidity."

Figure V complements Table VIII by presenting the distribution of race profits and price impact at different time horizons. The difference between the measures is that race profits are the difference between the price paid in the race and the midpoint

28. Because 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 data set obtained under a Section 165 request, there are two key limitations to this idea. First, we only have data from the LSE, so we 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 Korajczyk and Murphy 2019).

TABLE VIII

AVERAGE RACE PROFITS (PER SHARE AND PER RACE) FOR DIFFERENT MARK-TO-MARKET HORIZONS

Description	1 ms	10 ms	100 ms	1 s	10 s	30 s	60 s	100 s
Panel A: FTSE 100								
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								
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

Notes. For each race detected by our baseline method and for each race profits measure described in Table VII, we recompute the profits measure for different mark to market horizons, ranging from 1 millisecond to 100 seconds (Table VII used $T = 10$ seconds). We then report the mean at each horizon.

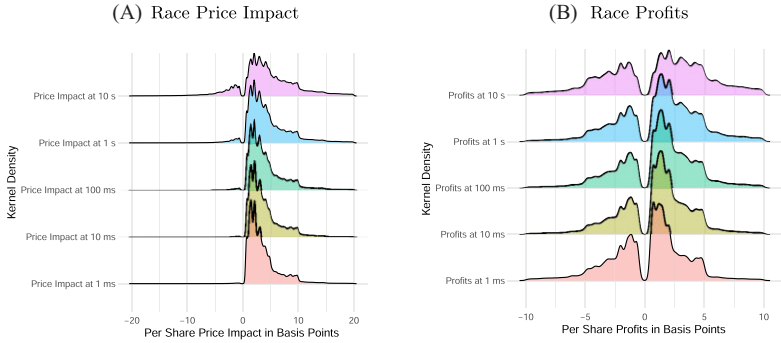

FIGURE V

Race Price Impact and Profits Distributions at Different Time Horizons

For each race detected by our baseline method, we obtain per share profits and price impact in basis points at different mark to market horizons ranging from 1 millisecond to 10 seconds. Profits at horizon T are defined as the signed difference between the race price and the midpoint price at time T, whereas 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 price impact (Panel A) and per share profits (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.

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 1 ms. 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 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 VIII 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.

2. *Remark: Races with Negative Realized Profits.* 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 IV.B and Table VI, we suspect the latter interpretation is more quantitatively important. To give a sense of magnitudes, at each of the 10 ms, 100 ms, and 1 s time horizons, about 80% of races are weakly profitable and about 20% of races have strictly negative realized profits. Conditional on race profits being nonzero, about 70% have positive profits and 30% have negative profits. See further discussion in Section V.B.

IV.D. Aggregate Profits and the "Latency Arbitrage Tax"

Table IX 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 these numbers are daily race profits in our data from the LSE; we extrapolate from these numbers to the full UK equities market and to global equities markets in Section VI.

Referring to Panel A, we see that the average symbol in the FTSE 100 has daily race profits of £1,047, and the 99th percentile symbol has daily race profits of £3,432. For the FTSE 250 the average and 99th percentile are £108 and £606, respectively. Referring to Panel B, we see that the average day in our data set has race profits of £105,734 for the FTSE 100, £26,643 for the FTSE 250, and £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 nonrace trading volume. The version based on all

TABLE IX

DAILY PROFITS IN GBP

Panel A: Daily profits by symbol

Description	Mean	Std. dev.	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	Std. dev.	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, we take per race profits in GBP based on displayed depth with prices marked to market at 10 seconds (see notes for Table VII). 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.

trading volume is both simpler to describe and more appropriate for out-of-sample extrapolation. However, the version based on all nonrace trading volume more closely corresponds to the theory, which shows that latency arbitrage imposes a tax on nonrace trading (both noise trading and nonrace informed trading).

Table X 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 nonrace 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 nonrace 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 nonrace volume measure.

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

1. *Relationship between Profits, Volume, and Volatility.* Figure VI presents scatterplots of latency arbitrage profits against trading volume (Panel A) and one-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 one-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.

Online Appendix Figure B.2 presents scatterplots of the latency arbitrage tax against these same measures: trading volume (Panel A) and one-minute realized volatility (Panel B). The 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 report further details on these relationships in Section VI, where they will be used for the purpose of out-of-sample extrapolation.

TABLE X

LATENCY ARBITRAGE TAX

Panel A: Distribution across symbols

Description	Mean	Std. dev.	Pct01	Pct10	Pct25	Median	Pct75	Pct90	Pct99
Measure 1, latency arbitrage tax based on all trading volume (basis points)									
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
Measure 2, latency arbitrage tax based on nonrace trading volume (basis points)									
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

Description	Mean	Std. dev.	Min	Pct10	Pct25	Median	Pct75	Pct90	Max
Measure 1, latency arbitrage tax based on all trading volume (basis points)									
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
Measure 2, latency arbitrage tax based on nonrace trading volume (basis points)									
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 VII). We then compute total regular-hours trading volume in GBP, and total nonrace regular-hours trading volume in GBP (see notes for Table III). Panel A, Measure 1 reports the distribution across symbols of race profits divided by all trading volume. Panel A, Measure 2 reports the distribution across symbols of race profits divided by nonrace trading volume. Panel B is the same except at the date level (with race profits, all volume, and nonrace volume each summed across all symbols) instead of the symbol level.

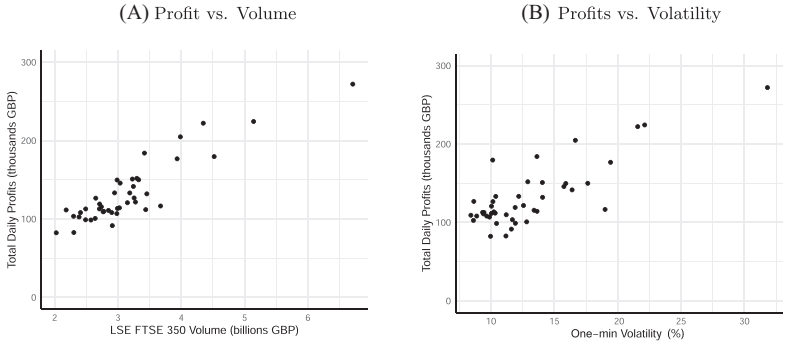

FIGURE VI

Latency Arbitrage Profits Correlation with Volume and Volatility

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

IV.E. Latency Arbitrage's Share of the Market's Cost of Liquidity

We quantify latency arbitrage as a proportion of the market's overall cost of liquidity. We present two complementary approaches.

