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

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

July 2021

Abstract

This paper models stock exchange competition and innovation in the modern electronic era. In the status quo, frictionless search and multi-homing cause seemingly fragmented exchanges to behave as a “virtual single platform.” As a result, exchange trading fees are competitive. But, exchanges earn economic profits from selling speed technology. We document stylized facts consistent with these results. We then analyze incentives for market design innovation. The novel tension between private and social innovation incentives is incumbents’ rents from speed technology. These rents create a disincentive to adopt market designs that eliminate latency arbitrage and the high-frequency trading arms race.

Keywords: market design, innovation, financial exchanges, industrial organization, platform markets, 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 Jason Abaluck, Nikhil Agarwal, Susan Athey, John Campbell, Dennis Carlton, Judy Chevalier, John Cochrane, Christopher Conlon, Shane Corwin, Peter Cranton, 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 suggestions. We are also very grateful to seminar audiences at Chicago, Yale, Northwestern, NYU, Berkeley, Harvard, MIT, UPenn, Columbia, HKUST, KER, the Economics of Platforms Workshop, NBER IO, and SEC DERA. 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

“We must consider, for example, whether the increasingly expensive search for speed has passed the point of diminishing returns. 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. ... A key question is whether trading venues have sufficient opportunity and flexibility to innovate successfully with initiatives that seek to deemphasize speed as a key to trading success in order to further serve the interests of investors. If not, we must reconsider the SEC rules and market practices that stand in the way.” (Securities and Exchange Commission Chair Mary Jo White, June 2014)

As of 2021, all 16 stock exchanges in the United States use a market design called the continuous limit order book. Academic research has shown that this market design gives rise to a phenomenon called “latency arbitrage” (Budish, Cramton and Shim, 2015), defined formally as arbitrage rents from symmetrically observed public information, as distinct from the rents from asymmetric private information that are at the heart of classic models of market microstructure (Kyle, 1985; Glosten and Milgrom, 1985). Latency arbitrage rents, in turn, lead to a socially wasteful arms race for trading speed. The race for trading speed is currently measured in microseconds (millionths of a second) and even nanoseconds (billionths). The “size of the prize” in the arms race for speed has recently been estimated at $5 billion per year in global equities markets (Aquilina, Budish and O’Neill, 2020). This same study estimates that eliminating latency arbitrage would reduce the cost of liquidity in equities markets by 17%.

This paper studies the incentives of stock exchanges to innovate to address latency arbitrage and the arms race for speed. Researchers and practitioners have identified two market design innovations that could, in theory, address the problem: asymmetric speed bumps and frequent batch auctions. Asymmetric speed bumps slow down orders that may potentially “snipe” stale quotes, thus giving a tiny head start to firms providing liquidity in the event of a race to respond to public information (Baldauf and Mollner, 2020). Frequent batch auctions slow down all orders, and then batch process them in discrete time using a uniform-price auction. The discrete-time interval gives firms providing liquidity time to cancel their stale quotes (like in the asymmetric delay mechanism) and the auction ensures that any trades that do occur are at a price that reflects the new public information, as opposed to a price that just became stale (Budish, Cramton and Shim, 2015).

Implicit in the quote at the top of the paper, delivered in a speech by then SEC Chair Mary Jo White, is the view that private and social incentives for market design innovation are aligned: if there is a market design innovation that is efficiency enhancing, then private market forces will naturally evolve towards realizing the efficiency if allowed to do so. In this view, “prescriptive regulation” of a specific market design is not necessary, and regulators should instead ensure that they do not
inadvertently “stand in the way” of “competitive solutions.” This is a natural instinct — the standard case in economics is that if there is a large inefficiency in a market, there will be private incentive to fix the inefficiency (e.g., Griliches, 1957). However, as is well known, there are numerous economic settings where private and social incentives for innovation diverge (Arrow, 1962; Nordhaus, 1969; Hirshleifer, 1971).

This paper’s main insight is that incumbent stock exchanges’ private incentives to innovate their market designs are misaligned with social interests because they earn economic rents from the arms race for speed. That is, even though there are many exchanges with little differentiation that appear to compete fiercely with one another for trading volume, they capture and maintain a significant share of the economic rents from latency arbitrage. We emphasize that in this paper’s theory the disincentive to adopt is not due to liquidity externalities, multiple equilibria due to coordination failure, chicken-and-egg, etc., as is central in the literature on network effects and platform competition (e.g., Farrell and Saloner, 1985; Katz and Shapiro, 1986; Rochet and Tirole, 2003; Farrell and Klemperer, 2007) and past market microstructure literature on financial exchange competition (see surveys by Madhavan, 2000 and Cantillon and Yin, 2011). Rather, our theory in the end is ultimately a novel version of the old economic ideas of incumbents protecting rents and missing incentives for innovation.

The first part of the paper builds a new theoretical model of financial exchange competition, tailored to the institutional and regulatory details of the modern U.S. stock market. The goal of the model is both to better understand the economics of stock exchange competition under the status quo, in which all exchanges employ the continuous limit order book market design, and to be able to analyze exchanges’ incentives for market design innovation. There are four types of players in our model, all strategic: exchanges, trading firms, investors, and informed traders. Exchanges are modeled as undifferentiated and they strategically set two prices: per-share trading fees, and fees for speed technology that enables trading firms to receive information about and respond more quickly to trading opportunities on a given exchange. In practice, speed technology includes co-location (the right to locate one’s own servers right next to the exchange’s servers) and proprietary data feeds (which enable trading firms to receive updates from the exchange faster than from non-proprietary data feeds). Trading firms choose the set of exchanges to buy speed technology from. They then choose whether and how to provide liquidity by choosing the exchange(s) on which to offer liquidity, the quantity to offer on each exchange, and a bid-ask spread on each exchange. The bid-ask spread trades off the benefits of providing liquidity to investors (thereby collecting the spread) versus the cost of either being adversely selected against by an informed trader (as in Glosten and Milgrom, 1985) or being on the losing end of a latency arbitrage race with other trading firms — i.e., being “sniped” (as in Budish, Cramton and Shim, 2015).

Our analysis of the status quo delivers three main results. First, as in Glosten (1994), although the market can be fragmented in the sense that trading activity is split across several exchanges, economically many aspects of trading activity behave as if there is what we call a “virtual single
Specifically: all liquidity is at the same prices and bid-ask spreads regardless of the
exchange on which it is offered, with the marginal unit of liquidity indifferent across exchanges due
to a one-for-one relationship between the quantity of liquidity on an exchange (i.e., market depth)
and the quantity of trade on that exchange (i.e., volume); and aggregate depth and volume are both
invariant to how trading activity is allocated across exchanges. This behavior is brought about by two
key sets of regulations in the U.S.: Unlisted Trading Privileges (UTP) and Regulation National Market
System (Reg NMS).\(^1\) UTP essentially implies that stocks are perfectly fungible across exchanges: i.e.,
a stock that is technically listed on exchange X can be bought on any exchange Y and then sold
on any exchange Z. Reg NMS ensures that searching among exchanges, and then transacting across
(“accessing”) them, are both frictionless. This frictionless search and access allows market participants
to costlessly “stitch together” the order books across the various exchanges, and yields investor demand
that is perfectly responsive to price differences across exchanges. This behavior also leads to our second
result: due to the same frictionless search and access, investor demand is perfectly elastic with respect
to trading fees as well; hence, fierce Bertrand-style competition yields competitive (zero) trading fees
on all exchanges.

Our third result is that exchanges can both capture and maintain substantial rents from the sale of
speed technology. This may appear surprising as exchanges are modeled as undifferentiated and search
and access is frictionless; as we have mentioned, these same features lead to competitive trading fees.
There are two reasons why exchanges earn supra-competitive rents for speed technology in equilibrium.
First, even though stocks are fungible across exchanges, latency-sensitive trading opportunities are not:
if there is a sniping opportunity that involves a stale quote on Exchange X, only trading firms that
have purchased Exchange X speed technology will be able to effectively compete in the sniping race.
Second, in contrast to basic models of add-on pricing whereby profits from add-on goods are dissipated
by firms selling the primary good below cost (cf. Ellison, 2005; Gabaix and Laibson, 2006), exchange
rents earned from the sale of speed technology are not dissipated via further competition on trading
fees. The reason is that trading fees are already at zero, and cannot become negative without creating
a “money-pump” wherein trading firms execute infinite volume to extract the negative fee.

We also prove that although exchanges are modeled as price setters who post take-it-or-leave-it
offers to trading firms for speed technology, exchanges nevertheless cannot extract all of the industry
rents from latency arbitrage. Taking our bound literally, and using realistic parameters for the numbers
of fast trading firms and exchanges, our model suggests that exchanges in aggregate can extract at
most 30% of the total latency arbitrage prize. The reason is that trading firms are able to influence
where volume is transacted, and this allows them to discipline exchanges that attempt to take too
much of the pie.

\(^1\)Glosten (1994) presciently foresaw that frictionless search and order-splitting across electronic markets (see his
Assumption 4) could generate what we refer to as a virtual single platform (see his Proposition 8), well over a decade
before the passage of Reg NMS. Please see Section 3.3 for discussion of the relationship between Glosten (1994) and this
aspect of our analysis.

\(^2\)These regulations are described in detail in Section 2 and Appendix A.
This model of the status quo is of course stylized, and in particular abstracts from many important aspects of real-world equity markets including agency frictions, tick-size constraints, asymmetric trading fees, the opening and closing auctions, and strategic trading over time as in Kyle-style models. Nonetheless, we establish that the model does reasonably well empirically, by documenting a series of stylized facts that relate to each of the model’s three main results. This work utilizes both the well-known trades-and-quotes (TAQ) dataset as well as information gleaned from various exchange-company financial documents (e.g., 10-K’s, S-1’s, merger proxies, fee schedules). The goal of the empirics is not to persuade the reader that the model is “correct” (no model is), but rather simply to suggest that our parsimonious model of a complicated industry is sensible.

The first set of stylized facts relates to our result that the market behaves as if trading activity occurred on a virtual single platform. Specifically, using a sample of highly traded stocks, we show that (Stylized Fact #1) all major exchanges typically have displayed liquidity at the same best bids and asks, and (SF#2) there is a one-to-one relationship between the quantity of liquidity on an exchange and its trading volume, which is what makes the marginal unit of liquidity indifferent across exchanges in our model. We also show (SF#3) that exchange market shares are interior and relatively stable (i.e., no tipping), both overall and at the level of individual stocks; this pattern is not a prediction of our model per se but makes the virtual single platform aspect of equilibrium more plausible. The second set of stylized facts relates to our results about trading fees. Trading fees are quite complicated, but using a variety of data sources to cut through this complexity, we compute that (SF#4) the average fee for regular-hours trading, across the three largest stock exchange families, is around $0.0001 per share per side. This is not zero, as the theory predicts, but is small. We also show (SF#5) that fees do in fact bump up against the money-pump constraint, as suggested by the theory. The last set of stylized facts relates to our results about exchange-specific speed technology. We document (SF#6) that exchanges earn significant revenues from the sale of co-location services and proprietary data feeds. We also document (SF#7) significant growth in these revenue sources during the Reg NMS era. We estimate that annual exchange-specific speed technology revenues are on the order of $1 billion.

