A Theory of Stock Exchange Competition and Innovation: Will the Market Fix the Market?*

Eric Budish[†], Robin S. Lee[‡], and John J. Shim[§]

April 2023

Abstract

Will stock exchanges innovate to address latency arbitrage and the arms race for speed? This paper models how exchanges compete in the modern electronic era and how this shapes incentives for market design innovation. In the status quo, exchange trading fees are competitive while exchanges earn economic rents from selling speed. These rents create a wedge between private and social incentives to innovate, and support the persistence of an inefficient market design in equilibrium of a market design adoption game. We discuss implications for policy and insights for the literatures on market design, innovation, and platforms.

Keywords: market design, innovation, financial exchanges, industrial organization, platform competition, high-frequency trading

*An early version of this research was presented in the 2017 AEA/AFA joint luncheon address. We are extremely grateful to the colleagues, policymakers, and industry participants with whom we have discussed this research over the last several years. Special thanks to Larry Glosten, Terry Hendershott and Jakub Kastl for providing valuable feedback as conference discussants, and to the editor and three anonymous referees for comments that significantly improved the paper. We are also grateful to Jason Abaluck, Nikhil Agarwal, Susan Athey, John Campbell, Dennis Carlton, Judy Chevalier, John Cochrane, Christopher Conlon, Shane Corwin, Peter Cramton, Doug Diamond, David Easley, Alex Frankel, Joel Hasbrouck, Kate Ho, Anil Kashyap, Pete Kyle, Donald Mackenzie, Neale Mahoney, Paul Milgrom, Joshua Mollner, Ariel Pakes, Al Roth, Fiona Scott Morton, Sophie Shive, Andrei Shleifer, Jeremy Stein, Mike Whinston, Heidi Williams and Luigi Zingales for valuable discussions, and to seminar audiences at Chicago, Yale, Northwestern, NYU, Berkeley, Harvard, MIT, NBER Market Design, UPenn, Columbia, HKUST, KER, the Economics of Platforms Workshop, NBER IO, SEC DERA, and the AFAs. Paul Kim, Cameron Taylor, Matthew O'Keefe, Natalia Drozdoff, and Ethan Che provided excellent research assistance. Budish acknowledges financial support from the Fama-Miller Center, the Stigler Center, and the University of Chicago Booth School of Business. Disclosure: the authors declare that they have no relevant or material financial interests that relate to the research described in this paper. John Shim worked at Jump Trading, a high-frequency trading firm, from 2006-2011.

[†]University of Chicago Booth School of Business and NBER, eric.budish@chicagobooth.edu

[‡]Harvard University and NBER, robinlee@fas.harvard.edu

[§]University of Notre Dame, jshim2@nd.edu

1 Introduction

Market design research usually focuses on designing the best possible market mechanism for a given problem.[1] This paper concerns a different, complementary, question. Suppose researchers have already designed an attractive mechanism—will it actually get adopted?

The context of the study is market design for financial exchanges. Financial exchanges are clearly important—they serve vital roles in the global economy generating price signals, helping companies and governments raise capital, and sharing risk. Stock exchanges alone execute over $200 trillion of transaction volume per year. Recent research has shown that the predominant financial exchange design used around the world, called the continuous limit order book, has an important design flaw. By treating time continuously, the market design gives rise to a phenomenon called "latency arbitrage," or arbitrage rents from symmetrically disseminated public information—rents that in principle are not supposed to exist in an efficient market, as opposed to the rents from asymmetric private information that are at the heart of classic models of market microstructure (Kyle, 1985; Glosten and Milgrom, 1985). These latency arbitrage rents, in turn, cause a socially wasteful arms race for speed and harm market liquidity. Latency arbitrage races are currently measured in millionths and even billionths of seconds, and have been estimated as generating about 20% of all trading volume and harming liquidity by from about 17% to 33% depending on the measure used. A simple (in theory) market design reform, that puts time into small discrete increments and batch processes trade requests that arrive at the "same time," would solve the problem.[2]

To date, while there has been a fair amount of innovative activity that is in some way related to latency arbitrage, it is fair to say that large incumbent financial exchanges have not embraced the new market design reform.[3] This paper tries to understand how financial exchanges compete and how this shapes their incentives to innovate—and ultimately whether exchanges' private incentives for market design innovation align with what is socially efficient. Will the market fix the market?

As is well known in the fields of industrial organization and innovation economics, there is no one answer to whether private and social innovation incentives align. The usual case, of course, is that if there is a large inefficiency in a market, and a private-sector innovation could address the inefficiency, then the private sector will innovate in a way that aligns with social welfare (Griliches, 1957). But there are many cases where private and social innovation incentives might diverge (Arrow, 1962; Nordhaus, 1969; Hirshleifer, 1971; Mankiw and Whinston, 1986). We will ultimately find that private and social incentives to innovate diverge here as well, for a mix of classical and novel reasons.

To analyze our questions, we build a model that is closely tailored to the institutional details of modern electronic financial exchanges. Some aspects of the model are tailored specifically to regula-

[1] For recent surveys of the market design literature, see Roth (2018), Milgrom (2021) and Agarwal and Budish (2021).

[2] See Budish, Cramton and Shim (2015) for the definition of latency arbitrage and the market-design reform of frequent batch auctions. Another market-design reform that can solve the latency arbitrage problem is an asymmetric speed bump; see Baldauf and Mollner (2020). See Aquilina, Budish and O'Neill (2022) for empirical magnitudes of latency arbitrage. See Indriawan, Pascual and Shkilko (2022) for an empirical comparison of continuous- and discrete-time trading.

[3] See discussion of innovation activity to date in Section 4.1 and Appendix H.

tions for the U.S stock market, which is both a canonical financial market and has been at the heart of the controversy around latency arbitrage. The players in our model are exchanges and three kinds of market participants: trading firms, informed traders and uninformed investors. Exchanges make a market design decision and set prices—prices for trading per se and prices for what we call speed technology. The trading firms decide whether to pay for speed technology and then, together with the other market participants, play a trading game. Two important details of the trading game are that stocks are fungible across exchanges (due to regulation called Unlisted Trading Privileges), and market participants are able to engage in frictionless search across exchanges (due to regulations such as Regulation National Market System in the U.S.). As will become clear, these details make stock exchange competition very different from other familiar forms of platform competition.

We start by studying equilibrium in the subgame where all exchanges choose the status quo market design ("Continuous"). We find that trading fees are perfectly competitive but that exchanges are able to extract rents from speed technology. The reason why trading fees are competitive is that frictionless search leads to Bertrand-like competition over the net price of a trade, and hence on trading fees. This stands in contrast to many other platform markets, which have network effects and supra-competitive transaction fees. The intuition for why speed technology fees are *not* competitive is that, if there is a speed-sensitive trading opportunity on a particular exchange, only that exchange's speed technology is useful for the opportunity. For example, Nasdaq speed technology is not useful for latency-arbitrage opportunities on the New York Stock Exchange, and vice versa. This creates market power. These outcomes of our theory model align with empirical facts that we document about trading fees and speed-technology revenues. The average trading fee in the U.S. stock market is just \$0.0001 per-share per-side, or just 0.0001% of a \$100 share of stock.[4] Further, speed technology revenues are several times larger than trading fee revenues and have been growing rapidly in the modern electronic era.

We then study equilibria in subgames where one or more exchanges use a market design that addresses latency arbitrage ("Discrete"). We obtain two sets of results. If there is a *single* exchange that adopts Discrete, then this exchange would win share and be able to charge supra-competitive trading fees, in any equilibrium. The same frictionless search that caused trading fees to be brutally competitive under the status quo enables an exchange with a better market design to get off the ground in any equilibrium. This too is in sharp contrast to many other platform settings, where there are chicken-and-egg equilibria in which the new market design can stay stuck with zero share, even if in principle it is better designed. Moreover, the exchange can charge a supra-competitive fee commensurate with the latency-arbitrage savings it creates. Intuitively, the innovator is getting compensated for solving the problem, as in the classic case where private and social incentives align.

However, if *multiple* exchanges use the new market design—as would be the case if there were a regulatory mandate or if the initial innovator is imitated—then trading fees become perfectly compet-

[4]This is significantly different from other platform markets with even modest search frictions. For example, internet-enabled platform markets for items such as tickets, ride-sharing, food delivery, and vacation rentals commonly have fees of about 10-30%, or about 100,000 times higher on a percentage basis. While there are many reasons why this comparison is not apples-to-apples (e.g., fraud costs), the contrast is nonetheless striking.

itive again, like under the status quo, but now exchanges no longer capture speed technology rents. Therefore, all exchanges are worse off than under the status quo.

This set of results has two major implications for the question of whether the market will fix the market. First, it implies that incumbent exchanges strictly prefer the status quo to a counterfactual in which they all use a market design that addresses latency arbitrage. Moreover, the structure of payoffs in the market-design adoption game is a prisoner's dilemma—a single Discrete exchange profits unilaterally, but if there are multiple Discrete exchanges they earn zero profits, and hence are worse off than under the status quo. We show formally that incumbent exchanges can maintain "cooperation" at Continuous as an equilibrium of the repeated game. This finding accords with the record to date.

Second, it implies that *if* there is an innovator, it would actually work. The new market design would gain share and help the market fix the market. The difficulty is not that the new market design would not get off the ground, as in other platform environments, but the lack of economic incentive. The same frictionless search that helps the innovator overcome the chicken-and-egg problem and get off the ground also makes the innovator very vulnerable to imitation and with that perfect competition.

This in turn has an important implication for policy. A natural prior coming into this analysis is that the relevant question for policy is whether (i) there will be a private-market solution to latency arbitrage and the arms race, or (ii) would some sort of market-design mandate be required to fix the problem. This is how the SEC Chair framed the issue in a 2014 speech, and the SEC Chair expressed reticence to impose a mandate.[5] In our analysis, a mandate to "fix the market" would certainly work, but there is a third option: a regulatory push. By mandate we mean requiring exchanges to play Discrete. By push we mean any policy that tips the balance of incentives sufficiently to entice a first adopter to choose to play Discrete. Two specific pushes we discuss are a modest exclusivity period or reducing risk-adjusted entry costs. Back-of-the-envelope calculations suggest that the magnitude of the push could be very modest relative to the stakes.

Zooming out, we emphasize novel insights from our study for three broader literatures. For the platforms literature, our contribution is the idea that market participants can stitch together multiple exchanges into a "virtual single platform" when there is frictionless search. This insight has important implications for other platform markets where search frictions could, in principle, be eliminated by regulation or technology. For the innovation literature, our study identifies a novel wedge between private and social innovation incentives: incumbent rents that arise from inefficiency in the status quo (cf. Bryan and Williams, 2021). For the market design literature, our study opens new ground by studying the question of whether a new market design will actually get implemented by the private sector. Our study also brings to the market design literature some classical themes from economics, such as incumbents protecting rents and issues of concentrated versus dispersed interests (Olson, 1965).

The remainder of this paper is organized as follows. Section 2 describes the empirical facts and

[5]SEC Chair Mary Jo White, in a 2014 speech, said: "I am personally wary of prescriptive regulation that attempts to identify an optimal trading speed, but I am receptive to more flexible, competitive solutions that could be adopted by trading venues. These could include frequent batch auctions or other mechanisms designed to minimize speed advantages."

Table 2.1: Estimate of Average Regular-Hours Trading Fees, U.S. Equities

Exchange Group	f
BATS	\$0.000089
NASDAQ	\$0.000105
NYSE	\$0.000128

Notes: The table reports our estimate of the average regular-hours trading fee per share per side for each of the three major exchange families in the U.S. stock market. Data are exchange financial filings and fee schedules from fiscal-year 2015. Please see Appendix A.1 and the associated spreadsheet for supporting details.

regulations that motivate the theoretical model. Section 3 presents the theoretical analysis. Section 4 discusses policy implications. Section 5 concludes.