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:

$$EffectiveSpread = PriceImpact + RealizedSpread, \tag{1}$$

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, 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 suggests two refinements to equation (1). First, we can decompose the price impact component of the spread into two components: one that reflects latency arbi-

trage and one that reflects traditional private information. 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; that is, 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

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

with the notation defined as follows. λ_{invest}, λ_{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 Online Appendix F.2 we show formally that equation (2) implies the spread decomposition:

$$\begin{aligned} EffectiveSpread &= PriceImpact_{Race} + PriceImpact_{NonRace} \\ &\quad + LossAvoidance + RealizedSpread, \end{aligned} \tag{3}$$

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 use 10 seconds), for trades in latency arbitrage races and trades not in latency arbitrage races. 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.

Table XI gives details for decomposition (3) at the symbol level. For the average symbol in the FTSE 100, averaged over

TABLE XI

SPREAD DECOMPOSITION: DISTRIBUTION ACROSS SYMBOLS (FTSE 100)

Description	Mean	Std. dev.	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

Note. Please see the text of Section IV.E for definitions of effective spread, price impact (PI), loss avoidance, and realized spread.

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).[29] 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.[30] This result suggests that liquidity provision is modestly profitable in nonrace 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.

For symbols in the FTSE 250 (see Online Appendix Table B.12), overall effective spreads are higher, at 8.06 basis

29. Realized spreads are slightly positive if price impact is measured at a shorter duration, such as 100 ms or 1 s rather than 10 s (see Online Appendix Tables B.13 and B.14). This is consistent with Conrad and Wahal (2020), who find that realized spreads decrease as the time interval decreases. Please note as well that at the LSE liquidity providers do not receive rebates, whereas in markets such as the United States, where rebates are common, this could lead to a negative realized spread being a rational feature of equilibrium liquidity provision (Battalio, Corwin, and Jennings 2016).

30. Note that the realized spread in races, multiplied by the roughly 22% of trading volume in races as reported in Table III, corresponds roughly to the all-volume latency arbitrage tax as reported in Table X. (The relationship is not exact due to loss avoidance, which we count as part of the latency arbitrage prize but does not count toward 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.

points, realized 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. As with FTSE 100 stocks, price impact is significantly higher in races than in nonrace trading (12.22 basis points versus 7.50 basis points), and consequently the realized spread is modestly positive in nonrace trading (0.72 basis points) and meaningfully negative in races (−5.48 basis points). Aggregated over all trading volume, price impact in races accounts 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.

Overall, these results suggest that latency arbitrage deserves a place alongside traditional adverse selection as one of the primary components of the market's cost of liquidity.

2. *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 six firms as defined by the proportion of races won (see Figure III) the realized spread in races is −1.699 basis points, versus −1.930 basis points for firms outside the top six. The difference between the takers and balanced firms in the top six is small as well: −1.493 basis points versus −1.775 basis points. See Table XII.

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

There is a more significant difference between faster and slower firms in their canceling behavior. The top six 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 six firms, the maximum cancel rate is 66% within the race horizon and 68% of the time within 1 millisecond. Firms outside of the top six 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

TABLE XII

REALIZED SPREADS IN RACES BY FIRM GROUP

	Realized spread (bps)			Cancel attempt rate (%)		
Firm group	Overall	Nonrace	Race	In race	Within 1 ms	Ever
All Firms	– 0.209	0.236	– 1.833	19.29	21.89	24.53
Fast versus slow						
Top six	– 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 six	0.016	0.455	– 1.493	45.16	47.56	47.82
Balanced in top six	– 0.120	0.311	– 1.775	30.97	35.09	36.33

Notes. Firm groups are as in Figure IV. The realized spread is calculated as described in the text and reported in Table XI. 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).

look beyond 1 millisecond to include any failed cancel attempts of quotes taken in a race, the top six cancel attempt rate goes up to 40% and the cancel rate for firms outside of the top six goes up to 13.35%.[31] 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 firms are better than slower firms at avoiding this cost, but even they get sniped with significant probability if their quotes become stale.[32]

31. For firms in the top six, essentially all of the incremental failed cancels come within 3 milliseconds after the race start (98.57% of all cancel attempts are within 3 milliseconds of the race start). For firms outside the top six 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%).

32. Our best guess for why slower firms rarely attempt to cancel, and even fast firms sometimes do not attempt to cancel, is that by the time the quote provider has figured out that their quote is stale and they should reprice, they already have received the update that their quote has been executed. That is, if the information horizon is 200 microseconds, they "lose the race" by more than 200 microseconds. We note as well that conditional on losing a race, the average cost is just half a tick, so this could easily be accounted for as a "cost of doing business". To confirm

3. *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:

$$\frac{\frac{s^{CLOB}}{2} - \frac{s^{FBA}}{2}}{\frac{s^{CLOB}}{2}}, \tag{4}$$

where s^{CLOB} is the bid-ask spread under the CLOB and s^{FBA} is the bid-ask spread under a counterfactual market design, frequent batch auctions (FBAs), which eliminates latency arbitrage. To turn expression (4) into something empirically measurable, we take the following steps. First, we multiply the numerator and denominator of expression (4) by $(\lambda_{invest} + \lambda_{private})$. Second, we use equation (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),

$$\lambda_{invest}\frac{s^{FBA}}{2} = \lambda_{private} \cdot L\left(\frac{s^{FBA}}{2}\right), \tag{5}$$

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 expression (4). Observe that the difference between the equilibrium bid-ask spread characterization for frequent batch auctions, equation (5), and the equilibrium bid-ask spread for continuous trading, equation (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 Online Appendix F.3 for completeness, shows that expression (4) can be reexpressed as:

$$\frac{\frac{s^{CLOB}}{2} - \frac{s^{FBA}}{2}}{\frac{s^{CLOB}}{2}} = \frac{\lambda_{public}L\left(\frac{s^{CLOB}}{2}\right)}{(\lambda_{invest} + \lambda_{private})\frac{s^{CLOB}}{2}}. \tag{6}$$

this hypothesis would require data from a broker-dealer execution algorithm, for instance.