The last part of the paper uses the model to study exchanges’ private incentives to adopt new market designs that address latency arbitrage. How do exchanges’ private innovation incentives relate to social incentives, for both incumbents and de novo entrants? To conduct this analysis, we extend our theoretical model to allow for exchanges to operate one of two market designs: either the continuous-time limit order book (Continuous), or discrete-time frequent batch auctions (Discrete). Importantly, in the context of competition with the Continuous market, we consider frequent batch auctions with a very short batch interval: long enough to effectively batch process if multiple trading firms react to the same public signal at the same time, but otherwise essentially as short as possible.\(^3\)

\(^3\)In practice, given advances in speed technology over the last several years, 500 microseconds to 1 millisecond would likely be more than sufficient to effectively batch process; some industry participants have argued to us that as little as 50 microseconds (i.e., 0.000050 seconds) might suffice. A short batch interval would also allow the frequent batch auction exchange to satisfy the SEC’s de minimis delay standard and have protected quotes under Reg NMS, which is significant. Please see Section 5 for the full details of how we model frequent batch auctions, including the important
We first study a market in which one exchange employs Discrete while all others employ Continuous. A natural prior is that there will be multiple equilibria, including an equilibrium in which the new exchange fails to take off. In many models of platform competition, there exist equilibria where a new platform fails to take off even if in principle it is better designed, if that is what market participants expect to happen. Instead, we find that there is a unique equilibrium in which the Discrete exchange attracts all trading volume. The reason is the frictionless search. Intuitively, eliminating latency arbitrage eliminates a tax on liquidity, and the fact that market participants can frictionlessly access and search across exchanges ensures that if there are two markets operating in parallel, one with a tax and one without, the one without the tax will take off. The Discrete exchange earns rents in this equilibrium by charging trading fees that are positive but less than the latency arbitrage tax that it eliminates.

We next study a market in which multiple exchanges employ Discrete. Unfortunately for the innovator, the frictionless search that enables an initial Discrete exchange to get off the ground is a double-edged sword. We show that in any equilibrium with multiple Discrete exchanges trading fees are competed down to zero, and trading volume is split among Discrete exchanges with zero fees. That is, we have the same Bertrand competition on trading fees as in the Continuous status quo, but now without the industry rents from the speed race.

Together, these two results imply that the market design adoption game among incumbent exchanges can be interpreted as a prisoner’s dilemma: while any one exchange has incentive to unilaterally adopt Discrete, all incumbents prefer the Continuous status quo, in which they share in latency arbitrage rents, to the counterfactual in which all exchanges are Discrete, and these rents are gone. This in turn implies that if an incumbent considering whether to adopt Discrete anticipates that imitation by other incumbents would be sufficiently rapid, it would prefer to remain Continuous. A de novo entrant weighing whether to enter as a Discrete exchange faces a similar tradeoff, except that they must overcome fixed costs of entry, rather than opportunity costs of losing latency arbitrage rents.

We emphasize that while this analysis identifies an important wedge between private and social incentives to innovate, it does not imply that the private incentives to innovate are strictly negative. This depends on parameters such as the cost of adoption, speed of imitation, and the magnitude of the latency arbitrage prize. That said, as of this writing, there has been just one proposal for an exchange design that eliminates latency arbitrage in a displayed, Reg NMS protected market. This proposal came from the smallest incumbent exchange (the Chicago Stock Exchange, with just 0.5% share at

---

4This result may seem to contradict the well-known result in Glosten (1994), referenced earlier, that finds that the electronic limit order book is in a certain sense “competition proof” (Proposition 9). The explanation is that the Glosten (1994) model implicitly precludes the possibility of latency arbitrage. The reason Discrete wins against Continuous in our model is precisely because it eliminates latency arbitrage. Please see Section 5.2.
the time), which had perhaps uniquely low costs of adoption: as an incumbent it did not have to pay fixed costs of entry, while given its small market share it did not face significant opportunity costs from the loss of speed-technology rents. CHX’s entry attempt thus suggests that if adoption costs are low enough, there could indeed be market design innovation that addresses latency arbitrage, which is consistent with our model. However, CHX ultimately withdrew its proposal after being acquired by the New York Stock Exchange group.5

Our analysis also yields a clear, and perhaps surprising, insight about the potential role for policy. A reasonable 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 — which of course raises all of the usual concerns about regulatory mandates as discussed by Chair White. Our results suggest a third possibility to consider: a regulatory “push.” By “push” we mean any policy that tips the balance of incentives sufficiently to get a de novo exchange to enter or an incumbent to adopt. Our analysis shows that such an initial entrant will gain share, which would not necessarily be the case in a coordination game environment. Back-of-envelope calculations suggest that the magnitude of the push could be very modest relative to the stakes.

**Contribution to the Platforms Literature: Virtual Single Platform** In traditional models of platform competition, some or all participants on one or both sides of the market single-home — that is, they interact exclusively with one platform (Rochet and Tirole, 2003; Caillaud and Jullien, 2003; Armstrong, 2006; Rochet and Tirole, 2006). This single homing has crucial implications for the economics of industries with two-sided platform competition, such as payment systems, online marketplaces, smart phones, and video games. Platforms derive network effects and market power from their single-homing participants, which gives them the ability to charge supra-competitive prices. The fight for single-homing participants and the associated network effects can lead to market tipping. See Figure 1.1, Panel A.

In our setting, because of the technological and regulatory environment for modern stock exchanges, participants on both sides of the market multi-home and search across the many exchanges is very low friction. As a result, even though stock exchange competition looks from a distance like two-sided platform competition, and indeed much past research on the economics of financial exchanges has used these kinds of ideas (see Section 4.4 for references and Cantillon and Yin (2011) for a survey), the industry economics instead collapses into what we call a “virtual single platform”, on which trading fees are competitive and network effects are nullified. See Figure 1.1, Panel B.

5To date, there have been three other exchange design proposals in the US that relate to latency arbitrage but for quotes outside of the displayed, Reg NMS protected market: the IEX 350 microsecond symmetric speed bump, the Cboe EDGA 4 millisecond asymmetric speed bump, and the Cboe BYX 100 millisecond user-initiated batch auctions. Also important is a recently-approved order type on IEX’s continuous exchange, D-Limit, which protects users of that order type against certain latency arbitrage opportunities detected by IEX based on price updates from other continuous exchanges. Please see Appendix B for brief details about each of these proposals.
It is striking empirically just how low the trading fees are as a result of this competition — we estimate $0.0001 per share per side, or about 0.0002% per share per side for a $50 stock. In aggregate, across the approximately 1 trillion shares and $50 trillion of value that is traded during regular hours each year, the entire US stock exchange industry earns just $200 million of trading fee revenue. In contrast, many internet-enabled platform markets, for items such as tickets, ride-sharing, food delivery, vacation rentals, and so forth, have fees of about 10-30% — or about 50,000 times higher on a percentage basis.\footnote{There are many reasons why this comparison is not apples-to-apples, such as fraud costs in online marketplaces, but still the contrast is striking. For example, StubHub, the largest secondary market venue for concert and sports tickets, has annual revenues exceeding $1 billion, or more than 5 times the regular-hours trading fee revenues for the entire US stock market, despite having less than 1/10000th the trading volume.}

Our theory suggests that where exchanges are able to capture supra-competitive profits is the sale of speed technology, with these rents protected from dissipation by a zero lower bound on trading fees. Even though search and multi-homing are frictionless and exchanges are modeled as undifferentiated, sniping opportunities have to arise somewhere, and each exchange can exclusively sell privileged access to the speed-sensitive trading opportunities that arise on its exchange. See Figure 1.1, Panel C.

**Contribution to the Market Design Literature: Incentives for Market Design Innovation**

The other literature most closely related to our paper is market design. Research in market design has been extremely active in the past few decades, since the seminal designs of the medical match and spectrum auctions.\footnote{For recent surveys of aspects of this literature, please see Roth (2018), Milgrom (2021), Pathak (2017), Kominers, Teytelboym and Crawford (2017), and Agarwal and Budish (2021).} However, to our knowledge, ours is the first paper in this literature to explicitly analyze market design adoption and innovation incentives. Often, the approach taken in this literature is to diagnose an existing market institution with some important flaw, and then to use theory and other complementary tools to design an alternative mechanism argued to be conceptually superior. But this approach elides the question of whether the new mechanism will be implemented in the world,
given the incentives of the parties with the power to change the market design. Notably, in many prior market design implementations that are well known to the literature — spectrum auctions, the medical match, school choice, kidney exchange, course allocation — the entities making adoption decisions are mainly governments or non-profit organizations. Here, the key players are for-profit exchanges, and the theory suggests that their private interests may be in tension with overall social welfare.

**Roadmap.** The remainder of this paper is organized as follows. Section 2 describes the institutional and regulatory details that inform the theoretical model. Section 3 presents and analyzes the theoretical model focusing on exchange competition under the status quo market design. Section 4 presents the seven stylized facts. Section 5 uses the model to analyze exchange competition when there are competing market designs. Section 6 discusses policy implications. Section 7 concludes.

## 2 Institutional Background

Readers of this paper — especially researchers who are less familiar with financial market microstructure — may have in mind, when thinking of stock exchanges and how they compete, the old New York Stock Exchange floor. As recently as the 1990s, if a stock was listed on the New York Stock Exchange, the large majority of its trading volume (65% in 1992) transacted on the New York Stock Exchange floor. Similarly, if a stock was listed on Nasdaq, a large majority of its volume transacted on the Nasdaq exchange (86% in 1993). In this earlier era, stock exchanges enjoyed valuable network effects and supra-competitive fees. The seminal model of Pagano (1989) — in which traders single-home, and there are liquidity externalities that can cause traders to agglomerate on an exchange with supra-competitive fees — was a reasonable benchmark for thinking about the industrial organization of the industry.

This model, however, is less applicable for the modern era of stock trading. In our data, from 2015, there are 12 exchanges, all stocks trade essentially everywhere, and market shares are both stable and interior (i.e., no tipping). There are 5 exchanges with greater than 10% market share each (83% in total), and the next 3 exchanges together have another 15% share. Please see our discussion of Stylized Fact #3 in Section 4.1 for further details. Trading fees, while complex and somewhat opaque, are ultimately quite small, as we will document as Stylized Fact #4 in Section 4.2.

There are two key sets of regulations that together shape the industrial organization of modern electronic stock exchanges. We describe them briefly here, and provide further details in Appendix A.

---

8Kidney exchange is a partial exception to this claim, and the private adoption incentives of hospitals may be in tension with overall social welfare as discussed in Ashlagi and Roth (2014) and Agarwal et al. (2019).

9Technically, stocks could not be “listed” on Nasdaq until it became an exchange in 2006, but the 1975 Exchange Act Amendments enabled stocks to trade over-the-counter via Nasdaq achieving something economically similar. For the NYSE market share claim, see the SEC study “Market 2000”, Exhibit 18 (U.S. Securities and Exchange Commission, 1994). For the Nasdaq market share claim, see the SEC Market 2000 study, Exhibit 12.