2 Institutional Background and Motivating Facts

This section documents three stylized facts about the economics of modern electronic stock exchanges: trading fees are small; exchanges earn significant and growing revenue from co-location and proprietary data feeds, which are forms of speed technology; and exchange market shares are interior and relatively stable over time. Together, these facts are at odds with exchange competition models, such as the seminal contribution of Pagano (1989), in which liquidity externalities can lead traders to agglomerate on a single exchange with supra-competitive trading fees. This section then discusses two key regulations that underlie modern stock exchange competition and are central to our theoretical model: Regulation National Market System (Reg NMS) and Unlisted Trading Privileges (UTP).

2.1 Exchange Trading Fees

Exchange trading fees are notoriously complicated.[6] Underneath this complexity, however, we find that exchange trading fees are economically small. The average regular-hours trading fee is just \$0.0001 per share per side (Table 2.1). For a \$100 share of stock, this average fee in percentage terms is just 0.0001%. We reach this conclusion using a combination of historical exchange fee schedules and exchange company financial filings. A challenge in this analysis is that each of the main exchange companies controls multiple exchanges. Appendix A.1 details how we use the financial filings and fee schedules to obtain an overall average fee for each exchange company.

Moreover, most exchanges charge fees that on average are slightly *negative* to participants with high-enough trading volume. This is consistent with exchanges being willing to lose money on trading fees to make money from other sources such as speed technology. That said, trading fees are not negative enough so that a market participant can extract revenue from the exchange by trading at the negative fee, once one accounts for regulatory fees charged by the SEC and FINRA. Appendix A.2

[6]Figure A.1 in the Appendix, from an investment bank research report, presents a tongue-in-cheek visualization of this complexity by depicting the hundreds of different fee scenarios that are possible for a particular trade.

Table 2.2: Estimate of Exchange Speed-Technology Revenue, U.S. Equities
($ millions, fiscal year 2015)

	BATS	NASDAQ	NYSE	Total
Market Data Revenue	114.1	222.4 – 267.3	218.9 – 241.5	555.4 – 623.0
Co-Location/Connectivity Revenue	64.3	121.0 – 139.0	251.6 – 281.5	436.8 – 484.8
Market Data + Co-Location Revenue	178.4	343.3 – 406.4	470.5 – 523.0	992.2 – 1107.8
CTA/UTP Tape Revenue				317.0
Market Data + Co-Lo Revenue net of Tape Revenue				675.2 – 790.8

Notes: The table reports speed-technology revenue for each of the three major exchange families in the U.S. stock market. Data are financial filings that cover fiscal-year 2015 and a CTA fee-change filing to the SEC. We estimate Nasdaq and NYSE revenue and report a range (BATS revenue comes directly from filings). Please see Appendix B.1 for supporting details.

provides further details.

Many models of platform competition have high fees in equilibrium reflecting market power from network effects. These data suggest a different model is needed.

2.2 Exchange Data and Co-Location Revenue

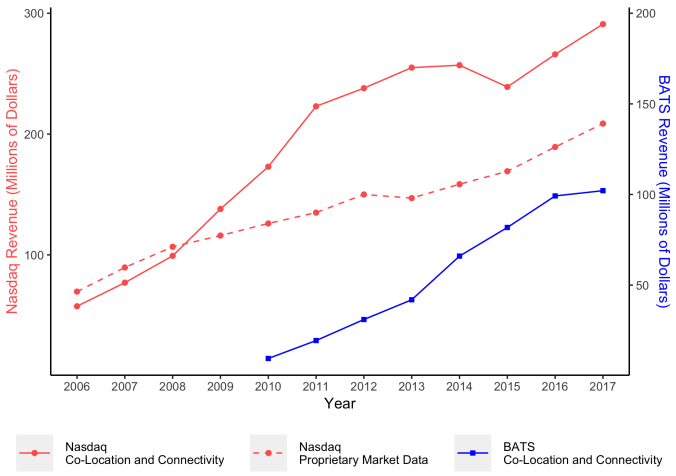
Exchanges earn revenue from selling proprietary fast data feeds and from selling a service called "co-location," which provides the right to locate one's own computer servers near the exchange's computer servers. These services provide a speed advantage to speed-sensitive traders.

Intuitively, exchanges have some market power over these services because the speed advantage is specific to a particular exchange. For example, only Nasdaq can sell the right to co-locate next to Nasdaq's servers, and Nasdaq co-location only provides a speed advantage for trading opportunities on Nasdaq. We will call these services "speed technology."

Public reporting about exchange speed-technology revenues is opaque.[7] We used a variety of sources of information, including exchange company 10-K filings, BATS's April 2016 initial public offering filing (form S-1), documentation related to NYSE's acquisition by ICE, and data from the entity that reports revenues from slower non-proprietary data feeds to obtain an estimate of overall exchange speed-technology revenues and their growth over time. Data availability are best for fiscal year 2015 because of the timing of BATS's initial public offering and NYSE's acquisition. We estimate that 2015 exchange speed-technology revenues across the three main exchange families were between $675-$790M (Table 2.2). We are also able to build meaningful time-series for some components of speed-technology revenue (Figure 2.1). We compute annual revenue growth rates of 16% for Nasdaq co-location/connectivity (2006-2017), 11% for Nasdaq proprietary market data (2006-2017), and 40% for BATS co-location/connectivity (2010-2017).[8] If we utilize 10% as a conservative overall growth rate since 2015, this implies annual exchange speed-technology revenues of $1.3-$1.5B in 2022.

[7] Jackson, Robert J., Jr., "Unfair Exchange: The State of America's Stock Markets," September 19, 2018.
[8] See Appendix B.2 for supporting details.

Figure 2.1: Growth in Exchange Speed-Technology Revenue: 2006-2017



Notes: Nasdaq data come from 2006-2017 10-K filings. BATS co-location/connectivity revenue data come from the 2012 S-1 filing (years 2010–2011), the 2016 S-1 filing (2012–2015), the 2016 CBOE/BATS Merger Proxy and the CBOE 2017 10-K. We omit BATS proprietary market data revenue from the figure because BATS only started charging for proprietary data in Q3 2014. For each time series we use the reporting category that contains U.S. equities revenue for that revenue source and make consistent assumptions over time to isolate estimated revenues from U.S. equities. Please see Appendix B.2 for further discussion of the data and methodology.

Overall, these data are suggestive of exchanges discovering a new source of revenue related to speed-sensitive trading. This new source of revenue plays an important role in our theoretical analysis.

2.3 Exchange Market Shares

Figure C.1 in Appendix C shows that exchange market shares are interior and relatively stable over time. There are 8 exchanges with meaningful market share, with the highest among them at 25%. As one simple measure of stability, if we regress the market share of each exchange-date on only a set of exchange fixed effects, the R^2 is 0.97. In the Appendix, we also show that exchange market shares are interior and relatively stable over time at the individual symbol level too.

While many models of platform competition have "tipping" (aka "winner take all") as a potential equilibrium outcome, it is clear from these data that this is not the case for U.S. stock exchanges.

2.4 Key Regulations

There are two key sets of regulations that help make sense of these empirical patterns and are central to our model. We describe them briefly here and provide further details in Appendix D.

The first set of regulations, Unlisted Trading Privileges (UTP), has its roots in the 1934 Exchange Act and in its modern incarnation enables all stocks to trade on all exchanges, essentially independently of where the stock is technically listed, with the exception of the opening and closing auctions which are proprietary to the listing exchange. For the purposes of our theoretical model, we incorporate UTP

in its current form by assuming that the security in the model is perfectly *fungible* across exchanges. This captures that regardless of where a security is listed, was last traded, etc., it can be bought or sold on any exchange.

The second, Regulation National Market System (Reg NMS), is a long and complicated piece of regulation implemented in 2007. For the purpose of the present paper, however, there are two core features to highlight. The first is the Order Protection Rule, or Rule 611. The Order Protection Rule prohibits an exchange from executing a trade at a price that is inferior to the best price on another exchange (called a "protected quote"). The second is the Access Rule, or Rule 610. Intuitively, in order to comply with the Order Protection Rule, exchanges and market participants must be able to efficiently obtain the necessary information about quotes on other exchanges and efficiently trade against them ("access" the protected quote). The Access Rule, and related rules that affect information provision, ensures that such efficient search and access is feasible. For our theoretical model, we capture these key provisions of Reg NMS by assuming what we will call *frictionless search and access*, on an order-by-order basis. That is, there is zero marginal cost of search across all exchanges, and there are zero additional marginal costs (beyond per-share trading fees) of accessing liquidity on a particular exchange or exchanges.

Intuitively, the combination of fungibility of assets across exchanges and frictionless search across exchanges nullifies the market power traditionally associated with network effects and platform markets. This provides some intuition for the low trading fees and interior market shares that we observe above and will capture in our model. Any model that does not take these key regulations seriously will misunderstand the industrial organization of the market.

3 Theoretical Analysis

The goal of the model is to illuminate the economic forces that underlie exchanges' incentives for market design innovation. What are the incentives to adopt a market design that addresses latency arbitrage and the arms race for trading speed? Also of theoretical interest is our model's characterization of the economics of exchange competition under the status quo market design.

The model consists of four kinds of players, all strategic: exchanges, trading firms, investors, and informed traders. They play a game with the following timing. First, exchanges choose their market designs. Next, exchanges set prices for trading and for speed technology. Third, trading firms decide from which exchanges, if any, to purchase speed technology. Last, the trading firms, investors, and informed traders play a repeated trading game.

The trading game played in the last stage of the model is a generalization of the Budish, Cramton and Shim (2015) model. The primary generalization is to multiple exchanges, in a competitive environment shaped by the key regulations discussed in Section 2. The trading game model also adds a stylized version of informed trading, in the spirit of Copeland and Galai (1983) and Glosten and Milgrom (1985), to parsimoniously incorporate adverse selection alongside latency arbitrage.

The section is organized as follows. Sections 3.1 and 3.2 present the formal description of the model and equilibrium solution concept. Section 3.3 provides a brief overview to the equilibrium analysis. Sections 3.4-3.6 analyze equilibria under subgames corresponding to various combinations of market design decisions. Section 3.7 synthesizes the results.

3.1 Formal Description of the Model

We describe the players and their actions in Section 3.1.1, the formal game timing in Section 3.1.2, payoffs in Section 3.1.3, and how the game incorporates key institutional details in Section 3.1.4.

3.1.1 Players

Exchanges. There are $M \geq 2$ exchanges, indexed by j. Exchanges are ex-ante undifferentiated and each make three strategic choices: (i) their market design, (ii) their trading fees, and (iii) their exchange-specific speed technology fees.

For the choice of market design, we focus on two options: the continuous limit order book (Continuous) and frequent batch auctions (Discrete). The continuous limit order book represents the status quo market design. At a high level, the continuous limit order book processes "limit orders"—i.e., messages specifying a price, quantity, and whether to buy or sell—or cancellations of previously submitted limit orders in a *serial* fashion, i.e., one at a time in order of their receipt.[9] The frequent batch auctions design is similar to the Continuous design in many respects, with the key difference being the way that it processes new messages. Instead of processing new messages (including cancellation requests) serially, frequent batch auctions process new messages in a batch process, in frequent pre-specified discrete-time intervals, using a uniform-price auction. We provide further details and formalize this difference in the description of the trading game below.