Both the numerator and denominator of the right-hand side of equation (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 nonrace 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 the numerator and denominator by nonrace trading volume, in basis points terms. Thus, we have the relationship:

$$\begin{aligned}\text{Proportional Reduction in Liquidity Cost} &= \frac{\text{Race Profits (GBP)}}{\text{Nonrace Effective Spread (GBP)}} \\ &= \frac{\text{Latency Arbitrage Tax (Nonrace Volume)}}{\text{Nonrace Effective Spread (bps)}}.\end{aligned} \tag{7}$$

Table XIII presents our computation of equation (7). For the average symbol in the FTSE 100, eliminating latency arbitrage would reduce the cost of liquidity by 19.95%.[33] 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 X), bid-ask spreads are several times wider for FTSE 250 symbols than for FTSE 100 symbols (see Online Appendix Tables B.10–B.12), so eliminating latency arbitrage would reduce the overall cost of liquidity by less for the FTSE 250 than for the FTSE 100.

33. It may at first be confusing why eliminating latency arbitrage reduces the spread by about 20% for FTSE 100 stocks in this exercise, whereas price impact in latency arbitrage races constituted 37% of the effective spread in Table XI. The difference is that latency arbitrage profits charge the aggressor the half spread, whereas the price impact calculation in effect does not. Here is the rough back-of-the-envelope math. The effective spread is on average about 3 basis points. Price impact in races is about 5 basis points and races constitute 23% of volume for the average FTSE 100 symbol (Online Appendix Table B.4). Therefore price impact in races as a proportion of the effective spread is $\frac{23\% \cdot 5\text{bps}}{3\text{bps}}$, which is about 37% as claimed. The latency arbitrage tax on nonrace volume is $\frac{(5\text{bps}-3\text{bps}) \cdot 23\%}{(100\%-23\%)}$, which is about 0.60 basis points, or about 20% of the nonrace effective spread, implying a roughly 20% reduction in the cost of liquidity as claimed.

TABLE XIII

PERCENTAGE REDUCTION IN LIQUIDITY COST, IF LATENCY ARBITRAGE WERE ELIMINATED

Panel A: Symbol level									
Description	Mean	Std. dev.	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	Std. dev.	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 (7) by summing total race profits in GBP, across all dates, and then dividing by total non race effective spread paid in GBP, summed across all dates. Race profits in GBP are as described in Table IX and effective spread paid in GBP is as described in Table XI. Analogously, for each date, we implement equation (7) by summing total race profits in GBP, across all symbols, and then dividing by total nonrace effective spread paid in GBP, summed across all symbols. We do both exercises separately for FTSE 100, FTSE 250, and the full sample. In Panel A, 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.

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%.[34]

V. Sensitivity Analysis and Robustness Checks

We performed a wide range of sensitivity analyses and robustness checks. First, we explored how our main results as presented in Section IV vary as we modify each component of the race definition presented in Section III. The insights from this work are discussed in Section V.A, with a summary table in the main text and full details in Online Appendix C. Second, we performed additional robustness analyses to better understand races in our data that do not fit as neatly in the paradigm of the BCS model: races with negative mark-to-market profits, and races where the best bid and offer (BBO) is volatile just before the race. This analysis is discussed in Section V.B with supporting details in Online Appendix D.

V.A. Sensitivity to Varying the Definition of a Race

As described in Section III, for each component of the race definition—multiple participants, at least some of whom are aggressive, at least some of whom succeed and some of whom fail, all at the "same time"—we performed our full analysis for a baseline and alternatives. In this subsection we report the main insights from these alternative specifications. Table XIV presents a range of sensitivity scenarios informed by this work.

1. *Finding 1: Effect of Race Horizon.* Our baseline method requires that a set of messages satisfying the baseline race requirements arrives within the information horizon of the first message of the race, which averages about 200 microseconds and is capped at 500 microseconds. In sensitivity analyses, we ex-

34. Both equation (2) for the spread in the continuous market and equation (5) for the spread in the FBA market assume no tick size constraints. A market design reform that both adopted FBA and eased tick size constraints, as advocated by Yao and Ye (2018), Kyle and Lee (2017), and others, would, based on these estimates, reduce the cost of liquidity by more than 16.7%. If tick size constraints bind, then the liquidity advantage of FBA relative to continuous trading can manifest in greater depth as opposed to a narrower spread. There can be more liquidity at a given price while keeping the marginal unit of liquidity provision indifferent between providing and not, because there is no sniping.

plored instead requiring that the set of messages satisfying the baseline race requirements arrives within a time window of T, with values of T ranging from 50 microseconds to 3 milliseconds. The longer horizons are intended to capture races among firms of varying technological sophistication that could still be considered racing one another.[35] We consulted with HFT industry contacts and FCA supervisors to agree on an appropriate horizon. Following these discussions, we determined that 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. We will also miss races where, by the time the loser or losers of the race detect the signal, they can already tell from their representation of the order book that they are too late and hence do not bother to send a message; such cases can be understood as a gray area between asymmetric private information (because the winner understood the signal far enough in advance of the losers) and symmetric public information (because the time differences are still quite small).

The main pattern that emerges from this analysis is that the longer is T, the more races we find, without much effect on the various measures of per race profits. The increase is especially steep up through 500 microseconds. For example, with $T = 100\,\mu$s the number of FTSE 100 races per symbol per day is 389, with $T = 500\,\mu$s the number is 720, with $T = 1$ ms the number is 768, and with $T = 3$ ms the number is 800.[36]

As a result, our measures of the effect of latency arbitrage on the cost of liquidity are all strongly increasing with the race horizon. At $T = 100\,\mu$s the full-sample latency arbitrage tax is 0.26 bps, price impact is 19.2% of the effective spread, and the

35. Sources of speed loss for firms that are sophisticated but not at the cutting edge of speed include not using code and hardware that is optimized for speed, not using the fastest colocation and connectivity options at exchanges, not using the fastest links to overseas markets such as the United States, and not using microwave connections where possible to do so.

36. The numbers reported in Table XIV at horizons of 500 microseconds and longer also reflect the sensitivity requirement that plain-vanilla (non-IOC) limit orders cannot count as fails, that is, only failed IOCs and failed cancels count as fails. See discussion under Finding 4. The Online Appendix includes a version of the time horizon sensitivity without this requirement (Table C.1); the numbers are about 10%–15% bigger.

implied reduction in the market's cost of liquidity from eliminating latency arbitrage is 9.5%. At $T = 500\,\mu$s the figures are 0.60 bps, 50.8%, and 28.1%, respectively; at $T = 1$ ms the figures are 0.68 bps, 59.3%, and 35.4%; and at $T = 3$ ms the figures are 0.74 bps, 66.1%, and 41.6%. That is, if the race window is defined as 3 milliseconds instead of the information horizon, latency arbitrage constitutes about 65% of the effective spread and eliminating latency arbitrage would reduce the market's cost of liquidity by about 40%.