10For surveys of modern electronic trading, focusing on a broader set of issues than stock exchanges per se, good starting points are Jones (2013), Menkveld (2016) and Fox, Glosten and Rauterberg (2019).
The first set of regulations, related to 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 that of a “protected quote” on another exchange. Sophisticated market participants can take on responsibility for compliance with the Order Protection Rule themselves, absolving exchanges of the responsibility for checking quotes on other exchanges, by using an order type denoted “intermarket sweep order” (ISO). 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. The Access Rule ensures that such efficient “search and access” is feasible — i.e., the Access Rule (and related rules that affect information provision, such as those governing slower, non-proprietary market data feeds) enables market participants to both search available quotes and then “access” them, i.e., trade against them, with the only marginal costs of accessing a particular quote on a particular exchange being that exchange’s per-share trading fees. 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.

3 Theory of the Status Quo

We now introduce our model of stock exchange competition. Section 3.1 presents the setup and timing of the model. Section 3.2 analyzes the model’s equilibria. Section 3.3 discusses the key economic aspects of equilibria. For this Section, we restrict all exchanges to employ the continuous limit order book market design; later, in Section 5 we extend the model and allow exchanges to be strategic with

---

11 A quote on a particular exchange is considered protected if (i) it is at that exchange’s current best bid or offer, and (ii) it is “immediately and automatically accessible” by other exchanges. Reg NMS does not provide a precise definition of “immediately and automatically accessible,” but the phrase certainly included automated electronic continuous limit order book markets and certainly excluded the NYSE floor system with human brokers. A June 2016 rules clarification issued by the SEC indicated that exchanges can use market designs that impose delays on the processing of orders and still qualify as “immediate and automatic” so long as (i) the delay is of a de minimis level of 1 millisecond or less, and (ii) the purpose of the delay is consistent with the efficiency and fairness goals of the 1934 Exchange Act (U.S. Securities and Exchange Commission, 2016b).
respect to their market design choice.

3.1 Model

Our model adapts and extends the framework introduced in Budish, Cramton and Shim (2015) (hereafter, BCS). We depart from it in the following ways. First and foremost, whereas BCS examined trading on a single non-strategic exchange, our model has multiple exchanges who strategically choose trading fees and speed technology fees in an environment shaped by UTP and Reg NMS. Second, we introduce a stylized version of informed trading in the spirit of Copeland and Galai (1983) and Glosten and Milgrom (1985), in order to parsimoniously incorporate traditional adverse selection from informed trading alongside latency arbitrage. Third, rather than working with a continuous-time model in which events occur according to exogenous Poisson processes, we instead work with an infinitely repeated two-period trading game where, in each play of the trading game, either 0 or 1 exogenous events occur. We view each trading game as lasting a sufficiently short amount of time — e.g., 1 millisecond or potentially even shorter — that the 0 or 1 exogenous events assumption reasonably approximates reality.\(^\text{12}\) This approach will retain the economic interpretability of the continuous-time Poisson model used in BCS while providing tractability when modeling trading behavior across multiple exchanges. Last, we develop and employ an alternative equilibrium solution concept for our trading game, order-book equilibrium. It is well known that Nash equilibria can fail to exist in environments with adverse selection, such as insurance markets (Rothschild and Stiglitz, 1976) and limit order book markets with private information (Glosten and Milgrom, 1985). Our alternative concept guarantees that an equilibrium exists for our trading game in a manner similar to alternative equilibrium notions developed to analyze insurance markets (Wilson, 1977; Riley, 1979).

3.1.1 Setup

There is a single security, \(x\), and a signal of the value of the security, \(y\). We make the purposefully strong assumption that the signal \(y\) is equal to the fundamental value of \(x\), and that \(x\) can always be costlessly liquidated at this fundamental value. The signal \(y\) evolves as a discrete-time jump process, where jumps occur with some positive probability per trading game and the value of the jumps is drawn from a symmetric distribution with bounded support and zero mean. What will matter economically is the absolute value of jumps, represented by random variable \(J\). We refer to the distribution of \(J\) as the jump-size distribution.

There are \(M\) exchanges, exogenously present in the market, across which security \(x\) can be bought or sold. Our main analysis will focus on the case of \(M \geq 2\) exchanges. Exchanges all use the continuous

\(^{12}\)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.
limit order book market design and are ex-ante undifferentiated. The security \(x\) is completely fungible across exchanges, that is, it can be bought or sold on any exchange 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.\(^{13}\)

There are four types of players: Investors, Informed Traders, Trading Firms, and Exchanges. We refer to the first three types of players as *market participants*. All players are risk-neutral.

An Investor arrives stochastically with probability \(\lambda_{\text{invest}}\) in each trading game, and has an inelastic need to buy or sell one unit of \(x\), with buying or selling equally likely.\(^{14}\) An investor can trade a single time across multiple exchanges using marketable limit orders (i.e., an investor is restricted to being a “taker,” and not a “maker,” of liquidity), and then exits the game. Formally, if an investor arrives to market needing to buy one unit of \(x\), buys a unit at price \(p\), and the fundamental value is \(y\), then her payoff is \(v + (y - p)\), 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 \(v + (p - y)\).\(^{15}\)

An Informed Trader with private information about the fundamental value of \(x\) also arrives stochastically to the market. In BCS, all jumps in \(y\) were public information. Here, we assume that jumps in \(y\) can be either public information, seen by all players at the same time, or private information, seen only by a single informed trader. Specifically, in each trading game, the probability that there is a jump in \(y\) that is public information is \(\lambda_{\text{public}}\), and the probability that there is a jump in \(y\) seen by an informed trader is \(\lambda_{\text{private}}\). For simplicity, both public and private jumps have the same jump-size distribution, with positive and negative changes being equally likely. 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 the informed trader exits and any privately observed information becomes public. The informed trader’s payoff, if he buys a unit of \(x\) at price \(p\) and the (new) fundamental value is \(y\), is \(y - p\); if he sells at price \(p\) his payoff is \(p - y\).\(^{16}\)

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. If they buy (or sell) a unit of \(x\) at price \(p\), and the fundamental value is \(y\) at the end of the trading game, their payoff is \(y - p\) (or \(p - y\)). Their objective is to maximize per-trading game profits. We assume that there are \(N\) “fast” trading firms that possess a general-purpose speed technology that enables their orders to be processed ahead of those without such technology. There is also a continuum of

---

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 BCS (pg. 1583-1586) it is possible to generalize the model to investors with varying-sized demands (e.g., some require “one” unit, some require multiple units) as long as all investors trade a single time upon arrival, but the model does not accommodate strategic trading over time as in Kyle (1985) or Vayanos (1999).

15 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.

16 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).
“slow” trading firms that do not possess such technology. 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.

**Exchanges**, indexed by $j$, simultaneously set two prices prior to play of the infinitely-repeated trading game: (i) a per-share trading fee denoted by $f_j$, and (ii) an exchange-specific speed technology fee denoted by $F_j$. The trading fee $f_j$ is assessed per share traded and is paid symmetrically by both sides of any executed trade.\(^\text{17}\) The exchange-specific speed technology (abbreviated ESST) fee $F_j$ represents the price of co-location (the right to locate one’s servers next to an exchange’s servers), access to fast exchange-specific proprietary data feeds, and connectivity/bandwidth fees.\(^\text{18}\) In reality, such technology allows trading firms to receive information about and respond more quickly to trading opportunities on a given exchange. In our model, we treat speed technology as a tie-breaker (as in Baldauf and Mollner, 2020), meaning that 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.\(^\text{19}\) We assume that the processing order on an exchange is uniformly random among firms in each of these speed groups. ESST fees are modeled as a rental cost per trading game charged to TFs, capturing that in practice exchanges typically assess these fees on a rental basis.

We also require that each exchange sell ESST to at least 2 trading firms or not sell ESST at all.\(^\text{20}\) We believe that this modest fair access requirement — which in essence prevents an exchange from auctioning off exclusive access to ESST — is consistent with the statutory requirement under the Exchange Act that fees are “fair and reasonable and not unreasonably discriminatory” (see Clayton, 2018). For this reason, we also assume that the number of TFs endowed with general-purpose speed technology is at least $N \geq 3$: with only two fast TFs, either one would be able to unilaterally deny usage of ESST on any exchange to the other one by not purchasing.

### 3.1.2 Timing

There are three stages to our game.

First, in Stage One (Exchange Price Setting), all $M$ exchanges simultaneously choose per-share trading fees $f = (f_1, \ldots, f_M)$ and per-trading game ESST rental fees $F = (F_1, \ldots, F_M)$.

Next, in Stage Two (Speed Technology Adoption), all $N$ TFs with general speed technology simultaneously decide which exchanges to purchase ESST from.

---

17 In practice exchanges often charge different fees for “making” liquidity as opposed to “taking” liquidity; see Section 4.2 and Chao, Yao and Ye (2019). However, the assumption of symmetric fees is without loss of generality in our model: since prices are continuous, only the net trading fee matters for determining equilibrium behavior.

18 In practice the dividing line between exchange-specific technology and general-purpose technology is not sharp — for example, latency sensitive code might be adapted to a particular exchange’s data protocol, and some communications links are specific to a particular exchange’s data center. The important thing to capture is that each exchange controls some but not all of the technology that is necessary to be fastest on their own exchange.

19 For simplicity we do not allow slow TFs to purchase ESST. In the equilibria we characterize they would not want to.

20 If only a single TF purchases ESST from an exchange, we assume that the TF is not allowed to use the speed technology on that exchange and both the TF and the exchange incur a strictly positive non-compliance cost.
Last, in Stage Three (Infinitely Repeated Trading Game), the following two-period trading game is repeated infinitely often.

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 ($\omega = (\omega_1, \ldots, \omega_M)$, where $\omega_j$ is also referred to as the state of exchange $j$’s order book).¹¹ TFs then simultaneously submit message sets to any subset of exchanges, where $\mu_{ij} \in S$ denotes the set of messages TF $i$ submits to exchange $j$, and $S$ denotes the set of all potential combinations of messages. We allow for three types of messages: standard limit orders, cancellations of existing limit orders, and immediate-or-cancel orders. Standard limit orders sent to an exchange take the form $(q_i, p_i)$, which indicates that the firm is willing to buy (if $q_i > 0$) or sell (if $q_i < 0$) up to $|q_i|$ units at price $p_i$. An immediate-or-cancel order (abbreviated IOC) behaves similarly to a standard limit order, but with proxy instructions to cancel the limit order at the end of the period if it is not executed (or to cancel whatever portion is not immediately executed). 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 $\omega_j$, if any exist. For each exchange $j$, all message sets sent to exchange $j$ in this period are serially processed by the exchange in a random sequence, with TF speed serving as a tie-breaker (in the manner described above).

After period-1 message sets have been processed by each exchange and each exchange’s order book has been updated, in period 2 of each trading game nature moves and selects one of four possibilities:

1. With probability $\lambda_{\text{invest}}$: 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. The investor’s activity may affect $\omega$; $y$ is unchanged.