Exchange trading fees, denoted f_j, are assessed per share traded and are paid symmetrically by both sides of any executed trade. In practice, exchanges often charge different fees depending on whether the order was the one resting in the limit order book ("making" liquidity) or the order was the one that executed against a resting order ("taking" liquidity).[10]

Exchange-specific speed technology (abbreviated ESST) fees, denoted F_j, represent the price for technology that allows a particular market participant to trade faster at a particular exchange. In the trading game stage of our model, we treat ESST as a tie-breaker meaning that if multiple market participants submit messages to the same exchange at the same time, and the exchange processes messages serially, then the ones with ESST are processed first. If the exchange processes messages in batch, ESST does not provide any advantage.[11] In practice, speed technology includes co-location

[9] Please see Harris (2002) for further details regarding the continuous limit order book design.

[10] The assumption of symmetric fees is without loss of generality since we assume that prices are continuous. Under this assumption, only the net trading fee matters for determining equilibrium behavior (see Chao, Yao and Ye (2019)).

[11] Practically, we have in mind that frequent batch auction exchanges would allow market participants to co-locate their servers and subscribe to proprietary market data, but would not be able to charge prices commensurate with their role, on continuous exchanges, in extracting sniping rents. For example, as of a few years ago Nasdaq offered four different

(the right to locate one's own servers next to an exchange's servers), access to fast exchange-specific proprietary data feeds, and connectivity/bandwidth fees. ESST fees are modeled as a rental cost paid per trading game, capturing that in practice exchanges typically assess these fees on a rental basis.

We require that at least two trading firms purchase speed technology on an exchange for it to be used on that exchange. This is a modest fair access requirement which prevents an exchange from auctioning off exclusive access to fast trading on its exchange.[12]

Market Participants. There are three other kinds of players in the model—(uninformed) investors, informed traders, and trading firms—whom we collectively refer to as *market participants*. Market participants play a repeated trading game that builds off of the Budish, Cramton and Shim (2015) model. Though the trading game is described in full below, we describe some aspects here to describe market participants' actions. In the trading game there is a single security, x, and the fundamental value of the security is given by y. We make the purposefully strong assumption that x can always be costlessly liquidated at this fundamental value. The value y evolves across trading games as a discrete-time jump process, where in each trading game there is a probability of a jump in y and the value of jumps is drawn from a symmetric distribution with bounded support and zero mean. What will matter economically is the distribution of the absolute value of jumps, represented by random variable J. The same security trades on all M exchanges, and its value does not depend on the exchange on which it is traded. We assume that prices are continuous and that shares are perfectly divisible.[13]

An *Investor* arrives stochastically with probability λ_{invest} in each trading game, with an inelastic need to buy or sell one unit of the security. Needing to buy or needing to sell are equally likely.[14] The investor trades a single time, potentially across multiple exchanges, and then exits the game.

An *Informed Trader* observes private information about the fundamental value of x. We assume that in each trading game, the probability that there is a jump in y that is public information and seen by all players at the same time is λ_{public} and the probability that there is a jump in y seen by an informed trader is $\lambda_{private}$. For simplicity, both public and private jumps have the same jump-size distribution J. If an informed trader observes a jump in y, he can trade on that information in the current trading game; regardless of the informed trader's actions, at the conclusion of the trading game any privately observed information becomes public.

---

levels of co-location services, with the most expensive version about 2 microseconds (0.000002 seconds) faster than the least expensive version, and about 10 times the price (IEX, "Re: Investors' Exchange LLC Form 1 Application (Release No. 34-75925; File No. 10-222)," 2015). A frequent batch auction exchange might be able to sell something akin to the cheapest version, but would not be able to extract rents from latency arbitrage by selling an ever-so-slightly faster connection.

[12] For example, former SEC Chair Jay Clayton emphasized that it has long been required, under the [1934] Exchange Act, that exchange fees be "fair and reasonable and not unreasonably discriminatory."

[13] Assuming continuous prices allows us to abstract from the queueing dynamics that are present in markets with binding tick-size constraints. Assuming that shares are perfectly divisible allows for any agent to split his desired order, regardless of size, across multiple exchanges. It is substantively important for the analysis, and also realistic, that agents can split orders across multiple exchanges.

[14] As in Budish, Cramton and Shim (2015, pgs. 1583-1586), it is straightforward to generalize the model to investors with varying-sized demands, as long as all investors trade a single time upon arrival.

Trading Firms, abbreviated as TFs and present throughout all iterations of the trading game, have no intrinsic demand to buy or sell x; rather they seek to buy x at prices lower than y and vice versa. Their objective is to maximize per-trading game profits (described below). We assume that there are N "fast" TFs and a continuum of "slow" TFs, where the difference reflects how messages are processed by the Continuous market design in a manner we formalize below.[15]

3.1.2 Formal Game Timing

Our game has four stages.

First, in Stage 1 (*Market Design Choice)*, all M exchanges simultaneously choose whether to operate a Continuous or Discrete market design.

Second, in Stage 2 *(Exchange Price Setting)*, all M exchanges simultaneously choose per-share trading fees $\mathbf{f} = (f_1, \ldots, f_M)$, and per-trading game ESST fees $\mathbf{F} = (F_1, \ldots, F_M)$.

Next, in Stage 3 *(Speed Technology Adoption),* all N TFs with general speed technology simultaneously decide from which exchanges, if any, to purchase ESST. All ESST purchase decisions are publicly observed.

Last, in Stage 4 (*Repeated Trading Game*), the following two-period trading game is played repeatedly T times, where T is a large, finite number. We interpret each trading game as lasting a very short amount of time (e.g., 1 millisecond).[16]

The Trading Game. In period 1 of each trading game, trading firms observe the public *state*, which consists of the current fundamental value of the security (y), and the current outstanding bids and asks in each exchange's limit order book ($\boldsymbol{\omega} = (\omega_1, \ldots, \omega_M)$, where ω_j is also referred to as the state of exchange j's order book).[17] TFs then simultaneously submit *message sets* to exchanges, where $\mu_{ij} \in \mathcal{S}$ represents the messages that TF i submits to exchange j, and $\mathcal{S}$ denotes the set of all potential combinations of messages. Denote by $\boldsymbol{\mu}_i \equiv \{\mu_{ij}\}_{j \in \mathcal{M}}$ the message sets submitted by TF i to all exchanges, where $\mathcal{M}$ represents the set of all exchanges.

Messages sent to each exchange j can affect that exchange's order book ω_j after they are *processed* by each exchange in a manner that we describe below. We allow for three types of messages that TFs can send to a particular exchange j: (i) standard limit orders, which take the form (q_i, p_i) and indicate that the TF is willing to buy (if $q_i > 0$) or sell (if $q_i < 0$) up to $|q_i|$ units at price p_i; (ii) cancellations of existing limit orders in ω_j; and (iii) immediate-or-cancel orders (IOCs), which are standard limit

[15] Note, practically, that what we mean by a slow trading firm is best interpreted as a sophisticated algorithmic trading firm not at the very cutting edge of speed, but still fast by non-high-frequency trading standards.

[16] We initially analyze a single play of our game—i.e., Stages 1, 2 and 3 are played once, and then the trading game in Stage 4 is played T times. Later, in Section 3.7, we examine repeated play of all four stages—i.e., after T iterations of the Stage 4 trading game are played, play returns to Stage 1 and the game repeats. When our game is repeatedly played, we interpret T to be large enough so that it captures the appropriate time horizon for relatively slower-moving decisions about market design and trading fees; e.g., T represents the equivalent of several months worth of millisecond-long trading games. Unlike in many other economic environments, stock exchange fee changes and market design changes require regulatory rule filings, so they cannot adjust without delay.

[17] Each exchange's order book is initially empty.

orders that, if not fully executed in a given period, have any portion that is remaining cancelled by the exchange at the end of the period. Standard limit orders remain in an exchange's order book across trading games until they are fully executed (i.e., all units q_i are traded against), or they are cancelled by a cancellation message in which case they are removed. A TF is also allowed to send no messages to a particular exchange j in a given period, in which case the TF simply maintains its existing limit orders in ω_j, if any exist.

We say that a TF *provides liquidity* if it offers to buy (or sell) some positive quantity at a price less than (or greater than) the current value of y and within the support of J. Since investors are equally likely to arrive needing to buy or sell and the distribution of jumps in y is symmetric about zero, it is convenient to focus on the provision of liquidity via pairs of limit orders: that is, for a given quantity q and fundamental value y, an order to buy quantity q at $y - \frac{s}{2}$ and an order to sell quantity q at $y + \frac{s}{2}$, where $s \geq 0$ represents the *bid-ask spread*.

After Period-1 message sets are submitted by TFs, they are processed by each exchange and each exchange j's updated order book ω_j is publicly observed. Next, in period 2, nature moves and selects one of four possibilities:

1. With probability λ_{invest}: an investor arrives, equally likely to need to buy or sell one unit of x. The investor has a single opportunity to send IOCs to all exchanges.

2. With probability $\lambda_{private}$: an informed trader privately observes a jump in y. The informed trader has a single opportunity to send IOCs to all exchanges.

3. With probability λ_{public}: there is a publicly observable jump in y. All TFs have a single opportunity to submit message sets consisting of IOCs and cancellation messages to each exchange.

4. With probability $1 - \lambda_{invest} - \lambda_{private} - \lambda_{public} \geq 0$: there is no event.

Period-2 messages are then processed, exchanges' order books are updated and publicly observed, and trading game payoffs are realized.

How the Continuous and Discrete Market Designs Process Messages. At the end of period 1 and period 2 of each trading game, each exchange j's order book ω_j is updated to reflect the processing of messages received by each exchange during that period, and ω_j is publicly observed for both Continuous and Discrete exchanges. Note that this means that the best bid and ask on a Discrete exchange is defined exactly the same way as on a Continuous exchange.

The only difference between Continuous and Discrete exchanges is how they process messages that are received in the same period. For an exchange that uses Continuous, all message sets sent to that exchange in the same period are serially processed by the exchange in a random sequence, with TF speed serving as a tie-breaker (as in Baldauf and Mollner, 2020). What this means is if multiple firms submit messages to an exchange in the same period of a trading game, the messages that are processed

first are those from fast TFs with ESST on that exchange; next are messages from fast TFs without ESST on that exchange; and last are messages from slow TFs.[18] Within each group, the processing order is uniformly random.

In contrast, Discrete exchanges first process all cancellations received in a period of the trading game, and then process any new limit or IOC orders received in that period, along with outstanding orders from previous periods, using a uniform-price auction, with price then discrete-time priority used to break any ties.[19]

Information policy is analogous between the Continuous and Discrete markets. In both cases, after any market participant actions, the exchange publicly announces (i) any trades that occurred (quantities and prices), and (ii) the updated state of the order book, reflecting any new orders and order cancellations.

3.1.3 Payoffs

At the conclusion of each trading game, market participants and exchanges earn the following per-period payoffs that depend on the value y (which may have changed during the trading game):

- If an investor buys a unit of x at price p on an exchange with trading fee f, then her payoff is $v + (y - p - f)$, where v is a large positive constant that represents her inelastic need to trade. If she needs to sell a unit and does so at p when the fundamental value is y, her payoff is $v + (p - y - f)$.[20]

- If an informed trader buys x at price p on an exchange with trading fee f, then his per-unit payoff is $y - p - f$; if he sells at price p his per-unit payoff is $p - y - f$.

- Similarly, if a trading firm buys (or sells) x at price p on an exchange that charges trading fee f, their per-unit payoff is $y - p - f$ (or $p - y - f$).

- Each exchange j earns trading fees of $2f_j$ per unit of x that is transacted on the exchange. Each exchange j also earns ESST fees of $N_j F_j$, where N_j is the number of TFs that have purchased ESST from the exchange in Stage 2.