2. *Finding 2: Number of Race Participants.* Our baseline method requires that there are at least two race participants in the information horizon. In sensitivity analysis we consider requiring 3+ participants and 5+ participants. Given the large effect that the race's time horizon had on the number of races and the harm to market liquidity, we perform this sensitivity for the baseline information horizon method and for fixed race horizons ranging from 50 microseconds to 3 milliseconds.

The main pattern that emerges from this sensitivity is that increasing the required number of race participants lowers the number of races found while increasing the various measures of profitability per race. For example, requiring 3+ participants reduces the number of races in the information horizon by about 60%, but increases per race profits by about 60%. The net effect is that measures of total race profits and harm to liquidity fall by about one-third: the latency arbitrage tax is 0.29 bps, race price impact's proportion of the spread is 20.5%, and the implied reduction in the market's cost of liquidity is 10.4%.

Increasing the race horizon increases the number of races detected, just as in the baseline case. The overall magnitudes for the total cost of latency arbitrage are similar among the baseline method, the sensitivity with 3+ participants within 500 microseconds, and the sensitivity with 5+ participants within 1 millisecond.

3. *Finding 3: Takes and Cancels.* Our baseline method defines a race to consist of either 1+ aggressors and 1+ cancels, or 2+ aggressors and 0 cancels. The former case corresponds to the equilibria studied in BCS, whereas the latter case, in which all race activity is aggressive, corresponds to equilibria in a modest extension of BCS presented in Online Appendix F.

There are two main findings that emerge from our sensitivity analysis of these criteria. First, requiring a cancel attempt within the race horizon significantly reduces the number of races and the associated harm to market liquidity. If we require at least 1 cancel within the information horizon, the number of races and the various harm to liquidity measures are each about 30% of the baseline. This is as expected given our findings in Section IV.B that most of the message activity in races is aggressive. That said, if we consider races with 1+ cancel over a 3 millisecond time horizon, then the results are closer to baseline, at about 85% of the number of races and the various harm to liquidity measures.

Second, races with just a single take attempt (i.e., 1 aggressor and 1+ cancels) have meaningfully lower profitability than races with 2+ aggressors. As a consequence, imposing the requirement that there are 2+ aggressors in a race lowers the number of races by about 20% but lowers the latency arbitrage tax by closer to 10%.

4. *Finding 4: Success and Fail Criteria.* Reassuringly, in our baseline analysis, varying the definition of success and fail does not move the needle too much. At longer time horizons, whether or not we treat plain-vanilla (non-IOC) limit orders as potential fails makes a bigger difference, affecting the number of races detected by on the order of 10%–15%. This makes sense because at longer horizons we should be more concerned about mistaking limit orders that post to the book with the intent to provide liquidity as failed race attempts. For this reason, in the summary of sensitivity scenarios presented as Table XIV, we do not allow non-IOC limit orders to count as fails at time horizons of 500 microseconds and longer, that is, at time horizons longer than the information horizon only failed IOCs and failed cancels count as fails.

5. *Selected Sensitivity Scenarios.* Based on what we have learned from the various sensitivity analyses, Table XIV 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 and to stricter requirements on the level of participation, we highlight: 2+ within 50 microseconds, 2+ within 100 microseconds, and 3+ within the information horizon.

TABLE XIV

SENSITIVITY ANALYSIS: SELECTED SCENARIOS

		Low scenarios			Middle scenarios			High scenarios		
Measure	Baseline	2+, 50 μs	2+, 100 μs	3+, IH	2+, 200 μs	2+, 500 μs	3+, 500 μs	2+, 1 ms	2+, 3 ms	3+, 3 ms
Panel A: Frequency and duration of races										
Races per day										
FTSE 100, per symbol	537.24	296.66	388.58	228.98	521.53	719.71	458.94	768.02	799.91	609.01
% of volume in races										
Full sample	21.46	9.77	13.32	12.30	19.21	34.89	26.42	40.03	43.72	38.14
Panel B: Per race profits										
Per share profits										
Ticks	0.55	0.54	0.53	0.71	0.51	0.52	0.60	0.54	0.55	0.64
GBX	0.17	0.16	0.16	0.23	0.16	0.16	0.19	0.16	0.16	0.19
Basis points	1.66	1.68	1.63	2.24	1.57	1.59	1.90	1.64	1.64	1.92
Per race profits GBP										
Displayed depth	1.85	1.58	1.59	2.98	1.60	1.91	2.54	2.04	2.09	2.67
Qty trade/cancel	1.76	1.38	1.44	2.87	1.51	1.92	2.59	2.07	2.16	2.76
Panel C: Latency arbitrage (LA) tax										
LA tax, all volume (bps)	0.42	0.20	0.26	0.29	0.35	0.60	0.50	0.68	0.74	0.70
LA tax, nonrace volume (bps)	0.53	0.22	0.30	0.33	0.44	0.92	0.70	1.14	1.31	1.14
Panel D: Spread decomposition										
Price impact in races / All price impact %	30.58	12.84	17.89	19.13	25.69	47.38	37.08	55.27	61.61	55.54
Price impact in races / Effective spread %	32.82	13.77	19.19	20.54	27.57	50.84	39.61	59.31	66.11	59.61
Panel E: Implied reduction in cost of liquidity										
Full sample, % reduction	16.73	6.96	9.49	10.43	13.62	28.12	20.96	35.37	41.64	36.20

Notes. For descriptions of the sensitivity scenarios, see the text. Descriptions of each item in this table can be found in the following table notes in Section IV. Races per day: Table II. % of volume in races: Table III. Per race profits: Table VII. Aggregate profits: Table IX. Latency arbitrage tax: Table X. Spread decomposition: Table XI. Implied reduction in cost of liquidity: Table XIII. All scenarios with a time horizon of 500 microseconds or longer use the strict fails criterion in which non-IOC limit orders cannot count as fails, as discussed in the text.

As medium scenarios, we highlight: 2+ within 200 microseconds, 2+ within 500 microseconds, and 3+ within 500 microseconds.

As high scenarios, we highlight: 2+ within 1 millisecond, 2+ within 3 milliseconds, and 3+ within 3 milliseconds.