2. With probability $\lambda_{\text{private}}$: an informed trader privately observes a jump in $y$. The informed trader has a single opportunity to send IOCs to all exchanges. The informed trader’s activity may affect $\omega$; the jump in $y$ is then publicly observed.

3. With probability $\lambda_{\text{public}}$: there is a publicly observable jump in $y$. All TFs have a single opportunity to submit an order consisting of IOCs and cancellation messages to each exchange. For each exchange $j$, orders sent to exchange $j$ in this period are serially processed in a random sequence by the exchange, with speed serving as a tie-breaker as in period 1. Orders sent in this period may affect $\omega$.

4. With probability $1 - \lambda_{\text{invest}} - \lambda_{\text{private}} - \lambda_{\text{public}} \geq 0$: there is no event; $y$ and $\omega$ are both unchanged.

3.1.3 Modeling Institutional Details

To close the setup of the model, we briefly discuss how the model incorporates institutional details described in the previous Section. First, we incorporate UTP into our model by having the same

¹¹If it is the first play of the trading game each exchange’s order book is initially empty.
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. Second, we capture key aspects of Reg NMS by implicitly assuming in our trading game that all market participants face, on an order-by-order basis, what we call frictionless search and access. More specifically, by frictionless search we mean that all market participants observe the current state of the order book on all exchanges, ω, at zero cost prior to taking any action. By frictionless access we mean that the marginal cost of sending any message to any exchange is zero; equivalently, the only per-order cost of transacting on any exchange is the per-share trading fee.

An additional institutional detail that we incorporate into the model is the ability for investors and informed traders to synchronize their orders across exchanges such that they can execute trades across multiple exchanges before other market participants can react. We allow 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 period 1 of the subsequent trading game.

3.2 Equilibrium Analysis

3.2.1 Stage Three Trading Game

We begin our equilibrium analysis by examining the Stage 3 trading game. We initially examine the special case where there is only a single exchange (M = 1) that sets trading fees of zero (f_1 = 0), and all N trading firms with general speed technology also have exchange-specific speed technology on this exchange so that they are equally fast. Focusing on this special case is helpful for two reasons. First, it is the most intuitive environment in which to introduce and motivate the solution concept of order book equilibrium. Second, the economics in this simpler environment will be a helpful guide to the economics of the full exchange competition game and our “virtual single platform” result.

For Stage 3 (both here and later with multiple exchanges), we restrict attention to pure Markov strategies: market participants play pure strategies that may only condition on the publicly observable state, and not on the history of play in previous trading games.

Period 2: Optimal Play. Regardless of which outcome nature chooses in period 2 of a given trading game, market participants’ optimal strategies in period 2 are straightforward to characterize:

---

22 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 BCS. Rather, our model is of regular-hours stock exchange trading (about 90% of exchange volume), for which UTP makes the listing exchange irrelevant.

23 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 (Lewis, 2014), and is modeled carefully in Baldauf and Mollner (2020).

24 For expositional simplicity, we initially analyze each trading game in isolation, thereby ignoring the possibility that actions in one trading game may affect continuation payoffs in subsequent games. In our formal proofs, we check and show that repeated play of the equilibrium that we construct for a single trading game remains an equilibrium for the infinitely repeated trading game when such interactions are accounted for.
• **Investor or Informed Trader Arrival.** If either an investor arrives or an informed trader arrives, they have essentially unique optimal strategies. An investor sends an IOC order to trade up to one unit in their desired direction; 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 (this latter case will not occur on the equilibrium path). An informed trader sends an IOC order to trade against any orders that are profitable to trade against based on their privately observed $y$.

• **Publicly Observed Jump.** If 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 to trade given the state of the order book (i.e., $y$ increases to a price lower than the best ask or decreases to a price higher than the best bid), then no trades occur. Second, if $y$ jumps to a value at which it is profitable to trade given the outstanding bids and asks in the exchange’s order book, there is a sniping race as described in BCS: any fast TFs that are providing liquidity at unprofitable prices send cancellation messages to the exchange to try to cancel these stale quotes, while at the same time all other fast TFs send IOCs to the exchange to try to trade against (“snipe”) these stale quotes. Note that fast TFs may simultaneously try to cancel their own quotes and snipe others’ quotes. Since the processing order among the $N$ fast TFs is uniformly random, fast TFs providing liquidity will get sniped with probability $\frac{N-1}{N}$. If a slow TF is providing liquidity, all $N$ fast TFs try to snipe them and the slow TF is sniped with probability 1. Either way, fast TFs attempting to snipe succeed with probability $\frac{1}{N}$.

We assume that market participants follow these optimal strategies in period 2, conditional on the stochastic decision by nature. The analysis of each trading game then simplifies to understanding TF behavior in period 1.

**Period 1: Bid-Ask Spread Indifference Condition.** 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 period-1 liquidity provision via pairs of limit orders: 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 $s \geq 0$ is the bid-ask spread.

Initially, consider a single fast TF who provides $q = 1$ of liquidity at spread $s$ in period 1, while no other TFs provide any liquidity in period 1. This fast TF’s liquidity has benefits and costs. The benefit is if an investor arrives and trades, paying the spread: in expectation, this is worth $\lambda_{\text{invest}} \cdot \frac{s}{2}$ per trading game. The costs arise from two sources. First, the liquidity may be adversely selected if an informed trader arrives with private information. Per trading game, this cost is $\lambda_{\text{private}} \cdot L(s)$, where $L(s) \equiv \Pr(J > \frac{s}{2}) \cdot E(J - \frac{s}{2} | J > \frac{s}{2})$ is the expected adverse selection loss to a liquidity provider provided $J > \frac{s}{2}$.

---

25If there are multiple optimal strategies, they differ only for prescribed actions over which agents are indifferent: e.g., whether or not to trade against orders that yield zero profits given the value of $y$.

26Any TF providing liquidity that wishes to replace an order is indifferent between canceling that order immediately and waiting until the beginning of the following trading game to do so.
upon arrival of a privately observed jump in \( y \). Second, if there is public information the TF may get sniped. Per trading game, this cost is \( \lambda_{\text{public}} \cdot \frac{N-1}{N} \cdot L(s) \), where the \( \frac{N-1}{N} \) term reflects the probability that a fast TF loses the sniping race, and the \( L(s) \) term is the same as above because we have assumed for convenience that private and public information jumps have the same distribution.\(^{27}\) For a fast TF to be indifferent between providing a unit of liquidity at spread \( s \) and sniping a rival TF offering liquidity at that same spread (succeeding with probability \( \frac{1}{N} \)), the spread must satisfy:

\[
\lambda_{\text{invest}} \cdot \frac{s^*_{\text{continuous}}}{2} = (\lambda_{\text{public}} + \lambda_{\text{private}}) \cdot L(s^*_{\text{continuous}}). \tag{3.1}
\]

At any spread wider than \( s^*_{\text{continuous}} \) fast TFs strictly prefer liquidity provision to sniping, and at any narrower spread they strictly prefer sniping to liquidity provision.

This same bid-ask spread \( s^*_{\text{continuous}} \) also leaves a slow TF with zero profits from liquidity provision — i.e., a slow TF is indifferent between providing liquidity and doing nothing.\(^{28}\) In equilibrium as described below, there can be a mix of fast and slow TFs providing liquidity at \( s^*_{\text{continuous}} \), and there can be multiple TFs each providing a fraction of the aggregate liquidity — e.g., one TF provides 0.6 at \( s^*_{\text{continuous}} \) 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.

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

**Order Book Equilibrium.** Given optimal period-2 play as described above and our restriction to pure Markov strategies, a natural solution concept for period-1 behavior in the infinitely repeated Stage 3 trading game would be pure-strategy Markov perfect equilibrium (MPE). For a single play of the Stage 3 trading game, MPE is equivalent to Nash equilibrium. However, neither a MPE of the repeated trading game, nor a Nash equilibrium of a single play of the trading game, exists in pure strategies.\(^{29}\) The key intuition is that if in period 1 there is some TF providing a single unit of liquidity at a hypothesized equilibrium spread, then, on the one hand, other TFs do not have incentive

\(^{27}\)If public and private information had different jump distributions, denoted \( J_{\text{public}} \) and \( J_{\text{private}} \), the right-hand side of (3.1) would be \( \lambda_{\text{public}} \cdot \Pr(J_{\text{public}} > \frac{s}{2}) \cdot E(J_{\text{public}} - \frac{s}{2} | J_{\text{public}} > \frac{s}{2}) + \lambda_{\text{private}} \cdot \Pr(J_{\text{private}} > \frac{s}{2}) \cdot E(J_{\text{private}} - \frac{s}{2} | J_{\text{private}} > \frac{s}{2}) \).

\(^{28}\)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 (2020) suggests that both fast and slow TFs providing liquidity that sometimes gets sniped are empirically relevant.

\(^{29}\)Equilibria in mixed strategies can exist when participants are able to provide liquidity at random prices (Baruch and Glosten, 2019). Such equilibria in our setting involve liquidity providers updating their quotes in *every* trading game. Since we examined repeated play of a trading game that is interpreted as lasting a short amount of time (e.g., 1 millisecond or shorter), and in the vast majority of milliseconds for real-world securities neither trades nor changes in the national best bid or offer occur (see footnote 12), we restrict attention to equilibria in pure strategies: as we will show, in such equilibria each exchange’s order book settles into a rest point after each period-2 arrival event (investor, informed trader or public information).
to offer additional liquidity at this spread (because they would suffer adverse selection and sniping without adequate compensation), but on the other hand, this leaves the TF who is providing liquidity incentive to deviate by widening their spread.

This non-existence result arises because of adverse selection. In the standard model of undifferentiated Bertrand competition without adverse selection, a Nash 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.

To address this non-existence issue, we introduce an alternative equilibrium solution concept, order book equilibrium, which strictly weakens MPE (or Nash equilibrium for a single play of the trading game). Whereas MPE requires that no players have profitable deviations, OBE allows for strictly profitable deviations to exist as long as they are rendered unprofitable by one of two specific reactions from rivals. By using anticipated reactions to counter otherwise profitable deviations, OBE captures the idea that each exchange’s limit order book settles into a rest point in which no trading firm wishes to add or remove any liquidity from any exchange’s order book, until the next arrival of an investor, informed trader, or public information. We provide the formal definition of OBE, as well as a detailed example that provides intuition for why it helps to restores equilibrium existence, in Appendix C.2.30

Equilibrium of the Trading Game. We can now formally characterize equilibrium of the Stage 3 trading game in this special case. Our solution concept is order book equilibrium for period 1 in anticipation of optimal play in period 2. Several aspects of equilibrium are unique:

Proposition 3.1. Consider the infinitely repeated Stage 3 subgame with a single exchange \((M = 1)\) charging zero trading fees \((f_1 = 0)\), and all \(N\) fast trading firms having purchased exchange-specific speed technology. Any equilibrium has the following properties. In period 1 of each trading game: in aggregate, a single unit of liquidity is provided (potentially by multiple trading firms) at bid-ask spread \(s^\text{continuous} \) (defined in (3.1)) around the current value of \(y\). In period 2 of each trading game: an investor, upon arrival, immediately transacts one unit at the best bid or offer; an informed trader,

---

30The two types of reactions that are allowed to discipline otherwise profitable deviations are (i) withdrawing liquidity that is no longer profitable to offer due to the deviation (e.g., it was just undercut), and (ii) providing new liquidity at a better price if profitable to do so, even if other liquidity is withdrawn (a “safe profitable price improvement”). The absence of profitable price improvements captures the spirit of “competitive liquidity provision” as assumed in Glosten and Milgrom (1985), but in a setting where fast TFs earn strictly positive profits from sniping. The concept also captures the spirit of “immediate responses” to deviations as assumed in the continuous-time model of BCS. Please see Appendix C.2 for further discussion. Our concept is also 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. 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.
upon arrival, immediately transacts one unit at the best bid or offer if their privately-observed jump in $y$ exceeds $s^*_\text{continuous}$; and if there is a publicly-observed jump in $y$ that exceeds $\frac{s^*_\text{continuous}}{2}$, there is a sniping race in which all fast trading firms attempt to trade against stale quotes provided by any trading firm other than themselves, and all fast trading firms providing liquidity attempt to cancel their stale quotes. Such an equilibrium exists.