[18] For simplicity, slow TFs cannot purchase ESST. In the equilibria that we characterize, they would not want to.

[19] More specifically, at the end of each time interval, the exchange aggregates all outstanding orders to buy and sell—both new orders submitted in that interval and orders that remain outstanding from previous intervals (i.e., neither executed nor canceled)—into demand and supply curves, respectively. If demand and supply cross, then trades are executed at the market-clearing price. If there is an interval of market-clearing prices, then we assume that the price chosen is the one closest to the prior midpoint. This tiebreaking will not occur in equilibrium of our game but will ensure that it is a weakly dominant strategy for TFs to bid their value in a sniping race. If it is necessary to ration quantity on either side of the market, priority is based first on price, then discrete time (i.e., orders that have been present in the book for strictly more intervals have higher priority if at the same price), with any remaining ties broken randomly.

[20] If an investor transacts strictly less than one unit, she receives v times her quantity traded; if an investor transacts strictly more than one unit, she receives v only for the first unit. In equilibrium, investors transact exactly one unit.

3.1.4 Additional Modeling Details

Batch Interval on Discrete. Our model of the trading game is only appropriate if the Discrete market's batch interval is very short. A very short batch interval makes it reasonable to assume that in each trading game, either 0 or 1 exogenous events occur. In practice, this amount of time is likely less than 1 millisecond (0.001 seconds).[21] In empirical evidence on races in Aquilina, Budish and O'Neill (2022), the modal latency arbitrage race lasts between 5-15 microseconds (0.000005-0.000015 seconds).

A very short batch interval also makes it reasonable for investors and informed traders to synchronize their orders across Continuous and Discrete exchanges, meaning that they can execute trades across multiple exchanges before other market participants can react. The model allows for this by assuming that an investor or informed trader can trade on all exchanges in period 2 before TFs see the updated state and respond in the following trading game.[22]

Key Regulations. Our trading game incorporates the key regulations described in the previous section as follows. First, we incorporate UTP by having the same security trade on all exchanges, and by having the value of the security be completely independent of the exchange on which it is bought or sold.[23] Second, we capture key aspects of Reg NMS by assuming that all market participants observe the current state of the order book on all exchanges at zero cost prior to taking any action (*frictionless search)*, and that the marginal cost of sending any message to any exchange is zero so that the only per-order cost of transacting on any exchange is the per-share trading fee (*frictionless access*).

3.2 Equilibrium Concept

For Stages 1, 2, and 3, our equilibrium solution concept is subgame perfect Nash equilibrium.

For Stage 4, we restrict market participants to use pure Markov strategies and condition their actions only on y and the publicly observable state of every exchanges' order book $\boldsymbol{\omega}$. We assume that in Period 1, market participants play what we refer to as an *order book equilibrium (OBE)* which we define below. In Period 2, we assume that market participants employ the following optimal (weakly

[21] Even for the highest activity symbol in all of US equity markets, SPY, on its highest-volume day of 2018 (February 6th), 95.2% of milliseconds have neither any trade nor change in the national best bid or offer (price or quantity). On an average day for SPY, 97.6% of milliseconds have neither a trade nor change in the national best bid or offer, and 99.4% of milliseconds have no trades. On an average day for GOOG, 99.6% of milliseconds have neither a trade nor change in the national best bid or offer, and >99.9% of milliseconds have no trade. These averages are computed based on a sample of 12 randomly selected trading days in 2018. Unless arrivals of exogenous events are highly correlated, these figures suggest that multiple events occurring within the same millisecond is very rare.

[22] Our impression, both from discussions with industry practitioners and our understanding of the relevant engineering details, is that while the ability to synchronize orders in this manner was pretty variable in the early days of Reg NMS, it is now widespread and commodified. Difficulty with such synchronization was at the heart of the narrative in Michael Lewis's book *Flash Boys*, and is modeled carefully in Baldauf and Mollner (2020).

[23] Our model is not designed to study the interesting and important role of the opening and closing auctions, which are proprietary to the exchange on which the stock is listed, and which are not subject to the market design criticism in Budish, Cramton and Shim (2015). Rather, our model is of regular-hours stock exchange trading (about 90% of exchange volume), for which UTP makes the listing exchange irrelevant.

dominant) strategies:

- Investors: upon arrival, an investor sends IOCs (at the prevailing bids or offers) to trade up to one unit in their desired direction, prioritizing their demand across exchanges based on the net price including trading fee; if there are any remaining orders that are profitable to trade against based on the publicly observed state, the investor trades against those as well.[24]

- Informed traders: upon a privately observed jump in y, an informed trader sends IOCs (at the prevailing bids or offers) to trade against any orders that are profitable to trade against based on their privately observed y and trading fees.[25]

- Trading firms: when there is a publicly observed jump in y, there are two cases to consider. First, if y jumps to a value at which it is not profitable for any TF to trade given the state of any exchange's order book and its trading fees, then TFs submit no messages.[26] Second, if y jumps to a value at which it is profitable to trade given the state of some exchange j's order book and its trading fees, then any TFs that are providing liquidity at unprofitable prices send cancellation messages to try to cancel their unprofitable ("stale") quotes, while at the same time all TFs send IOCs to try to trade against ("snipe") stale quotes offered by others.[27]

This behavior by TFs in period 2 when y jumps to a value at which it is profitable to trade is described as a "sniping race" in Budish, Cramton and Shim (2015). If there are K fast TFs with ESST on a Continuous exchange attempting to snipe a stale quote, each "wins" the race with probability $\frac{1}{K}$ (because a Continuous exchange processes messages serially in a uniformly random sequence). Hence, if a TF providing liquidity is fast and also has ESST on an exchange, then it is sniped with probability $\frac{K-1}{K}$; if a TF providing liquidity is slow (or does not have ESST and another TF does), then it is sniped with probability 1. On a Discrete exchange, on the other hand, any TF wishing to cancel its order can do so without being traded against. It is in this sense that Discrete eliminates latency arbitrage in our model.[28]

[24] When there is greater than one unit of liquidity offered across different exchanges at the best price (accounting for trading fees), investors use what we refer to as *routing table strategies* which dictate how they split their orders across exchanges (see Appendix E.3.2).

[25] Our assumption that informed traders act immediately if profitable to do so is in the spirit of Copeland and Galai (1983) and Glosten and Milgrom (1985); we abstract away from more sophisticated informed trading as in Kyle (1985).

[26] Any TF that wishes to cancel a limit order on any exchange's order book is indifferent between canceling that order immediately and waiting until period 1 of the following trading game to do so.

[27] When attempting to trade against stale quotes on exchange j, TFs use IOCs with a price of $y - f_j$ for attempts to buy and $y + f_j$ for attempts to sell. Given our tiebreaking rule for Discrete exchanges (see fn. 19) there is no advantage to reducing one's bid or increasing one's ask when attempting to trade against stale quotes on either Continuous or Discrete exchanges.

[28] On a Discrete exchange without trading fees, there are also Nash equilibria for Period 2 in which trading firms who are providing stale quotes do not cancel their orders upon the arrival of public information; in any such equilibria, if trade occurs on the Discrete exchange, it must be at the price $p = y$ as TFs competing to snipe stale quotes drive the price of the security to its fundamental value. Hence, if there were a TF with a stale quote that did not try to cancel, price competition in the auction would protect it from latency arbitrage as well. For expositional simplicity, we assume that liquidity providers on both Discrete and Continuous attempt to cancel any unprofitable quotes when there is a publicly observed jump in y in Period 2 (as doing so does not affect any of our economic conclusions).

Period 1 Equilibrium Play and Order Book Equilibrium. In Period 1, TFs simultaneously submit message sets $\boldsymbol{\mu}^* \equiv \{\boldsymbol{\mu}_i^*\}$ given the current state $(y, \boldsymbol{\omega})$. A natural solution concept would be pure strategy Markov perfect equilibrium (MPE) (or equivalently pure strategy Nash equilibrium for a single play of our trading game). However, because of adverse selection, a pure-strategy MPE (or Nash equilibrium for a single play of our game) does not exist. The key intuition is that if there is some TF providing a single unit of liquidity at a bid-ask spread that equates the benefits of liquidity provision (realized when an investor arrives seeking to trade one unit) to its costs (from being traded against by an informed trader or being sniped), then on the one hand, other TFs do not have incentive to offer additional liquidity at this spread (because they would suffer adverse selection or latency arbitrage costs without adequate compensation), but on the other hand, this leaves the TF who is providing liquidity incentive to deviate by widening its spread.[29]

To handle this non-existence issue, we introduce and employ an alternative equilibrium solution concept, *order book equilibrium*. OBE strictly weakens MPE by allowing for profitable unilateral deviations to exist, as long as they are rendered unprofitable by one of two specific reactions by rivals: *withdrawals* of liquidity, or *safe profitable price improvements*. Withdrawals are message sets that strictly reduce the amount of liquidity provided relative to a particular candidate equilibrium message set $\boldsymbol{\mu}^*$; that is, withdrawals only add cancellation messages to or eliminate limit orders from $\boldsymbol{\mu}^*$. Price improvements are message sets that, relative to $\boldsymbol{\mu}^*$, do not increase the cost of trading any quantity $q \in (0, 1]$ on any exchange $j \in \mathcal{M}$, but make it strictly cheaper to trade some quantity $q \in (0, 1]$ on some exchange j. For example, a price improvement may involve messages that offer liquidity at a narrower bid-ask spread on exchange j than what would be provided given $\boldsymbol{\omega}$ and $\boldsymbol{\mu}^*$. A safe profitable price improvement is a price improvement that is strictly profitable for a TF to engage in given the message sets submitted by other TFs and the state $\boldsymbol{\omega}$, and remains strictly profitable even if some other TF withdraws liquidity in response.

Definition 3.1. An *order book equilibrium* (abbreviated OBE) of our trading game is a set of message sets $\boldsymbol{\mu}^* \equiv \{\boldsymbol{\mu}_i^*\}$ submitted by all TFs in Period 1 given state $(y, \boldsymbol{\omega})$ that satisfies the following two conditions:

1. No TF i has a safe profitable price improvement.

2. No TF i has any other strictly profitable deviation (i.e., not a price improvement) that remains strictly profitable if, in response to TF i's deviation, some other TF engages in a profitable reaction that is either: (a) a withdrawal of liquidity; or (b) a safe profitable price improvement.

[29] For the standard model of undifferentiated Bertrand competition without adverse selection, a pure-strategy equilibrium exists with marginal-cost pricing: "excess liquidity provision" by any firm willing to sell as much as the market demands at marginal cost is riskless and constrains the price that other firms can charge. In contrast, in our environment the expected cost of providing liquidity depends on the mix of trading counterparties, which in turn depends on the liquidity provided by rivals. Hence, TFs are not willing to provide excess liquidity in the order book to constrain others' spreads, as they would be exposed to adverse selection and sniping risk without the full benefit of being filled by uninformed investors. (Equilibria in mixed strategies can exist when participants are able to provide liquidity at random prices (Baruch and Glosten, 2019).)