Over this set of scenarios, the latency arbitrage tax ranges from 0.20 to 0.74 basis points on the all-volume measure, and from 0.22 to 1.31 basis points on the nonrace volume measure. Latency arbitrage as a percentage of trading volume ranges from 9.8% to 43.7%. Latency arbitrage as a percentage of the effective spread ranges from 13.8% to 66.1%. The potential reduction in the market's cost of liquidity ranges from 7.0% 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 0 (e.g., 5+ within 50 microseconds yields very low numbers, see Online Appendix Table C.3). Similarly, at the high end, one could be even more inclusive (e.g., looking at horizons longer than 3 milliseconds). Or one could attempt to find a way to account for the gray-area case, between symmetric public information and asymmetric private information, where one firm's response to a trading signal is sufficiently faster than others' that by the time other firms observe the signal, they can already tell from the order book that they are too late and hence don't bother. 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 VI where we provide extrapolations for the total sums at stake in latency arbitrage races.

V.B. Additional Robustness Checks

We did additional robustness work to explore two aspects of the main results that do not fit neatly within the original BCS model.

1. *Races with Negative Profits Ex Post.* In our main specification, 20% of races have strictly negative profits when marked to market at 100 milliseconds after the race, 22% when marked at 1 second after the race, and 29% when marked at 10 seconds after the race. Although these numbers surely reflect some postrace noise, we find that 8% of races have strictly negative profits con-

tinuously for the 10 seconds after the race. (As seen in Figure V, when a race is not profitable, typically the pattern is that price impact is weakly in the direction of the race but not by enough to recover the half-spread.)

Even in our most strenuous sensitivity test, which requires 5+ participants within 50 microseconds, 10% of races have strictly negative profits 100 milliseconds after the race, 11% at 1 second, 19% at 10 seconds, and 3% of races have strictly negative profits continuously for the 10 seconds after the race. For full details, see Online Appendix D.1.

These results suggest to us that at least some races are based on noisy signals that turn out not to be profitable ex post. Although this squares with common intuitions about algorithmic trading more broadly, where it can of course be rational to trade on signals that are profitable in expectation but noisy—and our figures suggest that trading in races is a lot closer to pure arbitrage than the 51/49 odds described by Renaissance and Virtu[37]—it is inconsistent with a literal interpretation of the BCS model in which the public signal that triggers races is perfectly correlated to the fundamental value of the asset.

2. *Races Triggered by Order Book Activity.* In the BCS model, races are triggered by jumps in a public signal, interpreted for example as a change in the price of a highly correlated asset or the same asset on another venue. A recent paper of Li, Wang, and Ye (2021) extends the BCS model to incorporate discrete price increments (i.e., tick size constraints) and a stylized version of institutional investor execution algorithms and finds that races can be triggered by both public signals and order book activity by the execution algorithms. Specifically, if an execution algorithm places a limit order in the book that is sufficiently attractive (e.g., a new bid that is sufficiently close to the ask), and trading firms are sufficiently confident that this order does not reflect new information, the order could trigger a race.[38]

We find some evidence for this pattern in our data. In about 14% of races, there is a change in the race-side best bid or offer in

37. See note 8 in Section I on Renaissance and Virtu, and also see MacKenzie (2021) who discusses several different types of canonical HFT signals, some of which are closer to pure arbitrage and some of which are more statistical in nature.

38. See also Foucault, Kozhan, and Tham (2016), who call this "nontoxic arbitrage."

the 100-microsecond window just prior to the race, and of these, nearly all of the price changes (89% of the subset, or 12% of the total) are in the direction consistent with Li, Wang, and Ye (2021). These races have fewer cancellations than baseline races (0.24 versus 0.40) and a larger share of liquidity provided by non-top six firms (71% versus 58%), both of which also seem consistent with the theory in Li, Wang, and Ye (2021).

That said, the large majority of races have stable prices leading up to the race. This suggests that most races are triggered by some public signal external to the symbol's own order book, as in the BCS model. For full details, see Online Appendix D.2.

VI. Total Sums at Stake

VI.A. Extrapolation Models

Figure VI showed visually that daily latency arbitrage profits are highly correlated with market volume and volatility, as expected given the theory. Table XV presents these same relationships in regression form for the purpose of out-of-sample extrapolation.

Columns (1) and (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 column (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) and (4) regress daily in-sample latency arbitrage profits on daily realized one-minute volatility.[39] 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.[40] Here, including the constant term does provide a meaningfully better fit, which can also be seen visually in the scatterplot in Figure VI, Panel B. The coefficient of

39. In the Online Appendix, we report regression results for five-minute volatility and for a measure of volatility emphasized in BCS called distance traveled. Five-minute volatility has lower explanatory power than one-minute volatility. Distance traveled actually has greater explanatory power than one-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.

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

TABLE XV

EXTRAPOLATION MODELS

	Dependent variable:					
	Latency arbitrage profits (GBP)					
	(1)	(2)	(3)	(4)	(5)	(6)
Volume (GBP 10,000s)	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

Notes. The dependent variable in all regressions is daily race profits in GBP, for the full sample, as reported in Table IX. Volume is daily regular-hours LSE trading volume, in units of GBP 10,000s so that the coefficient is interpretable as a latency arbitrage tax in basis points. Volatility is realized one-minute volatility for the FTSE 350 index in percentage points, using TRTH data, as described in Figure VI. Volatility in percentage points is multiplied by average daily volume in GBP 10,000s 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 $1 - \frac{\mathrm{Var}(\hat{e})}{\mathrm{Var}(y)}$. p-values are computed using the student-t distribution. * $p < .1$; ** $p < .05$; *** $p < .01$.

0.023 in column (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) and (6) present results for a two-variable model in which daily latency arbitrage profits are regressed on trading volume and realized volatility. Again, to make the results interpretable, realized volatility is multiplied by average daily trading volume.[41] 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 with 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 one-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 sensitivity analysis when we considered different specifications for race detection. Therefore, we emphasize two 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), using both the baseline latency arbitrage tax and the range of latency arbitrage taxes across the various sensitivity analyses discussed in Section V.A.