(All proofs are contained in the appendix.)

3.2.2 Equilibrium of the Full Exchange Competition Game

We now turn to equilibrium analysis of the full exchange competition game. Our equilibrium concept is subgame perfect Nash equilibrium for Stages 1 and 2, order book equilibrium for Stage 3 period 1, and that participants play their essentially unique optimal strategies in Stage 3 period 2.

The main result of this Section is Proposition 3.2 (below), which establishes that there exist equilibria of the exchange competition game with the following key properties. First, all exchanges charge zero trading fees — i.e., trading fees are competitive. Second, exchanges charge positive ESST fees, and all fast TFs purchase ESST from all exchanges with positive market shares. These ESST fees are bounded above: i.e., exchanges cannot fully extract all latency arbitrage rents from fast TFs. Last, in each trading game, a single unit of liquidity is provided at spread $s^*_\text{continuous}$ across multiple exchanges, according to an arbitrary vector of market shares denoted $\sigma^*$. Investors and informed traders, when they arrive, route their orders across exchanges according to $\sigma^*$. That is, the exchange market share vector coordinates the liquidity provision actions of TFs and the routing decisions of investors and informed traders. In the event of a sniping race, it plays out in parallel across all $M$ exchanges, with all $N$ fast TFs racing on all $M$ exchanges. In essence, market participants use frictionless search and access to “synthesize” a single exchange from the $M$ parallel exchanges, and then act economically the same way as in Proposition 3.1.

Proposition 3.2. Consider the full exchange competition game with $M \geq 2$ exchanges. For any vector of market shares $\sigma^* = (\sigma^*_1, \ldots, \sigma^*_M : \sum_j \sigma^*_j = 1)$, and for any vector of exchange-specific speed technology (ESST) fees $F^* = (F^*_1, \ldots, F^*_M)$ that satisfies the condition given by (3.2) below, there exists an equilibrium where:

(Stage 1): Each exchange $j$ charges $F^*_j$ for ESST, and charges zero trading fees ($f^*_j = 0$);

(Stage 2): All $N$ fast trading firms purchase ESST from every exchange $j$ where $\sigma^*_j > 0$;

(Stage 3): The following occurs in every iteration of the trading game given state $(y, \omega)$. At the end of period 1, $\sigma^*_j$ quantity of liquidity is provided on each exchange $j$ at spread $s^*_\text{continuous}$ (defined in (3.1)) around $y$. In period 2: an investor, upon arrival, immediately transacts $\sigma^*_j$ at the best bid or offer on each exchange $j$; an informed trader, upon arrival, immediately transacts $\sigma^*_j$ at the best bid or offer on each exchange $j$ if their privately-observed jump in $y$ exceeds $\frac{s^*_\text{continuous}}{2}$; and if there is a publicly-observed jump that exceeds $\frac{s^*_\text{continuous}}{2}$, 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.

The condition on ESST fees is:

\[
\frac{\Pi^*_{\text{continuous}}}{N} - \sum_{j: \sigma_j^* > 0} F_j^* \geq \max(0, \pi_{N_{\text{lone-wolf}}} - \min_j F_j^*),
\]

where \(\Pi^*_{\text{continuous}} \equiv \lambda_{\text{public}} \cdot L(s_{\text{continuous}}^*)\) denotes the total “sniping prize,” and \(\pi_{N_{\text{lone-wolf}}}\) is a constant discussed below and defined in Appendix C.3.2, equation (C.3).

The proof of this result is constructive. We first establish our “virtual single platform” result: if all \(N\) fast trading firms purchase ESST from the same set of exchanges and all exchanges set zero trading fees, then any order book equilibrium of the multi-exchange trading game replicates the outcome of the single-exchange trading game described in Proposition 3.1 across multiple exchanges (Lemma C.1). Economically, the key feature of equilibrium of Stage 3 is that the marginal unit of liquidity is equally profitable across all exchanges, because each exchange's share of liquidity provided (“depth”) matches its share of volume from investors. For any arbitrary split of market shares \(\sigma^* = (\sigma_1^*, \ldots, \sigma_M^*)\), there is an equilibrium in which each exchange \(j\)'s share of depth and volume are each exactly \(\sigma_j^*\).

We next examine behavior in Stage 2, and prove that if each exchange \(j\) charges \(F_j^*\) for ESST fees and zero for trading fees, it is an equilibrium for all fast TFs to purchase ESST from all exchanges as long as condition (3.2) is satisfied and all exchanges charge zero trading fees. If all fast TFs purchase ESST from all exchanges, in any order book equilibrium of the subsequent trading game, each fast TF obtains (in expectation, gross of ESST fees) their share of the sniping prize, \(\Pi^*_{\text{continuous}}/N\). We analyze a specific deviation for fast TFs (which we refer to as a lone-wolf deviation), and show that it is the most attractive deviation to consider, hence ruling it out is sufficient. In a lone-wolf deviation, instead of purchasing ESST from all exchanges, 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 that is strictly narrower than \(s_{\text{continuous}}^*\) (which we prove to be an equilibrium of the Stage 3 subgame that follows; Lemma C.2). We prove that condition (3.2) in the statement of Proposition 3.2 ensures that this lone wolf deviation is not profitable, as each fast TF earns more in expectation by purchasing ESST from all exchanges and earning \(\Pi^*_{\text{continuous}}/N\) per trading game than purchasing ESST from just a single exchange and earning deviation profits of \(\pi_{N_{\text{lone-wolf}}}\) per trading game. Last, we examine Stage 1 and show that there is an equilibrium of the full game in which exchanges all charge zero trading fees and charge ESST fees that satisfy condition (3.2). Given equilibrium strategies in Stages 2 and 3, exchanges cannot profitably adjust their ESST fees, and any exchange that raises its trading fee from zero gets zero share.

### 3.3 Discussion

We now discuss the main features of the equilibria described in Proposition 3.2.
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. Specifically, all equilibria described in Proposition 3.2 share the following three features. First, in every trading game, all exchanges with positive depth have the same bid-ask spread \( s^\text{continuous} \), 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.

The key intuition behind these results is that, as long as depth and volume are equivalent across all exchanges, equation (3.1) which characterizes the equilibrium benefits and costs of providing liquidity, and hence the equilibrium bid-ask spread, 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.\(^{31}\)

We acknowledge that 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 3.2, investors break ties when indifferent across exchanges using what we refer to as routing table strategies (see Appendix C.3.4). 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.

Competitive Trading Fees. In the equilibria described in Proposition 3.2, trading fees are competitive and equal to 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 by market participants. 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. In a supporting Lemma for Proposition 3.2, 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 C.1 in Appendix C.3).

Money-Pump Constraint. In our model, exchanges may appear to lack an obvious source of market power: they are symmetric and undifferentiated, search is frictionless, and market participants

\(^{31}\)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. He shows that multiple exchanges can coexist in equilibrium if their liquidity schedules add up to what would have been provided on a single exchange. 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.
can costlessly participate on any exchange. Since add-on rents in competitive pricing models are often dissipated in competition to sell the pre-add-on good (cf. Ellison, 2005; Gabaix and Laibson, 2006), one might expect that exchanges would compete away any 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. However, this is not the case here. In the equilibria constructed in Proposition 3.2, exchanges are able to earn and maintain positive profits due to a binding money-pump constraint. 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 — i.e., a money-pump.\(^2\) In Appendix C.3.3, we show that an exchange’s losses from negative trading fees can be arbitrarily large without TFs engaging in any self-dealing.

ESST Fees and the Division of Latency Arbitrage Rents. 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). Using the analysis behind the lone-wolf bound, we are able to show that the proportion of sniping rents that exchanges must leave for TFs is economically significant:

**Proposition 3.3.** In the equilibria described by Proposition 3.2, exchanges’ total rents from ESST fees, \(N \times \sum_{j, \sigma_j > 0} F_j^*\), are strictly less than \(\frac{M}{(M-1)(N-1)} \Pi_{\text{continuous}}^*\).

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 Stylized Fact #3 in Section 4.1). Aquilina, Budish and O’Neill (2020) 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.\(^3\) Proposition 3.3 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.

We emphasize that while this particular division of latency arbitrage rents is specific to our model, what will ultimately matter for the analysis of market design innovation considered in Section 5 is simply that exchanges are able to capture and maintain some share of the rents generated from latency arbitrage activity in the status quo.\(^4\) A strength of our approach is that it highlights that even if

---

\(^{2}\)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.

\(^{3}\)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.

\(^{4}\)Other potential modeling frameworks for understanding the division of rents between TFs and exchanges include non-cooperative bargaining games and cooperative solution concepts for rent-splitting such as the Shapley value. Roth and Wilson (2019) discuss the complementary role non-cooperative and cooperative game theory can play in applied market design research. Potential non-cooperative bargaining games include the “Nash-in-Nash” solution for bilateral oligopoly in industrial organization settings (Collard-Wexler, Gowrisankaran and Lee, 2019).
exchanges can post prices — which, in many bargaining models, is akin to maximum bargaining power — they cannot extract all of the surplus.

Sources of Deadweight Loss. In our model, there are $N$ trading firms exogenously endowed with general-purpose speed technology, and $M$ exchanges exogenously present in the market and able to sell exchange-specific speed technology to TFs. TFs’ payments to the exchanges for this speed technology, represented by the $F^*_j$’s in our model, 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.

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

4 Stylized Empirical Facts

In this section we document a series of seven stylized facts regarding modern stock exchange competition. These facts relate to each of the three main results of Section 3’s model of the status quo: the virtual single platform in the Stage 3 trading game (Section 4.1), exchange trading fees (Section 4.2), and exchange-specific speed technology fees (Section 4.3). We also provide further details and robustness tests for each set of facts in Appendix D, E, and F. Section 4.4 provides discussion of the stylized facts with reference to other previous models of financial exchange competition.