To understand the role that Condition 1 plays by ruling out safe profitable price improvements, consider the following example. Suppose the value of the security y is 10, and that in equilibrium a liquidity provider offers a single unit of liquidity at a bid-ask spread of 2 (i.e., limit orders to buy at 9 and sell at 11) when there is no other liquidity being offered. Suppose that it is strictly profitable for another TF to engage in a price improvement, and offer an additional unit of liquidity at a narrower bid-ask spread—say by offering to buy at 9.1 and sell at 10.9, equivalent to a bid-ask spread of 1.8. By imposing the safe requirement, Condition 1 of OBE requires that for such a price improvement to challenge equilibrium existence, it must remain strictly profitable due to the act of liquidity provision alone, and not from also continuing to snipe any liquidity that is no longer profitable to offer. In the example, it means that offering a unit of liquidity at a spread of 1.8 is strictly profitable even if the original unit of liquidity were withdrawn. We believe that the absence of safe profitable price improvements (and not just strictly profitable price improvements) is a necessary condition for an exchange's order book to be at a rest point, whereby no TF wishes to modify or adjust its outstanding orders given the anticipation of likely reactions by rivals, and captures the spirit of competitive liquidity provision as discussed and assumed in Glosten and Milgrom (1985).[30]

Condition 2 of OBE imposes the additional requirement that there are no other strictly profitable deviations (i.e., not price improvements) that remain strictly profitable even if another TF profitably reacted with either (a) a withdrawal, or (b) a safe profitable price improvement. By allowing for safe profitable price improvements as reactions to other strictly profitable deviations (e.g., deviations that worsen liquidity), OBE requires that for all other deviations to challenge equilibrium existence, they must not incentivize the provision of new liquidity at more competitive prices. In our example, a deviation by the liquidity provider to widen its spread—say to 2.2, with an offer to buy at 8.9 and sell at 11.1—would not challenge OBE if the deviation would be rendered unprofitable by another TF engaging in a safe profitable price improvement (e.g., offering a unit of liquidity at a spread of 2.1). We believe this additional requirement is also a necessary condition for an exchange's order book to be at a rest point, and further captures the idea of competitive liquidity provision whereby equilibrium spreads are disciplined even without excess liquidity being present in exchanges' order books.

Our concept is related to, and borrows inspiration from, alternative solution concepts used by Wilson (1977) and Riley (1979) to study insurance markets. In these alternative concepts, deviations must remain profitable to the withdrawal (Wilson) or addition (Riley) of certain insurance policies to rule out equilibria. Our relation to the insurance literature is not accidental: both settings feature adverse selection, and firms that are "undercut" by a rival may wish to withdraw from the market rather than face an adversely selected set of trading partners.

In Appendix E.2, we provide additional details and an example that illustrates why OBE helps to ensure equilibrium existence.

[30] The concept also captures the spirit of "immediate responses" to deviations as assumed in the continuous-time model of Budish, Cramton and Shim (2015).

3.3 Overview of Equilibrium Analysis

In Section 3.4, we first analyze the subgame where all exchanges have chosen Continuous in Stage 1. This subgame represents "the status quo." We prove that there exist equilibria where exchanges maintain positive market shares, trading fees are competitive (i.e., zero), and, importantly, exchanges *capture and maintain economic rents* obtained through supra-competitive fees for ESST. We discuss how our model's predictions fit many of the empirical patterns documented in the previous section, providing support for the use of our model to predict equilibrium outcomes for novel market designs.

In Section 3.5, we then analyze the subgame where only one exchange has chosen Discrete in Stage 1, while the others have all chosen Continuous. We show that in *any* equilibria, all trading activity occurs on the sole Discrete exchange, and the Discrete exchange earns positive profits. In essence, the Discrete exchange is compensated for eliminating the tax that latency arbitrage imposes on trading.

In Section 3.6, we consider subgames where there are multiple Discrete exchanges. Here, we prove that in *any* equilibria, again trading activity occurs only on Discrete exchanges, but now all exchanges earn zero profits.

The results from Sections 3.4-3.6 are used to establish our main results in Section 3.7. We show that when exchanges choose market designs in Stage 1, the payoffs from their choices comprise a Prisoner's Dilemma: each exchange earns positive profits if all exchanges choose Continuous, but any single exchange has a unilateral incentive to deviate (if the game is played once) to choose Discrete. However, if more than one exchange chooses Discrete, all exchanges earn zero profits and are worse off than under the status quo. We then use this result to examine the market design adoption incentives facing exchanges when Stages 1–4 are played repeatedly, and derive a necessary and sufficient condition for the status quo—in which exchanges always choose Continuous—to persist as a possible equilibrium.

3.4 Equilibrium Analysis: All Exchanges Continuous ("The Status Quo")

In the subgame following Stage 1 where all exchanges have chosen Continuous, we will show that there exist equilibria with the following properties. First, all exchanges charge zero trading fees (i.e., trading fees are competitive). Second, ESST fees are strictly positive and fast TFs purchase ESST from all exchanges with positive market shares. Even so, ESST fees are bounded above, and exchanges cannot fully extract all latency arbitrage rents from fast TFs. Last, in Period 1 of each trading game, a single unit of liquidity is provided at an equilibrium spread denoted $s^*_{continuous}$ across multiple exchanges, according to an arbitrary vector of market shares denoted $\boldsymbol{\sigma}^*$; and in period 2 of each trading game, investors route their orders across exchanges according to $\boldsymbol{\sigma}^*$. That is, the exchange market share vector coordinates the liquidity provision actions of TFs and the routing decisions of investors.

The equilibrium spread $s^*_{continuous}$ is given by the solution to:

$$\lambda_{invest} \cdot \frac{s^*_{continuous}}{2} = (\lambda_{public} + \lambda_{private}) \cdot L(s^*_{continuous}), \tag{3.1}$$

where $L(s) \equiv \Pr(J > \frac{s}{2}) \cdot E(J - \frac{s}{2} | J > \frac{s}{2})$ is the expected loss to a liquidity provider upon a jump in y, if traded against due to either sniping or adverse selection. At this spread a sole liquidity provider would be indifferent between offering a unit of liquidity on an exchange with zero trading fees, and choosing to snipe a rival offering liquidity at the same spread.[31] To see this, note that the left-hand side represents per-trading game expected benefits earned from liquidity provision on an exchange with zero trading fees; such benefits arise whenever an investor arrives with probability λ_{invest} and trades, paying half the spread $s^*_{continuous}$. The right-hand side represents the expected costs of liquidity provision, which arise from the following three sources. First, there is traditional adverse selection whenever an informed trader arrives with private information. Per trading game, this cost is $\lambda_{private} \cdot L(s^*_{continuous})$. Second, there are latency arbitrage costs whenever there is a publicly observed jump in y and a resulting sniping race. Per trading game, this cost is $\lambda_{public} \cdot \frac{N-1}{N} \cdot L(s^*_{continuous})$, where the $\frac{N-1}{N}$ term reflects the probability that a fast TF loses the sniping race (assuming that all N TFs on the exchange are equally fast), and the $L(s^*_{continuous})$ term is the same because we have assumed for convenience that private and public information jumps have the same distribution. Third, there is the opportunity cost of not sniping a rival liquidity provider, equal to $\lambda_{public} \cdot \frac{1}{N} \cdot L(s^*_{continuous})$.[32]

Proposition 3.1. *Consider the subgame following Stage 1 in which all exchanges have chosen Continuous. For any vector of market shares* $\boldsymbol{\sigma^*} = (\sigma^*_1, \ldots, \sigma^*_M : \sum_j \sigma^*_j = 1)$*, and for any vector of ESST fees* $\boldsymbol{F}^* = (F^*_1, \ldots, F^*_M)$ *that satisfies the condition given by (3.2) below, there exists an equilibrium of this subgame where:*

(Stage 2): Each exchange j charges F^*_j for ESST, and charges zero trading fees ($f^*_j = 0$);

(Stage 3): All N fast trading firms purchase ESST from every exchange j where $\sigma^*_j > 0$;

(Stage 4): The following occurs in every iteration of the trading game given state $(y, \boldsymbol{\omega})$. At the end of period 1, σ^*_j quantity of liquidity is provided on each exchange j at spread $s^*_{continuous}$ (defined in (3.1)) around y. In period 2: an investor, upon arrival, immediately transacts σ^*_j at the best bid or offer on each exchange j; an informed trader, upon arrival, immediately transacts σ^*_j at the best bid or offer on each exchange j if their privately-observed jump in y exceeds $\frac{s^*_{continuous}}{2}$; and if there is a publicly-observed jump that exceeds $\frac{s^*_{continuous}}{2}$, a sniping race occurs on all exchanges, in which all fast trading firms attempt to trade against all stale quotes provided by trading firms other than themselves, and all fast trading firms providing any liquidity on any exchange attempt to cancel their stale quotes.

[31] Equation (3.1) has a unique solution since the left-hand side is strictly increasing, the right-hand side is strictly decreasing in $s^*_{continuous}$, and the left-hand side is less than the right-hand side when the spread is 0.

[32] This same bid-ask spread $s^*_{continuous}$ also leaves a slow TF with zero profits from liquidity provision—i.e., a slow TF is indifferent between providing liquidity and doing nothing. A slow TF who provides liquidity at (3.1) gets sniped with probability 1 in the event of a public jump as opposed to probability $\frac{N-1}{N}$ for a fast TF, but the slow TF does not need to be compensated in equilibrium for the opportunity cost of not sniping. Evidence in Aquilina, Budish and O'Neill (2022) suggests that both fast and slow TFs providing liquidity that sometimes gets sniped are empirically relevant. In equilibrium as described below, there can be a mix of fast and slow TFs providing liquidity at $s^*_{continuous}$, and there can be multiple TFs each providing a fraction of the aggregate liquidity—e.g., one TF provides 0.6 at $s^*_{continuous}$ while a second provides the remaining 0.4. A fast TF who provides a fraction of the aggregate liquidity earns liquidity provision profits on whatever they provide and sniping profits on whatever others provide.

The condition on ESST fees is:

$$\frac{\Pi^*_{continuous}}{N} - \sum_{j:\sigma^*_j>0} F^*_j \geq \max(0, \pi_N^{lone\text{-}wolf} - \min_j F^*_j), \tag{3.2}$$

where $\Pi^*_{continuous} \equiv \lambda_{public} \cdot L(s^*_{continuous})$ *denotes the total* "sniping prize" *(i.e., the expected amount of latency arbitrage rents), and* $\pi_N^{lone\text{-}wolf}$ *is a constant discussed below and defined in Appendix E.3, equation (E.3).*

(All proofs are contained in the Appendix.)

To prove the Proposition, we first prove that in any Stage 4 subgame where (i) all exchanges have chosen Continuous, (ii) all N fast trading firms purchase ESST from the same set of exchanges, and (iii) all exchanges set zero trading fees, any order book equilibrium has exactly one unit of liquidity at spread $s^*_{continuous}$ provided across exchanges according to some vector of market shares $\boldsymbol{\sigma^*} = (\sigma^*_1, \ldots, \sigma^*_M)$, and such an equilibrium exists (Lemma E.1). In such an equilibrium, investors upon arrival in period 2 route their demand across exchanges according to this same vector of market shares $\boldsymbol{\sigma^*}$. Economically, this means that the marginal unit of liquidity provided is equally profitable across all exchanges, because each exchange's share of liquidity provided ("depth") matches its share of trading volume from investors.

We next examine behavior in Stage 3, and prove that if each exchange j charges F^*_j for ESST fees and zero for trading fees, there is a subgame equilibrium for all fast TFs to purchase ESST from all exchanges in $\mathcal{M}^* \equiv \{j : \sigma^*_j > 0\}$ as long as condition (3.2) is satisfied. This condition imposes an upper-bound on ESST fees, and is key to our finding that exchanges do not extract all sniping rents. To derive this condition, we analyze a specific deviation for fast TFs (which we refer to as a *lone-wolf deviation*), and show that because it is the most attractive deviation for TFs to consider, ruling it out is sufficient for establishing equilibrium existence.[33] We prove that condition (3.2) ensures this lone-wolf deviation is not profitable, as each fast TF earns more in expectation by purchasing ESST from all exchanges in $\mathcal{M}^*$ and earning $\frac{\Pi^*_{continuous}}{N}$ per trading game (gross of ESST fees) than purchasing ESST from just a single exchange and earning deviation profits of $\pi_N^{lone-wolf}$ per trading game.