VI.B. Out-of-Sample Extrapolation: UK Equity Markets

Table XVI presents our estimates of the annual sums at stake in latency arbitrage races in the United Kingdom for the five-year

41. 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 left-hand-side 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 XVI

ANNUAL LATENCY ARBITRAGE PROFITS IN UK EQUITY MARKETS (GBP MILLIONS)

Year	Volume-volatility (1)	Volume-only (2)	Low scenario (3)	High scenario (4)
2014	52.0	56.7	27.1	99.1
2015	58.9	61.6	29.4	107.7
2016	63.3	63.8	30.4	111.4
2017	51.0	57.5	27.4	100.4
2018	55.8	60.6	28.9	105.9

Notes. 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 one-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 XV. Model (2) uses the coefficient from regression (2) in Table XV. Model (3) and Model (4) use the min and max latency arbitrage taxes found in Table XIV, of 0.20 bps and 0.74 bps, respectively.

period 2014–2018. In column (1) we present the estimate based on the volume and volatility regression model. 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. In columns (3) and (4) we present the range of estimates implied by the sensitivity analyses discussed in Section V.A; these are based on latency arbitrage taxes of 0.20 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 £51.0 million to £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 £30 million and at the high end the annual profits are about £100 million.

VI.C. 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. We emphasize 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 (WFE). The advantage of WFE data is that they cover 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, these data do 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 XVI, Table XVII, 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) and (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 in 2018 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 sensitivity analyses, the annual latency arbitrage profits for global equity markets are about $2.3 billion, and at the high end the annual profits are about $8.4 billion.[42]

42. As yet another approach to extrapolation: Virtu recently started publishing global bid-ask spreads data (Virtu 2021). If we take the Virtu spreads data from January 2020, which is the one month in their data that is both globally comprehensive and prepandemic, and we use WFE global equity volumes from that same month, we obtain a global value-weighted effective spread of 3.78 basis points. If we then apply our 16.7% overall reduction in liquidity cost to the nonrace effective spread paid, we get about $6 billion per year instead of about $5 billion per year.

TABLE XVII

ANNUAL LATENCY ARBITRAGE PROFITS IN GLOBAL EQUITY MARKETS IN 2018 (US$ MILLIONS)

Exchange group	Volume-volatility (1)	Volume-only (2)	Low scenario (3)	High scenario (4)
NYSE Group	1,006	1,023	488	1,787
BATS Global Markets - U.S.	895	910	434	1,590
NASDAQ - U.S.	847	862	411	1,505
Shenzhen Stock Exchange	327	336	160	588
Japan Exchange Group	281	286	136	500
Shanghai Stock Exchange	260	268	128	468
Korea Exchange	118	120	57	209
London Stock Exchange Group	109	119	57	207
BATS Chi-X Europe	110	119	57	207
Hong Kong Exchanges and Clearing	102	104	50	182
Euronext	89	96	46	168
Deutsche Börse Group	78	85	40	148
TMX Group	56	61	29	107
National Stock Exchange of India	47	49	24	86
SIX Swiss Exchange	40	43	21	76
Global total (WFE data universe)	4,674	4,799	2,289	8,383

Notes. London Stock Exchange Group includes London Stock Exchange as well as Borsa Italiana. As discussed in the text, this analysis does not attempt to account for differences in market structure across countries and exchanges that may affect the level of latency arbitrage. Rather, its goal is to provide the reader with a sense of global magnitudes. Trading volume is from the World Federation of Exchanges (2021) (WFE). 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 XVI 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).

Because of the COVID-19 pandemic, 2020 was an exceptionally high-volume and high-volatility year for financial markets. Our volume-only model applied to 2020 implies annual latency arbitrage profits in global equity markets of $6.5 billion, while our volume-and-volatility model yields a slightly higher estimate of $7.0 billion. At the low end of our sensitivity analyses, the figure is $3.1 billion and at the high end the figure is $11.4 billion for global equity markets in 2020 (Online Appendix Table E.2).

VII. Conclusion

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

The article's first contribution is methodological: using 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 III.[43]

The article's second—and we think main—contribution is the set of empirical facts we document about latency arbitrage in Section IV. 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 1/10,000th 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. But even just half a tick, over 20% of trading volume, adds 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 nonrace volume. This amounts to about £60 million annually in the United Kingdom. Extrapolating from our UK data,

43. 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 article required about three 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. We expect that future research using message data will be a lot more efficient than our study for at least two reasons. First, our study can be used as a blueprint. Second, some code reoptimization we have incorporated in the code that is posted publicly reduces the computational run time by about 75%.

our estimates imply that latency arbitrage is worth on the order of $5 billion annually in global equity markets alone.

A third contribution, more technical in nature but we hope useful to the 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 market designs that eliminate latency arbitrage would reduce the cost of liquidity for investors by 17%.

One natural direction for future research is to use this article's method for detecting latency arbitrage races to 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, and so on. Such a study could use machine learning methods, treating races as the outcome variable and then trying to understand what preceding market conditions explain the observed races, and would ideally use data across many different exchanges and asset classes to cast a wide net in the search for race triggers.

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 U.S. equities that are more fragmented than the UK market and other kinds of assets such as ETFs, futures, treasuries, 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 here is relatively straightforward.[44]

To our knowledge, most regulators do not currently capture message data from exchanges, and exchanges seem to preserve message data somewhat inconsistently. We hope this will change.

44. To this end, our codebase and a user guide will be made publicly available upon publication of this paper by the *Quarterly Journal of Economics* (https://doi.org/10.7910/DVN/ZFDWDZ) and at https://github.com/ericbudish/HFT-Races. Updates to the code will be provided via the GitHub site. Please contact the authors with any questions or suggested improvements.

Limit order book data have historically been viewed as the official record of what happened in the market, but our study suggests that message data, especially the "error messages" that indicate that a particular participant has failed in their request, are key to understanding speed-sensitive trading.

FINANCIAL CONDUCT AUTHORITY, UNITED KINGDOM AND FINANCIAL STABILITY BOARD, SWITZERLAND
UNIVERSITY OF CHICAGO BOOTH SCHOOL OF BUSINESS AND NATIONAL BUREAU OF ECONOMIC RESEARCH, UNITED STATES
FINANCIAL CONDUCT AUTHORITY, UNITED KINGDOM

SUPPLEMENTARY MATERIAL

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online.

DATA AVAILABILITY

The LSE message data used in this article were obtained by the Financial Conduct Authority under a Section 165 request and cannot be shared publicly. Code and a detailed data appendix can be found in Aquilina, Budish and O'Neill (2021) in the Harvard Dataverse, https://doi.org/10.7910/DVN/ZFDWDZ, and on GitHub at https://github.com/ericbudish/HFT-Races. This code and documentation allows the reader to understand step-by-step how all tables and figures in this article were produced and can be used by future researchers to conduct studies using their own message data sets.