4.1 Evidence on the Trading Game and Virtual Single Platform Theory

There are three main features of the multi-exchange trading game equilibria, discussed in Section 3.3, that we will assess empirically. First, all active exchanges have the same equilibrium bid and offer, i.e., quoted prices are identical across exchanges. Second, each exchange’s share of market depth (i.e., its share of liquidity) at this common best bid and offer equals its share of market volume. Third, exchange depth and volume shares can be interior and stable, i.e., there need not be tipping.

Data. We use the Daily NYSE Trade and Quote (“TAQ”) dataset. The data contain every trade and every top-of-book quote update for every exchange, for all U.S. listed stocks and exchange-traded funds (ETFs), timestamped to the millisecond. The key advantage of this data, for our purposes, is that it is comprehensive across exchanges and labels every trade and quote update by exchange.

For the results presented in this section, we make three types of sample restrictions. First, we
use data from all trading days in 2015.\textsuperscript{35} Second, we focus primarily on the top 5 exchanges by market share which together constitute 83\% of trading volume. The top 5 exchanges all utilize what is commonly referred to as the “maker-taker” pricing model in which the taker of liquidity is charged a fee and the provider of liquidity is paid a rebate. The next 3 exchanges, which constitute 15\% of trading volume, all use the “taker-maker” pricing model in which the taker is paid a rebate and the maker pays a fee. This difference in fee structure raises some subtleties for the analysis that we discuss in Appendix D. Third, we restrict attention to the 100 most heavily traded stocks and ETFs. These 100 symbols constitute about one-third of daily volume.\textsuperscript{36}

**Stylized Fact #1 (Many Exchanges at the Best Bid and Best Offer):** At any given moment in time, for highly traded stocks and ETFs, the modal number of exchanges at the best bid and best offer is “all of them.” Of the Top 5 exchanges, in about 85\% of milliseconds all exchanges are at the best bid (similarly, best offer). It is rare for there to be just one exchange at the best bid or best offer.

Figure 4.1 presents the distribution of the number of exchanges at the best price at a given moment in time. For each symbol-millisecond in our sample, we compute the number of exchanges that are at the best bid or ask. We then compute the percent of time there were a given number of exchanges out of the top 5 that were at the best price, aggregating over all symbols, millisecond, and sides (bid/ask). We present the results separately for NYSE-listed symbols and non-NYSE listed symbols since, at the time of our data, non-NYSE listed symbols did not trade on NYSE (but did trade everywhere else), whereas NYSE listed symbols traded everywhere. Hence, for NYSE listed symbols the maximum number of exchanges out of the Top 5 that could be at the best bid or offer is 5, whereas for non-NYSE listed symbols the maximum is 4. For NYSE-listed symbols, all Top 5 exchanges are at the best bid or offer in 86.1\% of milliseconds, and for non-NYSE symbols all 4 exchanges are at the best bid or offer in 84.6\% of milliseconds.

**Stylized Fact #2 (Depth-Volume Relationship):** Among the Top 5 exchanges there is a one-for-one relationship between depth share and volume share. The coefficient from regressing daily volume share on daily depth share is 0.99 (statistically indistinguishable from 1) and the $R^2$ is 0.87.

Figure 4.2 presents a scatterplot of the volume share against depth share for each symbol-date-exchange in our sample. Volume share is an exchange’s fraction of regular-hours volume for each symbol-date-exchange. Depth share is an exchange’s fraction of shares available at the best quote for a given symbol-date, averaged over all milliseconds in the day on both the bid and ask. The figure shows that the depth-volume data fall along the 45 degree line for the Top 5 exchanges. The slope of a regression

\textsuperscript{35}2015 is the best year in terms of data availability for the analysis of ESST revenues (see Section 4.3 and Appendix F).

\textsuperscript{36}We also require that the symbols in our sample satisfy a set of data-cleaning filters: trading continuously throughout the year under the same ticker, having a share price of at least $1, not having a listing change, and having at least $10 million in average daily volume. We have also conducted robustness tests in which we look at the top 1000 symbols, which constitute roughly three-quarters of total volume. Results are qualitatively similar but with more noise.
Notes: The data is from NYSE TAQ. Percent of time indicates the percent of symbol-side-milliseconds (e.g., SPY-Bid-10:00:00.001) for which the number of exchanges at the best bid or offer was equal to $N$. An exchange was at the best price for a symbol-side-millisecond if the best displayed quote on that exchange was equal to the best displayed quote on any of the Top 5 exchanges, all measured at the end of the millisecond. The best bid or offer on the Top 5 exchanges was also the best bid or offer across the Top 8 exchanges in over 99.9% of milliseconds; see Appendix D for details. Sample is 100 highest volume symbols that satisfy data-cleaning filters on all dates in 2015.

37 The depth-volume relationship also holds at the individual symbol level. Regressing daily exchange market shares on daily exchange depth shares for each symbol yields an average regression coefficient close to one (mean 0.991, st. dev. 0.026) and a high average $R^2$ (mean 0.840, st. dev. 0.130).

38 Our model assumes all investors demand exactly “1” unit of perfectly-divisible liquidity and in equilibrium exactly 1 unit of liquidity is offered across exchanges so investors must spread their demand across multiple exchanges. In reality, investors demand varying amounts of liquidity. Investors who only wish to trade a small amount (e.g., 100 shares) often do so with a single small trade on a single exchange. Investors who wish to trade a larger amount often break their total desired quantity into smaller individual orders spread out over time. So, volume shares at the trade-by-trade level are often 100% for a single exchange and 0% for all others. However, the logic of our model suggests that, at a higher level of aggregation, volume shares should match depth shares — else, the marginal unit of liquidity will be too adversely selected on some exchanges and too favorable on others.

39 We start the time period in 2011 since that is the first full year of data after the original BATS exchanges (BZX...
Notes: The data is from NYSE TAQ. The dark line depicts the 45-degree line which is the depth share to volume share relationship predicted by the theory. The results are presented for the Top 5 maker-taker exchanges, and includes the 100 highest volume symbols that satisfy data-cleaning filters on all dates in 2015. Observations are symbol-date-exchange shares, with shares calculated among the Top 5 exchanges.

interior, with no exchange’s market share ever rising above 30%. Additionally, these market shares are relatively stable in the sense that in the 2011-2015 period, if we regress the market share of each exchange-date on only a set of exchange fixed effects, the $R^2$ is 0.967.\textsuperscript{40}

4.2 Evidence on Exchange Trading Fees

We now examine the two predictions of our theoretical model regarding exchange trading fees. First, trading fees are competed down to zero (i.e., fees are perfectly competitive), and second, fees are bounded below by a money-pump constraint.

\textbf{Data.} We use two types of data sources for our analysis of exchange trading fees. First, we use historical fee schedules from exchange websites retrieved using the Internet Archive. Second, we use exchange company financial filings such as 10-K’s and S-1’s. We focus on 2015 for consistency with the other analyses.\textsuperscript{41} It is important to clarify that exchange companies each control several exchanges, and BYX) and Direct Edge exchanges (EDGX and EDGA) were approved as exchanges (prior to that they operated Alternative Trading Systems, or ATS’s). Appendix D presents the same figure as above but starting from October 2007, the start of the Reg NMS era.

\textsuperscript{40}In Appendix D, we present related results at the individual-symbol level. As at the exchange level, shares are interior and relatively stable.

\textsuperscript{41}The specific historical fee schedules we use range from Feb 2015 to Sept 2015 depending on the Internet Archive’s coverage. The specific financial filings we use are the BATS April 2016 S-1 filing, Nasdaq’s fiscal year 2015 10-K report, Intercontinental Exchange’s (NYSE’s parent) fiscal year 2015 10-K report, and NYSE’s fiscal year 2012 10-K filing (2012 was its last full fiscal year as a stand-alone company).
Figure 4.3: Exchange Market Shares, 2011 - 2015

Notes: The data is from NYSE TAQ and covers January 2011 to Dec 2015 for the Top 8 exchanges. The market shares are based on all on-exchange trading volume in shares.

and while the historical fee schedules are at the exchange level, most of the financial data in the annual report is at the exchange company level.

Stylized Fact #4 (Trading Fees are Economically Small): Exchange trading fees are economically small. The observed range of regular hours trading fees is, on a per-share per-side basis, -0.00015 to 0.00080 on the Top 8 exchanges. The average per-share per-side fee paid for regular hours trading is about 0.0001. For a $100 share of stock, the fee in percentage terms is 0.0001%.

Our theoretical model says that exchange trading fees, denoted $f$ in the model, will be perfectly competitive and bounded below by a money-pump constraint. In practice, however, there is no single number to look up that represents “$f$.” We take two approaches to provide a sense of $f$ in the data. First, we report the range of trading fees for the top 8 exchanges from historical fee schedules. Second, we use exchange family financial filings to compute average trading fees, i.e., average $f$.

Table 4.1 Panel A presents the observed range for per-share per-side regular-hours trading fees for the top 8 exchanges. As can be seen, many of the exchanges have minimum fees on a per-share per-side basis that are actually slightly negative. Appendix Table E.1 provides a more complete table, which takes into account additional details for special fee programs, and shows that 7 of the 8 exchanges have a minimum per-share per-side fee that is negative. The maximum fee per-side is always strictly positive and typically about $0.0005. Roughly, low-volume market participants pay the maximum fee and the highest-volume market participants pay the minimum fee; see the Appendix for details.

Table 4.1 Panel B provides an estimate for average per-share per-side trading fees at the exchange family level using annual financial filings. The average trading fee across the 3 major exchange families is about 0.0001 per-share per-side.\footnote{Please see Appendix E.2 for full details regarding these calculations.} For a $100 share of stock, the fee in percentage terms is 0.0001%.
<table>
<thead>
<tr>
<th>Exchange</th>
<th>Fee Type</th>
<th>Taker Fee</th>
<th>Maker Fee</th>
<th>Total fee per share per side</th>
</tr>
</thead>
<tbody>
<tr>
<td>NASDAQ</td>
<td>Maker-Taker</td>
<td>0.00300</td>
<td>-0.00305</td>
<td>0.00003</td>
</tr>
<tr>
<td>BATS BZX</td>
<td>Maker-Taker</td>
<td>0.00300</td>
<td>-0.00320</td>
<td>0.00010</td>
</tr>
<tr>
<td>EDGX</td>
<td>Maker-Taker</td>
<td>0.00300</td>
<td>-0.00350</td>
<td>0.00025</td>
</tr>
<tr>
<td>NYSE</td>
<td>Maker-Taker</td>
<td>0.00270</td>
<td>-0.00220</td>
<td>0.00025</td>
</tr>
<tr>
<td>NYSE Arca</td>
<td>Maker-Taker</td>
<td>0.00280</td>
<td>-0.00300</td>
<td>-0.00010</td>
</tr>
<tr>
<td>BATS BYX</td>
<td>Taker-Maker</td>
<td>-0.00160</td>
<td>0.00140</td>
<td>-0.00010</td>
</tr>
<tr>
<td>EDGA</td>
<td>Taker-Maker</td>
<td>-0.00020</td>
<td>0.00030</td>
<td>0.00005</td>
</tr>
<tr>
<td>NASDAQ BX</td>
<td>Taker-Maker</td>
<td>-0.00150</td>
<td>0.00165</td>
<td>0.00008</td>
</tr>
</tbody>
</table>

Panel B: Estimate of Average Trading Fees

<table>
<thead>
<tr>
<th>Exchange Group</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>BATS</td>
<td>$0.000089</td>
</tr>
<tr>
<td>NASDAQ</td>
<td>$0.000105</td>
</tr>
<tr>
<td>NYSE</td>
<td>$0.000128</td>
</tr>
</tbody>
</table>

Notes: Panel A summarizes the fee schedules for the top 8 exchanges retrieved from the Internet Archive (Wayback Machine) dated from February 28, 2015 to September 1, 2015. In general, we determine the max rebates based on what a trading firm that satisfies the exchange’s highest volume tier would pay or receive, and the min rebates and fees tend to be the baseline for adding or taking liquidity. We omit fees associated with special programs or differences based on tape plans. Please see Appendix E.1 for a complete table of estimated fees for both regular and special programs and for tape A, B, and C stocks. Panel B shows the average trading fee for each of the three major exchange families estimated from financial filings. Please see Appendix E.2 and the associated spreadsheet for supporting details for these calculations.