Last, we turn to behavior in Stage 2. Given equilibrium strategies that we construct in Stages 3 and 4, no exchange j has an incentive to adjust ESST fees from F^*_j: as all TFs are already purchasing ESST from j, lowering F^*_j does not affect the amount of trading volume that j receives in Stage 4 and strictly reduces profits; and raising F^*_j induces TFs to no longer purchase from, or provide liquidity on, exchange j. Last, trading fees are zero because any exchange that raises its trading fee from zero

[33] In a lone-wolf deviation, instead of purchasing ESST from all exchanges in $\mathcal{M}^*$, a fast TF purchases ESST from just a single exchange. The lone-wolf then becomes the sole liquidity provider on this single exchange at a spread $\tilde{s}_N < s^*_{continuous}$ that we characterize analytically, and by doing so attracts all trading volume to this single exchange in equilibrium (Lemma E.2). In this equilibrium, the spread charged by the lone-wolf needs to be sufficiently narrower than $s^*_{continuous}$ so that other fast TFs prefer to snipe the lone-wolf rather than to undercut the lone-wolf's spread on a different exchange where there is one less TF (the lone-wolf) who is able to competitively snipe.

receives zero share in all subsequent trading games.[34]

3.4.1 Features of the Status Quo

The equilibria described in Proposition 3.1 have the following features, which are consistent with the empirical facts presented in Section 2 and in Appendix F.

Virtual Single Platform. Due to frictionless search and access, market participants can "stitch" together multiple exchanges into what we refer to as a *virtual single platform*. By this, we mean the following. First, in every trading game, all exchanges with positive depth have the same bid-ask spread $s^*_{continuous}$, resulting in a common market-wide best bid and offer. Second, each exchange's share of market depth at this spread is equal to its equilibrium share of market volume. Last, multiple exchanges are able to maintain positive market shares without the market tipping to any one exchange (consistent with exchange market shares shown in Section 2.3).[35]

The intuition is that as long as depth and volume are equivalent across all exchanges, the equilibrium bid-ask spread (3.1) applies equally to all liquidity on all exchanges. As long as the depth to volume ratio is the same across all exchanges, the marginal unit of liquidity is equally well off across all exchanges. If some exchange has too much depth relative to its volume, liquidity providers will suffer too much adverse selection and sniping relative to the benefits of liquidity provision. If some exchange has too little depth relative to its volume, the reverse is true.[36] In Appendix F we show that exchanges with positive depth have the same bid-ask spread, and the depth-volume relationship obtains robustly in the data.

Competitive Trading Fees. Trading fees are competitive and zero on all exchanges. Any exchange j, given that all other exchanges set zero trading fees, cannot charge a positive trading fee and attract positive trading volume due to frictionless search. This is true even if investors broke ties in j's favor (all else equal), and even if j charged lower ESST fees than other exchanges.[37]

[34] In the Appendix, we show that an exchange's losses from negative trading fees can be arbitrarily large without TFs engaging in any self-dealing. For this reason, we assume that exchanges cannot charge negative trading fees.

[35] Our model does not yield much insight into the determination of equilibrium exchange market shares. That said, it does provide some insight into why they might be interior and relatively stable over time. In the equilibria described in Proposition 3.1, investors break ties when indifferent across exchanges using what we refer to as *routing table strategies* (see Appendix E.3.2). Such strategies, in turn, coordinate where TFs provide liquidity. Thus, if investor routing tables are relatively stable over time, then exchange market shares will be as well.

[36] These results are closely related to Glosten (1994) and Ellison and Fudenberg (2003). Glosten (1994) models multiple limit order book exchanges under the assumption that "an investor can costlessly and simultaneously send separate orders to each exchange" (pg. 1146), i.e., frictionless search and access. Ellison and Fudenberg (2003) study a model of platform competition for single-homing buyers and sellers that encompasses elements of the classic Pagano (1989) exchange competition model. Ellison and Fudenberg show there can exist a "plateau" of equilibria with interior market shares, where all platforms with positive market share in these equilibria have the same seller-buyer ratio.

[37] In a supporting Lemma for Proposition 3.1, we prove that in any equilibrium of a Stage 3 subgame where trading fees are zero for some exchanges and strictly positive elsewhere (and where all TFs purchase ESST from the same set of exchanges), no trading volume occurs on any exchange with positive trading fees (see Lemma E.1 in Appendix E.3).

ESST Fees and the Division of Latency Arbitrage Rents. In contrast to competitive pricing models where add-on rents are dissipated in competition to sell the pre-add-on good (cf. Ellison, 2005; Gabaix and Laibson, 2006), here exchanges do not compete away rents earned from the sale of ESST (an add-on service that is only valuable if an exchange has positive trading volume) by charging lower trading fees in competition for transaction volume. This is the case even though exchanges are assumed to be symmetric and undifferentiated, search is frictionless, and market participants can costlessly participate on any exchange. The reason is that trading fees are zero across all exchanges. Any dissipation of ESST rents via trading fees in order to attract trading volume would require such fees to be negative, which in turn would create an incentive for market participants to execute an unlimited number of trades and make unlimited profits.[38]

Moreover, even though exchanges are able to "post prices" and make take-it-or-leave-it offers to TFs, they cannot capture *all* latency arbitrage rents: fast TFs have bargaining leverage with exchanges because they can steer liquidity provision, and hence trading volume, to rival exchanges. This gives rise to the condition on ESST fees given by (3.2).[39] Using the analysis behind this bound, we are able to show that the proportion of sniping rents that TFs obtain is economically significant:

Proposition 3.2. *In the equilibria described by Proposition 3.1, exchanges' total rents from ESST fees,* $N \times \sum_{j:\sigma_j^*>0} F_j^*$*, are strictly less than* $\frac{M}{(M-1)(N-1)}\Pi^*_{continuous}$*.*

Proposition 3.2 implies that if $M \geq 3$ and $N \geq 6$, then exchanges in total are able to extract at most 30% of sniping rents, with the remainder accruing to fast trading firms.[40]

Sources of Deadweight Loss. In our model, there are N "fast" TFs (who can be thought of as exogenously endowed with "general-purpose" speed technology that makes them faster than "slow" TFs), and M exchanges exogenously present in the market and able to sell ESST to TFs. TFs' payments to the exchanges for ESST are transfers as opposed to deadweight loss.

We emphasize that, outside of the model, there is significant deadweight loss associated with the development of both general-purpose and exchange-specific speed technology. This includes investments in communications links between exchanges, proprietary speed-optimized hardware and software, and significant high-skilled human capital.

Moreover, standard excess entry and business stealing incentives (Mankiw and Whinston, 1986) may also be present in our environment. Specifically, if a potential entrant exchange has a way

[38] Although exchanges theoretically could dissipate rents via fixed payments to investors or broker-dealers for trading volume, our understanding is that this would not be legal.

[39] As with equilibrium exchange market shares, our model does not deliver a prediction for how ESST revenues are split across exchanges. In the equilibria described in Proposition 3.1, ESST fees are not required to be proportional to the volume traded at an exchange as long as condition (3.2) is satisfied.

[40] In our empirical setting there are 12 exchanges in total, of which 8 have significant market share and are owned by 3 exchange families (see Section 2.3). Aquilina, Budish and O'Neill (2022) found that the top 6 trading firms win over 80% of latency arbitrage races in the UK equities market in data from 2015; this number is consistent with our anecdotal understanding of the rough magnitude for N in U.S. equities. For example, the CEO of one of the largest high-frequency traders in the U.S. described in a conversation with two of the authors that there are 7 firms in the "lead lap" of the speed race in the U.S. equities market.

to obtain positive market share, then it has incentive to enter to capture ESST rents, even if it is completely undifferentiated from incumbent exchanges, including using the same market design.

3.5 Equilibrium Analysis: A Single Discrete Exchange

We next examine the subgame following Stage 1 where there is a single Discrete exchange.

First, suppose that trading fees on all exchanges are set to zero, and all TFs have purchased ESST from the same set of Continuous exchanges. A reasonable prior might be that there are multiple equilibrium outcomes for the Stage 4 trading game: for example, there might be an equilibrium where all liquidity is provided and taken from Continuous exchanges, and another where all liquidity is provided and taken from the Discrete exchange. However, this is not the case:

Proposition 3.3. *Consider the repeated Stage 4 trading game with a single Discrete exchange, assuming that in Stage 2 all exchanges set trading fees to zero and in Stage 3 all fast trading firms have purchased ESST from the same set of Continuous exchanges. Any equilibrium has the following properties. In period 1 of each trading game: exactly one unit of liquidity is provided on Discrete at bid-ask spread $s^*_{discrete}$, which solves:*

$$\lambda_{invest} \frac{s^*_{discrete}}{2} = \lambda_{private} \cdot L(s^*_{discrete}) \,, \tag{3.3}$$

around the value of y, and no liquidity is provided on any Continuous exchange. In period 2 of each trading game: an investor, upon arrival, immediately transacts one unit at the best bid or offer; an informed trader, upon arrival, immediately transacts one unit at the best bid or offer if their privately-observed jump in y exceeds $\frac{s^*_{discrete}}{2}$; if there is a publicly-observed jump in y that exceeds $\frac{s^*_{discrete}}{2}$, either all TFs with stale quotes cancel their stale quotes, or if the auction results in trade the auction price is the new value of y. Such an equilibrium of the trading game exists.

That is, liquidity cannot be offered on any Continuous exchange in any equilibrium. To understand why, note that if a trading firm was to provide liquidity on a Continuous exchange and not lose money, it would have to charge at least a "zero-variable profit spread" on a Continuous exchange, denoted $\bar{s}_{continuous}$, which we prove is strictly greater than $s^*_{discrete}$.[41] Since investor demand is perfectly elastic with respect to the bid-ask spread, if any liquidity provider on a Continuous exchange was weakly profitably offering liquidity at some spread $s \geq \bar{s}_{continuous}$, that provider could be strictly profitably undercut on Discrete at a strictly smaller spread $s' \in (s^*_{discrete}, s)$. Furthermore, any liquidity cannot be offered on Discrete at any spread other than $s^*_{discrete}$ in equilibrium: any greater, and it could be profitably undercut by another TF; any lower, and the liquidity provider would be losing money and be better off withdrawing.

[41] The difference between (3.3) and the equilibrium spread on Continuous exchanges, given by (3.1), is the $\lambda_{public} L(s^*)$ term missing from the equation defining $s^*_{discrete}$: this reflects that Discrete eliminates latency arbitrage rents, and hence the associated cost for liquidity providers. For this reason, $s^*_{discrete} < s^*_{continuous}$.