REFERENCES

Acharya, Viral V., and Lasse H. Pedersen, "Asset Pricing with Liquidity Risk," *Journal of Financial Economics*, 77 (2005), 375–410.
Amihud, Yakov, "Illiquidity and Stock Returns: Cross-Section and Time-Series Effects," *Journal of Financial Markets*, 5 (2002), 31–56.
Angel, James J., Lawrence E. Harris, and Chester S. Spatt, "Equity Trading in the 21st Century: An Update," *Quarterly Journal of Finance*, 5 (2015), 1550002.
Aquilina, Matteo, Eric Budish, and Peter O'Neill, "Quantifying the High-Frequency Trading 'Arms Race': A New Methodology and Estimates," Financial Conduct Authority Occasional Paper no. 50, 2020.
———, "Replication Data for: 'Quantifying the High-Frequency Trading "Arms Race"'," (2021), Harvard Dataverse, https://doi.org/10.7910/DVN/ZFDWDZ.
Aquilina, Matteo, Sean Foley, Peter O'Neill, and Thomas Ruf, "Asymmetries in Dark Pool Reference Prices," Financial Conduct Authority Occasional Paper no. 21, 2016.

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

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

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

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

Breckenfelder, Johannes, "Competition among High-Frequency Traders, and Market Quality," European Central Bank Working Paper no. 2290, 2019.

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

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

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

Brunnermeier, Markus K., and Lasse H. Pedersen, "Market Liquidity and Funding Liquidity," *Review of Financial Studies*, 22 (2009), 2201–2238.

Budish, Eric, Gérard P. Cachon, Judd B. Kessler, and Abraham Othman, "Course Match: A Large-Scale Implementation of Approximate Competitive Equilibrium from Equal Incomes for Combinatorial Allocation," *Operations Research*, 65 (2017), 314–336.

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

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

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

Castillo, Juan Camilo, Amrita Ahuja, Susan Athey, Arthur Baker, Eric Budish, Tasneem Chipty, Rachel Glennerster, Scott Duke Kominers, Michael Kremer, and Greg Larson et al., "Market Design to Accelerate COVID-19 Vaccine Supply," *Science*, 371 (2021), 1107–1109.

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

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

Cochrane, John, "Volume and Information," Technical report, 2016. https://johnhcochrane.blogspot.com/2016/10/volume-and-information.html.

Commodity Futures Trading Commission, "Concept Release on Risk Controls and System Safeguards for Automated Trading Environments," *Federal Register*, 78(2015), 56541. 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, "The Term Structure of Liquidity Provision," *Journal of Financial Economics*, 136 (2020), 239–259.

Diamond, Douglas W., and Philip H. Dybvig, "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, 91 (1983), 401–419.

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

Engle, Robert F., and Jeffrey R. Russell, "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data," *Econometrica*, 66 (1998), 1127–1162.
European Securities Market Authority, "High-Frequency Trading Activity in EU Equity Markets," 2014, https://www.esma.europa.eu/system/files_force/library/2015/11/esma20141_-_hft_activity_in_eu_equity_markets.pdf.
Financial Conduct Authority, "Algorithmic Trading Compliance in Wholesale Markets," 2018, https://www.fca.org.uk/news/press-releases/fca-publishes-report-supervision-algorithmic-trading.
Foucault, Thierry, Roman Kozhan, and Wing W. Tham, "Toxic Arbitrage," *Review of Financial Studies*, 30 (2016), 1053–1094.
Frazzini, Andrea, Ronen Israel, and Tobias Moskowitz, "Trading Costs," 2018; available from SSRN, https://ssrn.com/abstract=3229719.
Glosten, Lawrence R., "Components of the Bid-Ask Spread and the Statistical Properties of Transaction Prices," *Journal of Finance*, 42 (1987), 1293–1307.
Glosten, Lawrence R., and Lawrence E. Harris, "Estimating the Components of the Bid/Ask Spread," *Journal of Financial Economics*, 21 (1988), 123–142.
Glosten, Lawrence R., and Paul R. Milgrom, "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders," *Journal of Financial Economics*, 14 (1985), 71–100.
Hagströmer, Björn, "Bias in the Effective Bid-Ask Spread," *Journal of Financial Economics*, 142 (2021), 314–337.
Hanson, Samuel G., Anil K. Kashyap, and Jeremy C. Stein, "A Macroprudential Approach to Financial Regulation," *Journal of Economic Perspectives*, 25 (2011), 3–28.
Hasbrouck, Joel, "Measuring the Information Content of Stock Trades," *Journal of Finance*, 46 (1991a), 179–207.
———, "The Summary Informativeness of Stock Trades: An Econometric Analysis," *Review of Financial Studies*, 4 (1991b), 571–595.
Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld, "Does Algorithmic Trading Improve Liquidity?," *Journal of Finance*, 66 (2011), 1–33.
Hong, Harrison, and Jeremy C. Stein, "Disagreement and the Stock Market," *Journal of Economic Perspectives*, 21 (2007), 109–128.
ICE Futures, "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)," 2019, https://www.cftc.gov/sites/default/files/2019-02/ICEFuturessPassiveOrder020119.pdf?mod=article_inline.
Indriawan, Ivan, Roberto Pascual, and Andriy Shkilko, "On the Effects of Continuous Trading," 2020; available from SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3707154.
Investors' Exchange, "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, 2015, File no. 10-222. https://www.sec.gov/rules/other/2015/investors-exchange-form-1.htm.
Joint Staff Report, "Joint Staff Report: The U.S. Treasury Market on October 15, 2014," 2015. 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. https://www.treasury.gov/press-center/press-releases/Documents/Joint_Staff_Report_Treasury_10-15-2015.pdf.
Jones, Charles M., "What Do We Know about High-Frequency Trading?," Columbia Business School Research Paper 13-11, 2013.
Kominers, Scott D., Alexander Teytelboym, and Vincent P. Crawford, "An Invitation to Market Design," *Oxford Review of Economic Policy*, 33 (2017), 541–571.
Korajczyk, Robert A., and Dermot Murphy, "High-Frequency Market Making to Large Institutional Trades," *Review of Financial Studies*, 32 (2019), 1034–1067.
Kyle, Albert S., "Continuous Auctions and Insider Trading," *Econometrica*, 53 (1985), 1315–1335.