While not zero, this figure is arguably economically small. Across the approximately 1 trillion shares traded during regular hours each year, this adds up to about $200M. As a point of comparison, the operating expenses for BATS’s U.S. equities business alone were $110M in 2015 — and BATS is generally viewed as more cost-effectively run than Nasdaq or NYSE (each have about a third of regular-hours volume). NYSE’s operating expenses for its equities business in 2012, its last full-year of operation before the ICE acquisition, were $718M. In other words, regular-hours trading revenues do not nearly cover exchange operating expenses.

Stylized Fact #5 (Money-Pump Constraint Binds): Exchange trading fees for high-volume traders are often slightly negative on a per-share per-side basis. For 4 of the top 8 exchanges the fee is negative for the highest volume tier. For another 3 of the 8 exchanges, the fee is negative for traders with high-enough volume who satisfy additional requirements. However, trading fees do not get negative enough to create a money pump once we account for SEC + FINRA fees.

In the language of our model, exchanges are in principle willing to lose money on $f$ (trading fees)
in order to make more money from $F$ (speed technology). However, trading fees are bounded below by a money-pump constraint: if $f < 0$ on some exchange, trading firms would engage in infinite volume and extract infinite dollars. In practice, the money-pump boundary is below zero, because of SEC Section 31 fees and FINRA fees. These fees increase with the nominal share price: for example, the per-share per-side fees are $0.00017 for a $10 stock and $0.00060 for a $50 stock.\(^{43}\)

Table 4.1 shows that the per-share per-side exchange trading fee is negative for Nasdaq, BZX, EDGX, and BYX for the highest volume tier, with the lowest observed fee being -$0.00015. For another 3 of the 8 exchanges (NYSE, NYSE Arca, Nasdaq BX), the fee is negative for traders with high-enough volume who satisfy additional requirements, with the lowest observed such fee being -$0.00040 per-share per-side (see Table E.1 in the Appendix). These negative fees are consistent with exchanges being willing to lose money on trading fees ($f$) to make money on exchange-specific speed technology fees ($F$). However, trading fees are not negative enough to create a money pump once we account for SEC and FINRA fees, with the exception of very-low priced stocks.

### 4.3 Evidence on Exchange-Specific Speed Technology Revenue

The last set of stylized facts relates to our theoretical prediction about exchange-specific speed technology (ESST) revenues. Our model shows that exchanges can earn supra-competitive rents from ESST in equilibrium. The intuition is that exchanges have market power over speed technology that is specific to their exchange, e.g., only Nasdaq can sell the right to co-locate next to Nasdaq’s servers. Notably, our model does not pin down the exact level of ESST, but does indicate that total ESST revenue cannot be too large of a fraction of the total sniping pie (see Proposition 3.3).

**Data.** Our evidence on the magnitude and growth of ESST revenues comes from exchange company financial filings (10-K’s, S-1’s, and merger proxies). We also use a Consolidated Tape Association fee filing to get an estimate for aggregate tape revenues (revenues that come from a data feed not used by latency sensitive traders).

**Stylized Fact #6 (Exchanges Earn Significant Revenues from ESST):** Using exchange parent company financial filings and applying sensible assumptions, the data suggest that in 2015 total ESST revenue was between $675-790M. This is several times larger than regular-hours trading revenues.

Table 4.2 provides estimates for U.S. market data and co-location/connectivity revenue associated with each of the three major exchange families. BATS’s figures come directly from financial filings.\(^{44}\) NYSE and Nasdaq do not separately report ESST revenue for their U.S. equities businesses. Thus, we

---

\(^{43}\)At the time of our data, the SEC fee was $21.80 per $1M traded and the FINRA fee was $0.000119 per share traded. Both fees are assessed on sales but not purchases, i.e., they are assessed on one side of each transaction. For the purpose of calculating the money-pump boundary, we look at the SEC + FINRA fees on a per-share per-side basis because an exploiter of a money pump would need to both buy and sell.

\(^{44}\)BATS’s April 2016 IPO filing (i.e., form S-1) provides an unusually clear window into how exchange revenues break down across trading revenue, market data, and co-location/connectivity. BATS was acquired by CBOE later in 2016 and following that acquisition no longer reported their revenues with such granularity. Neither Nasdaq nor NYSE have ever reported their U.S. equities revenue with the granularity of BATS’s IPO filing.
Table 4.2: Estimated Market Data and Co-Location Revenues for U.S. Equities Market in 2015
(Millions of Dollars)

<table>
<thead>
<tr>
<th></th>
<th>BATS</th>
<th>NASDAQ</th>
<th>NYSE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market Data Revenue</td>
<td>114.1</td>
<td>222.4</td>
<td>–</td>
<td>267.3</td>
</tr>
<tr>
<td>Co-Location/Connectivity Revenue</td>
<td>64.3</td>
<td>121.0</td>
<td>–</td>
<td>139.0</td>
</tr>
<tr>
<td>Market Data + Co-Location Revenue</td>
<td>178.4</td>
<td>343.3</td>
<td>–</td>
<td>406.4</td>
</tr>
<tr>
<td>CTA/UTP Tape Revenue</td>
<td>317.0</td>
<td>317.0</td>
<td>–</td>
<td>317.0</td>
</tr>
<tr>
<td>Market Data + Co-Lo Revenue net of Tape Revenue</td>
<td>675.2 – 790.8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: BATS data is from its April 2016 S-1 filing, which contains data up through the end of 2015. Nasdaq data is from its 2015 10-K filing. NYSE data uses both ICE’s 2014 and 2015 10-K filings. BATS directly reports a U.S. equities revenue breakdown. For Nasdaq and NYSE some assumptions are needed to estimate U.S. equities revenue from the market data and co-location/connectivity revenue items they report; therefore we report a range of estimates. For full details see Appendix F.1. The CTA/UTP tape revenue number is obtained from a CTA fee-change filing to the SEC, in which they report the total CTA/UTP market data revenue (allocated to exchanges) annualized through March of 2014.

adjust revenue figures reported in filings by making some assumptions about revenue not associated with U.S. equities and report a range. We provide a detailed description of our calculations for Nasdaq and NYSE in Appendix F.

Across all three major exchange families, we estimate $555-$623M for market data revenue and $436-$485M for co-location/connectivity revenue. The market data revenue figures include revenue from exchanges’ proprietary data feeds as well as from market-wide “Tape Plans,” sometimes known as the SIP feed. Proprietary data feeds are utilized by latency-sensitive market participants, whereas the market-wide SIP feed is not as fast, and therefore should be deducted from our estimate of overall ESST revenues. If we subtract the $317M in tape revenue reported by the Consolidated Tape Authority (CTA) from the total, we have proprietary market data revenue of $238.4-306.0M, and total ESST revenue of $675.2-790.8M.\(^{45}\)

For context, note that our estimate of 2015 ESST revenue is roughly 3 to 4 times larger than the estimated revenue from regular-hours trading fees of $200M as reported in Stylized Fact 4 above. If we take the lone-wolf bound from the theory seriously, and assume that exchanges extract at most 30% of latency arbitrage rents (see Section 3.3), our estimated range of ESST revenues yields a lower bound on the total size of the latency-arbitrage pie of $2.25 billion in 2015.

Stylized Fact #7 (Exchange Revenue from ESST has Grown Significantly in the Reg NMS Era): Exchanges’ revenues from exchange-specific speed technology have grown significantly in the Reg NMS era. We compute annual growth rates of: 15.9% for Nasdaq co-location/connectivity (2006-2017),

\(^{45}\)For related empirical evidence, see also a recent paper of Jones (2018) commissioned by the New York Stock Exchange. While our interpretation is different, our numbers are mostly consistent with those documented in Jones (2018). One important exception is that Jones (2018) considers exchange trading revenues gross of exchange rebates rather than net of exchange rebates.
10.5% for Nasdaq proprietary market data (2006-2017), and 40.4% for BATS co-location/connectivity (2010-2017). If we utilize 10% as a conservative overall growth rate since 2015, this implies annual ESST revenue in 2018 on the order of $1 billion per year.

We can get a sense of magnitudes for U.S. equities ESST revenue growth over time by examining the financial reporting categories that contain ESST revenue. As in Stylized Fact #6, we make some assumptions to isolate revenue from U.S. equities. We are able to build meaningful time-series for Nasdaq co-location/connectivity revenues and proprietary market data revenues from 2006 to 2017 and for BATS co-location and connectivity revenues from 2010 to 2017. Complete details of the data and methodology used are provided in Appendix F.2.

Figure 4.4 presents ESST revenue growth for the three series we construct. Overall the data, while imperfect, are suggestive of exchanges discovering that they could charge significant money in the Reg NMS era for something they used to not charge for. If we use 10% as a conservative overall growth rate for ESST revenue since 2015, and apply this growth rate to our estimates from Stylized Fact #6, this implies that 2018 ESST revenues are between $899M-$1,053M.

4.4 Discussion of Alternative Models of Exchange Competition

There are several other models of exchange competition that are inconsistent with aspects of the data. It is important to note that many of these models are designed to study other significant aspects of

---

46BATS only began charging for the proprietary market data that we think of as part of ESST relatively late in our sample period (Q3 2014) and we discuss the limited data that is available in the appendix. NYSE’s financial reporting segments unfortunately changed too frequently in the Reg NMS era for the exercise to be instructive.
exchange competition and not specifically the modern U.S. stock market.

First, there is a class of models in which some market participants “single home,” thereby generating exchange-specific network effects (Pagano, 1989; Cantillon and Yin, 2008; Pagnotta and Philippon, 2018; Cespa and Vives, 2019). In all of these models, exchanges charge supra-competitive fees in equilibrium (exploiting network effects), which stands in contrast to Stylized Facts #4-#5. Furthermore, in many of these models, these exchange-specific network effects often lead to tipping which stands in contrast to Stylized Facts #1-#3.