These same arguments also imply that no liquidity can be offered on any Continuous exchange in any Stage 4 trading game even if Discrete were to charge a strictly positive (but small enough) trading fee. We thus obtain the following result, which characterizes equilibria in any subgame following Stage 1 with a single Discrete exchange:

Proposition 3.4. *Consider the subgame following Stage 1 with a single Discrete exchange. Any equilibrium has the following properties: (i) in period 1 of each Stage 4 trading game, exactly one unit of liquidity is provided on Discrete and no liquidity is provided on any Continuous exchange; (ii) every Continuous exchange earns zero profits; and (iii) Discrete charges strictly positive trading fees and earns expected per-trading-game profits that exceed* $\frac{N-1}{N}\Pi^*_{continuous}$*. Such an equilibrium exists.*

In essence, when a single Discrete exchange competes against Continuous exchanges, the Discrete exchange is compensated for the elimination of the tax that latency arbitrage imposes on trading: as long as the Discrete exchange charges a trading fee that is less than this tax, by enough to account for the zero-variable profit deviation described above, it tips the market.[42]

3.6 Equilibrium Analysis: Multiple Discrete Exchanges

Last, consider the subgame following Stage 1 where more than one exchange chooses Discrete and the rest (if any) choose Continuous. When there are at least two Discrete exchanges and potentially one or more Continuous exchanges, the resulting equilibrium has similar features to the equilibria with only Continuous exchanges, described in Proposition 3.1:

Proposition 3.5. *Consider the subgame following Stage 1 where there are at least two Discrete exchanges. Any equilibrium has the following properties: (i) at least one Discrete exchange charges zero trading fees; (ii) in every iteration of the trading game, exactly one unit of liquidity is provided in aggregate across only Discrete exchanges with zero trading fees at bid-ask spread* $s^*_{discrete}$ *around the value of y following Period 1; (iii) no liquidity is provided on Discrete exchanges with positive trading fees or on Continuous exchanges; (iv) all exchanges earn zero profits. Such an equilibrium exists.*

Just as in the case of the status quo with only Continuous exchanges, multiple Discrete exchanges also operate as a virtual single platform: a single unit of liquidity is always provided in each trading game across only Discrete exchanges, the depth-volume relationship ensures that the marginal unit of liquidity is indifferent across these exchanges, and equilibria differ from one another only in exchange market shares. However, there are two key differences. First, the bid-ask spread is $s^*_{discrete}$, not $s^*_{continuous}$, which is better for investors and informed traders because $s^*_{discrete} < s^*_{continuous}$. Second, there are no longer latency arbitrage rents for exchanges or trading firms.

[42] Propositions 3.3-3.4 may at first seem in tension with Glosten (1994) (Proposition 9), which finds that the limit order book is in a sense "competition proof." The explanation for this apparent contradiction is that the Glosten (1994) model precludes latency arbitrage—traders arrive to market one-at-a-time, so it is not possible for there to be public information that multiple traders try to act on at the same time. The reason Discrete "wins" against Continuous in our model is precisely that it eliminates the latency arbitrage tax on liquidity.

3.7 Market Design Choice

We have now analyzed subgames following Stage 1 when there are multiple Continuous exchanges (Section 3.4), a single Discrete and one or more Continuous exchanges (Section 3.5), and multiple Discrete exchanges (Section 3.6). Under the equilibria that we have described, we have shown that for a single play of our overall game:

- If all exchanges are Continuous: each exchange j earns (per trading game) profits of NF_j^* (Proposition 3.1).
- If there is a single Discrete exchange and all other exchanges are Continuous: the Discrete exchange earns economic profits denoted $\Pi^*_{discrete}$, where $\Pi^*_{discrete} > \frac{N-1}{N}\Pi^*_{continuous}$, and the Continuous exchanges earn zero profits (Proposition 3.4).
- If there are multiple Discrete exchanges: all exchanges earn zero profits (Proposition 3.5).

Proposition 3.2 places an upper bound on exchange ESST revenues in {all Continuous}, while Proposition 3.4 places a lower bound on the Discrete exchange's profits in {a single Discrete, the remainder Continuous}. These bounds and some simple algebra (Lemma E.4 in the appendix) yields that $\Pi^*_{discrete} > NF_j^*$ for all exchanges j, for any equilibrium ESST revenues consistent with Proposition 3.2, and for $\Pi^*_{discrete}$ as characterized in Proposition 3.4. Discrete is thus a dominant strategy, but all exchanges prefer {all Continuous}, where they earn profits from speed technology, to {all Discrete} where they do not. We summarize these results in the following Proposition.

Proposition 3.6. *Assume that following Stage 1 market design choices, subgame equilibria are characterized by either Proposition 3.1 for {all Continuous}, Proposition 3.4 for {a single Discrete, the remainder Continuous}, or Proposition 3.5 for {multiple Discrete, the remainder Continuous}. Then anticipated exchange profits in Stage 1 as a function of their market designs constitute a Prisoner's Dilemma: Discrete is a dominant strategy, but all exchanges make greater profits in the subgame in which all exchanges are Continuous than in the subgame in which all exchanges are Discrete.*

In our analysis Discrete is a weakly dominant strategy, because an exchange's profits are zero if they are Continuous while there are one or more Discrete exchanges, and are also zero if they are one of many Discrete exchanges.

Last, with these results, we analyze infinitely-repeated play of our game: i.e., after Stages 1, 2, 3 and T iterations of our trading game are played, play returns to Stage 1.[43] The following Proposition states a necessary and sufficient condition for there to exist an equilibrium in which the status-quo equilibrium described in Proposition 3.1 (with all exchanges choosing Continuous) is repeatedly played.

[43] As noted above, we interpret trading games as each lasting a very short amount of time (e.g., one millisecond), whereas we interpret market design choices and fee adjustments as changing much less frequently (e.g., on the order of months). We thus assume that exchanges engage in discounting over the large number T of trading games played between opportunities to alter their market designs or fees, but, for notational simplicity, assume that there is no discounting between Stages 1, 2, and 3.

Proposition 3.7. *All exchanges repeatedly choosing Continuous in Stage 1 and playing a subgame equilibrium as described in Proposition 3.1 (in which each exchange j earns NF_j^* in ESST fees per trading game) in Stages 2–4 is an equilibrium of infinitely-repeated play of our game if and only if:*

$$\rho\Pi^*_{discrete} \leq NF_j^* \tag{3.4}$$

for all exchanges j, where $\rho \equiv (\sum_{t=0}^{T} \delta^t)/(\sum_{t=0}^{\infty} \delta^t)$ represents the share of net present value *represented by the initial T trading games out of an infinite series and $\delta < 1$ is the per-trading-game discount factor.*

This result follows directly from the Prisoner's Dilemma structure of Stage 1 established in Proposition 3.6.[44] Hence, as long as condition (3.4) holds and rents from the sale of ESST in perpetuity are larger than the short-term gains from eliminating latency arbitrage, it is possible for all exchanges to maintain the status-quo and repeatedly choose an inefficient market design.

4 Policy Implications

The question for policy is whether private-market forces will fix latency arbitrage and the arms race for speed (i.e., "will the market fix the market?"), or would a regulatory intervention be necessary, and if so, of what form. Section 4.1 synthesizes insights from the theoretical analysis for this policy question. Our theory suggests that although regulatory intervention may be necessary, this intervention need not take the form of a market design mandate: a regulatory "push" would be enough to induce private-sector market design innovation, which our analysis shows would attract share and help the market fix the market. Section 4.2 briefly discusses two potential forms such a push might take.

4.1 Insights from the Theory

Insight #1: Private innovation incentives may not be sufficient to induce market design innovation, even if social incentives for market design innovation are high. Proposition 3.7 shows that innovation might not occur—and the inefficient status quo could persist—even if the latency-arbitrage pie $\Pi^*_{continuous}$ is large. This can occur if other exchanges are able to imitate an innovator quickly (ρ is small) and/or speed-technology rents (represented by F_j^*) are sufficiently large.

We can capture formally that private and social incentives for innovation may diverge as follows. Add another parameter to the model, $DWL \leq \Pi^*_{continuous}$, that represents the portion of the latency-arbitrage prize that is dissipated as deadweight loss in the arms race for speed.[45] Sources of deadweight

[44] For necessity, if condition (3.4) were violated for some exchange j, then that exchange would have a profitable deviation to instead choose Discrete in Stage 1: doing so would earn it at least $\Pi^*_{discrete}$ in profits for at least T iterations of the trading game (Proposition 3.4), whereas in equilibrium it would earn NF_j^* in rents in perpetuity. For sufficiency, since all exchanges choosing Discrete in Stage 1 is an equilibrium for a single play of our overall game (Proposition 3.6), there exists a "grim-trigger" equilibrium whereby all exchanges choose Continuous in Stage 1 unless any exchange has previously deviated in which case exchanges always play Discrete.

[45] In the Budish, Cramton and Shim (2015) model, the entire latency-arbitrage pie is dissipated by investments in speed, i.e., $DWL = \Pi^*_{continuous}$. In richer models, inframarginal participants can earn economic rents—such as the

loss are discussed in Section 3.4, and include investments in speed technology, communications links, and specialized human capital. Social incentives for innovation are positive, but private incentives can be negative—i.e., "the market will not fix the market"—if deadweight loss from the arms race for speed is positive and the conditions of Proposition 3.7 obtain. Formally, if

$$\text{Social incentives are positive: } DWL > 0$$
$$\text{Private incentives are negative: } \rho\Pi^*_{discrete} < NF^*_j \text{ for all } j.$$

We can distinguish two ways in which private and social innovation incentives diverge. The first is that a private innovator only earns profits from their innovation temporarily, while society benefits from it permanently. This difference is captured by the parameter ρ, which affects private innovation incentives but not social innovation incentives—society enjoys the elimination of deadweight loss in perpetuity. This is the same wedge between private and social innovation incentives as in standard patent models (Nordhaus, 1969; Williams, 2017), though a difference here is that imitation may be especially rapid, meaning ρ is very small.

The second way in which incentives diverge is that exchanges, by innovating to fix the arms race for speed, lose the speed rents they currently enjoy. Formally, incumbent j would lose the net present value of their speed-technology revenues, $\frac{NF^*_j}{1-\delta}$. This is the sense in which the the industry rents from the speed race create a wedge between private and social innovation incentives. This wedge is conceptually novel relative to the extant innovation literature (cf. Bryan and Williams, 2021).

SEC Chair White's policy address on market design reform assumed that private and social innovation incentives align. In that case, the role of policy makers, as the SEC Chair described, is simply to ensure that they do not inadvertently "stand in the way" of "competitive solutions" to the problem.[46] However, if private and social incentives are misaligned, then there is a potential role for a policy intervention.

Insight #2: <u>If</u> an exchange adopts Discrete, it wins significant share in any equilibrium. Propositions 3.3-3.4 shows that, <u>if</u> an exchange adopts Discrete, it wins significant share against Continuous. The reason is the frictionless search environment under Reg NMS. Frictionless search ensures that, if there are two markets running in parallel, one with a tax and one without, the one without the tax wins in any equilibrium. This result is in contrast to many other models of platform competition, in which there exist equilibria where a new platform fails to get any share even if in principle it is better designed; the so-called "chicken-and-egg" problem (Farrell and Saloner, 1985; Katz and Shapiro, 1986; Caillaud and Jullien, 2003).

Our theory also suggests, however, that frictionless search is a double-edged sword for the innovator: it makes the innovator vulnerable to imitation, and, once imitated, fees are competed down to zero

incumbent exchanges in our analysis. For this reason, it is possible that $DWL < \Pi^*_{continuous}$.

[46] White, Mary Jo, "Enhancing Our Equity Market Structure," June 5, 2014.

(Proposition 3.5).

These results have an important implication for the form that a policy intervention might take: a "push" might be a viable alternative to a "mandate." By push, we mean any policy intervention that tips the balance of incentives sufficiently that an exchange will choose to innovate, despite their vulnerability to imitation and with that competitive pricing.

Insight #3: The incentives to adopt are highest for exchanges with low speed-technology rents (potentially including entrants). In our model, with M exchanges exogenously present in the market, the adoption incentives are largest for the exchange with the lowest speed-technology rents NF_j^*. In a richer model that considers entry, potential de novo entrants would not face any opportunity costs associated with losing speed technology rents, but they would face entry costs.

The policy implication is that the "push" implied by Insights #1-#2 could focus on either incumbents with low speed-technology rents or de novo entrants. These are the parties with the lowest opportunity cost of market design innovation.