Kyle, Albert S., and Jeongmin Lee, "Toward a Fully Continuous Exchange," *Oxford Review of Economic Policy*, 33 (2017), 650–675.
Lewis, Michael, *Flash Boys: A Wall Street Revolt* (New York: Norton, 2014).
Li, Sida, Xin Wang, and Mao Ye, "Who Provides Liquidity and When?," *Journal of Financial Economics*, 141 (2021), 968–990.
London Metals Exchange, "Technical Change to LMEselect FIX Message Processing for the LMEprecious Market to Introduce a Fixed Minimum Delay," 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, *Trading Services Price List (On-Exchange and OTC)*, 2015. https://docs.londonstockexchange.com/sites/default/files/documents/trading_services_price_list-11_may_2015_0.pdf.
MacKenzie, Donald, "How Fragile Is Competition in High-Frequency Trading," Tabbforum, March 26, 2019. https://tabbforum.com/opinions/how-fragile-is-competition-in-high-frequency-trading/.
———, *Trading at the Speed of Light: How Ultrafast Algorithms Are Transforming Financial Markets* (Princeton, NJ: Princeton University Press, 2021).
Malinova, Katya, Andreas Park, and Ryan Riordan, "Do Retail Traders Suffer from High Frequency Traders?," 2018, available from SSRN, https://ssrn.com/abstract=2183806.
Mamudi, Sam, "Virtu Touting Near-Perfect Record of Profits Backfired, CEO Says," *Bloomberg*, June 4, 2014. https://www.bloomberg.com/news/articles/2014-06-04/virtu-touting-near-perfect-record-of-profits-backfired-ceo-says?sref=OlHJrDFy.
Menkveld, Albert J., "High Frequency Trading and the New Market Makers," *Journal of Financial Markets*, 16 (2013), 712–740.
———, "The Economics of High-Frequency Trading: Taking Stock," *Annual Review of Financial Economics*, 8 (2016), 1–24.
Michaels, Dave, "Chicago Stock Exchange Targets Rapid-Fire Traders With Speed Bump, Echoing IEX," *Wall Street Journal*, August 30, 2016. https://www.wsj.com/articles/chicago-stock-exchange-targets-rapid-fire-traders-with-speed-bump-echoing-iex-1472591832.
Milgrom, Paul, "Auction Research Evolving: Theorems and Market Designs," *American Economic Review*, 111 (2021), 1383–1405.
New York Attorney General's Office, "Remarks on High-Frequency Trading & Insider Trading 2.0," 2014, New York Law School Panel on "Insider Trading 2.0 - A New Initiative to Crack Down on Predatory Practices." https://ag.ny.gov/pdfs/HFT_and_market_structure.pdf.
O'Hara, Maureen, "High Frequency Market Microstructure," *Journal of Financial Economics*, 116 (2015), 257–270.
Osipovich, Alexander, "High-Frequency Traders Eye Satellites for Ultimate Speed Boost," *Wall Street Journal*, April 1, 2021. https://www.wsj.com/articles/high-frequency-traders-eye-satellites-for-ultimate-speed-boost-11617289217.
Pástor, Ľuboš, and Robert F. Stambaugh, "Liquidity Risk and Expected Stock Returns," *Journal of Political Economy*, 111 (2003), 642–685.
Pathak, Parag A., "What Really Matters in Designing School Choice Mechanisms," *Advances in Economics and Econometrics*, 1 (2017), 176–214.
Ring, John H. IV, Colin M. Van Oort,, David R. Dewhurst, Tyler J. Gray, Christopher M. Danforth, and Brian F. Tivnan, "Scaling of Inefficiencies in the U.S. Equity Markets: Evidence from Three Market Indices and More than 2900 Securities," arXiv Preprint, arXiv:1902.04691 (2019).
Roth, Alvin E., "The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics," *Econometrica*, 70 (2002), 1341–1378.
———, "What Have We Learned from Market Design?," *Innovations: Technology, Governance, Globalization*, 3 (2008), 119–147.

———, "Marketplaces, Markets, and Market Design," *American Economic Review*, 108 (2018), 1609–1658.
Roth, Alvin E., Tayfun Sönmez, and M. Utku Ünver, "Kidney Exchange," *Quarterly Journal of Economics*, 119 (2004), 457–488.
Securities and Exchange Commission, "Concept Release on Equity Market Structure," Release no. 34-61358, 2010, File no. S7-02-10. https://www.sec.gov/rules/concept/2010/34-61358.pdf.
Shkilko, Andriy, and Konstantin Sokolov, "Every Cloud Has a Silver Lining: Fast Trading, Microwave Connectivity and Trading Costs," *Journal of Finance*, 75 (2020), 2899–2927.
Shleifer, Andrei, and Robert W. Vishny, "Liquidation Values and Debt Capacity: A Market Equilibrium Approach," *Journal of Finance*, 47 (1992), 1343–1366.
———, "The Limits of Arbitrage," *Journal of Finance*, 52 (1997), 35–55.
———, "Fire Sales in Finance and Macroeconomics," *Journal of Economic Perspectives*, 25 (2011), 29–48.
Stoll, Hans R., "Inferring the Components of the Bid-Ask Spread: Theory and Empirical Tests," *Journal of Finance*, 44 (1989), 115–134.
———, "Friction," *Journal of Finance*, 55 (2000), 1479–1514.
Van Kervel, Vincent, and Albert J. Menkveld, "High-Frequency Trading around Large Institutional Orders," *Journal of Finance*, 74 (2019), 1091–1137.
Virtu "Virtu Financial Global Equities Market Structure Monthly," 2021. https://www.virtu.com/thinking/thought-leadership.
Wah, Elaine, "How Prevalent and Profitable Are Latency Arbitrage Opportunities on US Stock Exchanges?," 2016, available from SSRN, https://ssrn.com/abstract=2729109.
Weller, Brian M., "Does Algorithmic Trading Reduce Information Acquisition?," *Review of Financial Studies*, 31 (2018), 2184–2226.
World Federation of Exchanges, "World Federation of Exchanges Database," 2021. https://www.world-exchanges.org/our-work/statistics.
Yao, Chen, and Mao Ye, "Why Trading Speed Matters: A Tale of Queue Rationing under Price Controls," *Review of Financial Studies*, 31 (2018), 2157–2183.
Zuckerman, Gregory, *The Man Who Solved the Market: How Jim Simons Launched the Quant Revolution* (New York: Penguin 2019).