Notably, the record of innovation attempts to date that relate to latency arbitrage lines up with this insight of our theory. The one case of an exchange proposal that addressed latency arbitrage in a theoretically comprehensive way came from an incumbent with very low market share and speed-technology revenues: the Chicago Stock Exchange's (CHX) proposal of an asymmetric speed bump in 2017. CHX's proposal generated significant opposition from larger incumbents. CHX was eventually acquired by the New York Stock Exchange Group, which officially withdrew the proposal in 2018.[47]

4.2 Potential "Pushes"

One potential push would be for the regulator to provide innovators with a modest exclusivity period — a ρ large enough to ensure that equation 3.4 does not hold. During this time, other exchanges would not be allowed by the regulator to imitate the design (either identically or with designs judged to be essentially similar). This policy could be modeled on a practice of the Food and Drug Administration, wherein it grants a period of market exclusivity for certain kinds of drugs that are not patentable.[48]

In Appendix G, we provide a back-of-envelope calculation that suggests an exclusivity period on the order of 1-2 years might be sufficient to induce entry. This exercise attempts to take into consideration some frictions left out of the main analysis, namely tick-size constraints, a maker-taker fee structure,

[47] Other innovative activity that relates to latency arbitrage has come from either entrants (most prominently IEX) or incumbents with low share (most prominently BYX). In all of these cases, the proposed designs addressed just a subset of latency arbitrage. While outside our model, this has the strategic flavor of the "puppy dog ploy" in Fudenberg and Tirole (1984) or the "judo economics" in Gelman and Salop (1983)—purposefully being small enough to avoid provoking a fierce competitive response. We discuss these proposals in detail in Appendix H.

[48] For legal reasons, patents do not seem a viable way to create market exclusivity in this context. First, the specific market design of frequent batch auctions is in the public domain. Second, even if frequent batch auctions were patented, to be effective the intellectual property protection would have to cover all possible market designs that eliminate latency arbitrage. As evidence of the difficulty of this, consider that the Chicago Mercantile Exchange filed for a patent for a market design that is in essence a form of batch auction without using the word "auction" (Hosman et al., "Mitigation of Latency Disparity in a Transaction Processing System," US Patent Application No. 14991654, 2017).

and agency frictions between investors and brokers trading on their behalf.

A second potential push would be simply to reduce entry costs or otherwise subsidize entry. Entry costs are meaningful in practice, and seem in large part to reflect legal costs related to gaining regulatory approval.[49] Moreover, an entrant proposing a new market design would face risk that its design is not approved. Policymakers could encourage entry by either reducing or subsidizing the cost of the regulatory approval process for useful new market designs, or by reducing risk of the regulatory approval process by proactively clarifying what kinds of exchange design innovations would be welcomed. Formally, let c denote risk-adjusted entry costs—i.e., the dollar cost of the entry process divided by the perceived probability of regulatory approval. If these costs can be lowered to the point where $c < \rho\Pi^*_{discrete}$, then a de novo entrant has incentive to adopt. Propositions 3.3 and 3.4 then tell us the better market design would take off.

4.3 Can Investors Fix the Market?

Another possibility is that the parties harmed by the current market design—investors, who ultimately bear the cost of latency arbitrage—can find a market-based solution to the problem without a need for policy (Kilenthong and Townsend, 2021). For example, large institutional investors could fund a frequent batch auction exchange. Speculatively, it seems to us that the most likely explanation for why this has not happened is the nature of the magnitudes involved and the concentrated-dispersed dynamics of the problem. Aquilina, Budish and O'Neill (2022) find that latency arbitrage imposes a roughly one-half of one basis point tax on investors, i.e., roughly 0.005%. This certainly seems "small" to a typical investor. Yet, it adds up to about $5 billion per year in equities markets alone. At a 5% discount rate, this has a net present value of $100 billion. So, latency arbitrage is very important to the concentrated parties that enjoy a share of the pie—high-frequency trading firms and exchanges—but at the same time imposes only a modest tax on the widely-dispersed set of end investors. As Olson (1965) emphasizes, it is precisely the role of policy to act on behalf of dispersed interests (while resisting being co-opted by concentrated interests).

5 Conclusion

Our paper has put forth a theoretical model of stock exchange competition that clarifies why, even if allowed, exchanges may not wish to innovate: they profit from the speed race generated by the existing market design. Our story is not about new market designs failing to gain traction if introduced, but rather one of incumbents protecting rents. The modest policy proposals put forth in the last section are designed with this perspective in mind. Rather than mandate a particular market design, these

[49] The Investors' Exchange (IEX) is estimated to have raised over $100M of venture capital in advance of its approval as a stock exchange in June 2016 (see the IEX Group organization page on Crunchbase.com). The Chicago Stock Exchange was acquired for $70M and many industry observers speculated that its main asset was its exchange license, i.e., that it had paid the entry costs necessary to exist as a formal exchange (Michaels and Osipovich, "NYSE in Talks to Buy Chicago Stock Exchange," *Wall Street Journal,* March 30, 2018).

proposals attempt to alter the incentives for private innovation to better align private incentives with social interests, to encourage "the market to fix the market."

A standalone contribution of this paper, separable from our motivating question about market design innovation, is the development of an industrial organization (IO) model of the modern stock exchange industry. This model may prove to be a useful starting point for other research on financial exchange competition, and perhaps platform competition more generally. We also hope that future research can take inspiration from the style of analysis in this paper, with a mix of theory and empirical work guided by institutional and regulatory details.

The ideas in this paper are already having some modest policy impact. In October 2019, the SEC issued a statement inviting market design proposals for the thinly-traded segment of the U.S. stock market. In this proposal, the SEC explicitly points to batch auctions as a potential market design alternative it encourages, and signals willingness to suspend Unlisted Trading Privileges for stocks listed on exchanges that so innovate (thereby creating a form of exclusivity for the innovator).[50] In February 2020, the SEC issued a proposed reform to the market for exchange data. The proposed rule cited our theoretical finding that each exchange has market power in the sale of proprietary market data and related speed technology, as well as our empirical finding that exchanges earn significant revenue from selling these products.[51] In a policy address on the topic of market data and exchange governance at around that time, Commissioner Robert J. Jackson Jr. cited our work and said "Without changing [the] incentives, we cannot and should not expect the market to fix the market."[52]

References

Agarwal, Nikhil, and Eric Budish. 2021. "Market Design." *Handbook of Industrial Organization*, 5(1): 1–79.

Aquilina, Matteo, Eric Budish, and Peter O'Neill. 2022. "Quantifying the High-Frequency Trading 'Arms Race'." *The Quarterly Journal of Economics*, 137(1): 493–564.

Arrow, Kenneth. 1962. "Economic Welfare and the Allocation of Resources to Invention." In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, ed. 609-626. Princeton University Press.

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

Baruch, Shmuel, and Lawrence R. Glosten. 2019. "Tail expectation and imperfect competition in limit order book markets." *Journal of Economic Theory*, 183: 661–697.

Bryan, Kevin A., and Heidi L. Williams. 2021. "Innovation: Market Failures and Public Policies." *Handbook of Industrial Organization*, 5(1): 281–388. NBER Working Paper No. 29173.

Budish, Eric, Peter Cramton, and John Shim. 2015. "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response." *The Quarterly Journal of Economics*, 130(4): 1547–1621.

Caillaud, Bernard, and Bruno Jullien. 2003. "Chicken & Egg: Competition Among Intermediation Service Providers." *The RAND Journal of Economics*, 34(2): 309–328.

[50] U.S. Securities and Exchange Commission, "Commission Statement on Market Structure Innovation for Thinly Traded Securities," Release No. 34-87327 (2019).

[51] U.S. Securities and Exchange Commission, "Market Data Infrastructure," Release No. 34-88216 (2020), pp. 365-366.

[52] Jackson, Robert J., Jr., "Statement on Reforming Stock Exchange Governance," January 8, 2020.

Chao, Yong, Chen Yao, and Mao Ye. 2019. "Why Discrete Price Fragments U.S. Stock Exchanges and Disperses Their Fee Structures." *The Review of Financial Studies*, 32(3): 1068–1101.

Copeland, Thomas E., and Dan Galai. 1983. "Information Effects on the Bid-Ask Spread." *The Journal of Finance*, 38(5): 1457–1469.

Ellison, Glenn. 2005. "A Model of Add-On Pricing." *The Quarterly Journal of Economics*, 120(2): 585–637.

Ellison, Glenn, and Drew Fudenberg. 2003. "Knife-Edge or Plateau: When Do Market Models Tip?" *The Quarterly Journal of Economics*, 118(4): 1249–1278.

Farrell, Joseph, and Garth Saloner. 1985. "Standardization, Compatibility, and Innovation." *The RAND Journal of Economics*, 16(1): 70–83.

Fudenberg, Drew, and Jean Tirole. 1984. "The Fat-Cat Effect, the Puppy-Dog Ploy, and the Lean and Hungry Look." *American Economic Review: Papers and Proceedings*, 74(2): 361–366.

Gabaix, Xavier, and David Laibson. 2006. "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets." *The Quarterly Journal of Economics*, 121(2): 505–540.

Gelman, Judith R., and Steven C. Salop. 1983. "Judo Economics: Capacity Limitation and Coupon Competition." *The Bell Journal of Economics*, 14(2): 315–325.

Glosten, Lawrence R. 1994. "Is the Electronic Open Limit Order Book Inevitable?" *The Journal of Finance*, 49(4): 1127–1161.

Glosten, Lawrence R., and Paul R. Milgrom. 1985. "Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders." *Journal of Financial Economics*, 14(1): 71–100.

Griliches, Zvi. 1957. "Hybrid Corn: An Exploration in the Economics of Technological Change." *Econometrica*, 25(4): 501–522.

Harris, Larry. 2002. *Trading and Exchanges: Market Microstructure for Practitioners.* Oxford University Press.

Hirshleifer, Jack. 1971. "The Private and Social Value of Information and the Reward to Inventive Activity." *The American Economic Review*, 61(4): 561–574.

Indriawan, Ivan, Roberto Pascual, and Andriy Shkilko. 2022. "On the Effects of Continuous Trading." Working Paper.

Katz, Michael L., and Carl Shapiro. 1986. "Technology Adoption in the Presence of Network Externalities." *Journal of Political Economy*, 94(4): 822–841.

Kilenthong, Weerachart T., and Robert M. Townsend. 2021. "A Market-Based Solution for Fire Sales and Other Pecuniary Externalities." *Journal of Political Economy*, 129(4): 981–1010.

Kyle, Albert S. 1985. "Continuous Auctions and Insider Trading." *Econometrica*, 53(6): 1315–1335.

Mankiw, N. Gregory, and Michael D. Whinston. 1986. "Free Entry and Social Inefficiency." *The RAND Journal of Economics*, 17(1): 48–58.

Milgrom, Paul. 2021. "Auction Research Evolving: Theorems and Market Designs." *American Economic Review*, 111(5): 1383–1405.

Nordhaus, William D. 1969. *Invention, Growth, and Welfare: A Theoretical Treatment of Technological Change.* The MIT Press.

Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups.* Harvard University Press.

Pagano, Marco. 1989. "Trading Volume and Asset Liquidity." *The Quarterly Journal of Economics*, 104(2): 255–274.

Riley, John G. 1979. "Informational Equilibrium." *Econometrica*, 47(2): 331–359.

Roth, Alvin E. 2018. "Marketplaces, Markets, and Market Design." *American Economic Review*, 108(7): 1609–1658.

Williams, Heidi L. 2017. "How Do Patents Affect Research Investments?" *Annual Review of Economics*, 9(1): 441–469.

Wilson, Charles. 1977. "A Model of Insurance Markets with Incomplete Information." *Journal of Economic Theory*, 16(2): 167–207.