Another class of models are those where exchanges are meaningfully differentiated (Pagnotta and Philippon, 2018; Baldauf and Mollner, 2019), which allows exchanges to charge supra-competitive trading fees and is inconsistent with Stylized Facts #4-#5. Also, such models suggest that there may be segmentation of market participants and securities across venues, which is at odds with Stylized Facts #1-#3. Lastly, Chao, Yao and Ye (2019) provide a model in which tick-size frictions are central to understanding exchange fragmentation and competition. Their model is inconsistent with the fact that the Top 5 exchanges, which control 83% of volume, all use essentially the same fee structure, as we show in Stylized Fact #4.

5 Market Design Innovation: Will the Market Fix the Market?

In Section 3, we introduced a theoretical model of competition among multiple continuous limit order book exchanges (the status quo) and proved that there exist equilibria with the following key features: many exchanges maintain positive market shares, with liquidity at the same bid-ask spread and with trading firms indifferent at the margin across exchanges due to the depth-volume relationship; exchange trading fees are competitive and bounded below by the money-pump constraint; and exchanges capture and maintain economic rents via supra-competitive fees for exchange-specific speed technology (ESST), which trading firms need to purchase to participate in speed-sensitive trading. In Section 4, we established that this model does a reasonable job empirically, documenting stylized facts that correspond to each of the model’s main results.

In this section, we use our theoretical model to examine exchanges’ incentives for market design innovation. Our discussion will focus on frequent batch auctions (FBAs) with a very short batch interval as the specific market design alternative to the continuous limit order book, though our analysis applies equally to the asymmetric delay market design with a very short delay interval.\footnote{In “asymmetric speed bump” or “asymmetric delay” market designs, an exchange processes cancellations immediately upon receipt but processes marketable orders only after a fixed small delay. This market design also eliminates latency arbitrage in the BCS model, and captures one aspect of FBAs in that orders can be canceled at any time while executions can only occur with some delay (i.e., at the end of the batch interval), but it does have some weaknesses relative to FBAs that are outside the model. Specifically, because it serially processes new orders, there still is a race to the top of the book, and there still can be sniping races if there are stale limit orders provided by market participants who are not fast enough to update within the delay window. Recent evidence in Aquilina, Budish and O’Neill (2020) suggests that stale quotes taken in races are supplied by firms outside of the fastest HFTs more than 50% of the time. Such orders would be vulnerable to sniping in an asymmetric delay market if not cancelled within the delay window, but would trade at a price that reflects new public information in a FBA market. See Section VIII.C-D of Budish, Cramton and Shim (2015)}
Formally, for our theoretical analysis, we assume that the alternative market design eliminates rents from symmetric public information but does not have any additional benefits or costs.

Section 5.1 presents modeling details. Section 5.2 analyzes equilibrium of our exchange competition model if there is a single FBA exchange and one or more continuous limit order book exchanges. Section 5.3 analyzes equilibrium if there are multiple FBA exchanges and one or more continuous exchanges. Sections 5.4-5.5 analyze what the equilibria for these different configurations of market designs implies about exchanges’ private innovation incentives.

5.1 Modeling Frequent Batch Auctions (FBAs)

The FBA market design, proposed and analyzed in Budish, Cramton and Shim (2015), is similar to the continuous limit order book market design in many key respects. In both market designs: (i) orders consist of a price, side, and quantity; (ii) orders can be submitted, modified or canceled at any moment in time; (iii) orders remain outstanding until either executed or canceled; (iv) priority, if necessary to break ties, is based on price then time; and (v) information policy is that orders are received by the exchange, economically processed by the exchange, and then the updated economic state is disseminated publicly.

There are two key differences. First, FBAs divide the trading day into frequent pre-specified discrete-time intervals, and treat all orders received in the same interval as having been received at the same time. A way to think about this is that time is treated as a discrete variable rather than as a continuous variable. Second, orders are batch processed at the end of each discrete-time interval, using a uniform-price auction, rather than being serially processed upon receipt as in a limit order book.  

The exchange then publicly announces (i) any trades that occurred (quantities and prices, just as in the continuous market), and (ii) the updated state of the order book, i.e., any orders that remain outstanding (just as in the continuous market).

Formal Model Details. We make the following adjustments to the exchange competition game introduced in Section 3 to allow for the possibility that exchanges use either the continuous limit order book (“Continuous”) or FBA (“Discrete”) market design.

First, in Stage 1 of our game, we assume that a Discrete exchange does not sell exchange-specific speed technology. Practically, we have in mind that an FBA exchange would allow market participants to co-locate their servers and subscribe to proprietary market data, but would not be able to charge

---

48 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. In case there is an interval of market-clearing prices the midpoint of this interval is utilized; this case is not relevant for our analysis. If necessary to break ties 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.

---
prices commensurate with their role, on continuous exchanges, in extracting sniping rents.\textsuperscript{49} Second, we modify the Stage 3 trading game as follows. A Discrete exchange first processes all cancellations received in a period of the trading game (reflecting that in an FBA orders can be canceled at any moment in time), and then processes any new limit or IOC orders received in that period, along with outstanding orders from previous periods, using a uniform-price auction as described above, with price then discrete-time priority used to break any ties. Note that, unlike on a Continuous exchange, TF speed does not affect the order in which messages are processed. Everything else about the Stage 3 trading game is the same as in Section 3.

### 5.2 A Discrete and a Continuous Exchange

We first examine a single Discrete exchange competing against a single Continuous exchange; the case of a single Discrete exchange competing against multiple Continuous exchanges will be economically equivalent.\textsuperscript{50} Recall that if there was only a single Continuous exchange in operation charging zero trading fees (see Section 3.2.1), a single unit of liquidity would be provided in equilibrium each trading game by fast trading firms at a spread $s^*_\text{continuous}$ given by (3.1): $\lambda_{\text{invest}} s^*_\text{continuous} = (\lambda_{\text{public}} + \lambda_{\text{private}}) \cdot L(s^*_\text{continuous})$. In contrast, if there was only a single Discrete exchange also charging zero trading fees, arguments developed in BCS imply that a single unit of liquidity would be provided in equilibrium at the spread $s^*_\text{discrete}$ which solves:

$$\lambda_{\text{invest}} \frac{s^*_\text{discrete}}{2} = \lambda_{\text{private}} \cdot L(s^*_\text{discrete}).$$

(5.1)

The difference between (5.1) and (3.1) is the $\lambda_{\text{public}} L(s^*)$ term missing from the equation defining $s^*_\text{discrete}$: this reflects that Discrete eliminates latency arbitrage rents, and hence the associated cost for liquidity providers. For this reason, $s^*_\text{discrete} < s^*_\text{continuous}$.

Now consider the multiple exchange trading game between a Continuous and a Discrete exchange. Suppose initially that trading fees on both exchanges are zero, and all TFs have purchased ESST from the Continuous exchange. A reasonable prior might be that there are multiple equilibrium outcomes: for example, there might be an equilibrium where all liquidity is provided and taken from Continuous, and another where all liquidity is provided and taken from Discrete. However, this is not the case:

**Proposition 5.1.** Consider the infinitely repeated Stage 3 subgame with a single Continuous and a single Discrete exchange, assuming that in Stage 1 both exchanges set trading fees to zero and in Stage 2 all fast trading firms have purchased exchange-specific speed technology from Continuous. Any

\textsuperscript{49}For example, as of a few years ago Nasdaq offered four different 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, 2015b). An FBA 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.

\textsuperscript{50}As discussed in Section 3.1, our theoretical analysis has shown that frictionless search and access enable multiple Continuous exchanges to operate as if they were a virtual single platform. It will become clear from the equilibrium that it makes no difference whether there is a single Continuous exchange or multiple Continuous exchanges that operate as a virtual single platform.
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 \( s^*_{\text{discrete}} \) (defined in (5.1)) around the current value of \( y \), and no liquidity is provided on the 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 \( s^*_{\text{discrete}} \); if there is a publicly-observed jump in \( y \) that exceeds \( s^*_{\text{discrete}} \), 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 the Continuous exchange in any equilibrium. To understand why, note that if a trading firm was to provide liquidity on Continuous and not lose money, it would have to charge at least a “zero-variable profit spread” on Continuous, denoted \( \tilde{s}_{\text{continuous}} \), which is strictly greater than \( s^*_{\text{discrete}} \). As a result, since investor demand is perfectly elastic with respect to the bid-ask spread, if any liquidity provider on Continuous was weakly profitably offering liquidity on Continuous at some spread \( s \geq \tilde{s}_{\text{continuous}} \), that provider could be strictly profitably undercut on Discrete at a strictly smaller spread \( s' \in (s^*_{\text{discrete}}, s) \). Furthermore, any liquidity cannot be offered on Discrete at any spread other than \( s^*_{\text{discrete}} \) 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.

These same arguments also imply that no liquidity can be offered on Continuous in any Stage 3 trading game even if Discrete were to charge a strictly positive (but small enough) trading fee. We thus obtain the following characterization of the full exchange competition game:

**Proposition 5.2.** Consider the full exchange competition game with a single Continuous exchange and single Discrete exchange. Any equilibrium has the following properties: (i) in period 1 of each Stage 3 trading game, exactly one unit of liquidity is provided on Discrete and no liquidity is provided on Continuous; (ii) Continuous earns zero profits; and (iii) Discrete charges strictly positive trading fees and earns expected per-trading-game profits that exceed \( N^{-1} \Pi^*_{\text{continuous}} \). Such an equilibrium exists.

In essence, when a single Discrete exchange competes against a Continuous exchange, Discrete is compensated for the elimination of the tax that latency arbitrage imposes on trading: as long as Discrete 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.

Propositions 5.1-5.2 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.
5.3 Multiple Discrete Exchanges

Now consider the case of multiple Discrete exchanges. With at least two Discrete exchanges (and potentially one or more Continuous exchanges), the resulting equilibrium has similar features to the equilibria with multiple Continuous exchanges, described in Proposition 3.2:

**Proposition 5.3.** Consider the full exchange competition game with 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 \( s_{\text{discrete}}^* \) around the current 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 with multiple continuous exchanges as studied in Section 3, in equilibrium multiple Discrete exchanges also operate as a virtual single platform: a single unit of liquidity is always provided in each trading game, the depth-volume relationship ensures that the marginal unit of liquidity is indifferent across exchanges, and equilibria differ from one another only in exchange market shares. However, there are two key differences. First, the bid-ask spread is \( s_{\text{discrete}}^* \), not \( s_{\text{continuous}}^* \), which is better for investors and informed traders because \( s_{\text{discrete}}^* < s_{\text{continuous}}^* \). Second, there are no longer latency arbitrage rents for exchanges or trading firms.

5.4 Prisoner’s Dilemma

We have now analyzed equilibrium of the exchange competition game with multiple Continuous exchanges (Section 3), a single Discrete and one or more Continuous exchanges (Section 5.2), and multiple Discrete exchanges (Section 5.3). We next establish that exchanges’ economic profits as a function of their market designs constitute a prisoner’s dilemma:

- If all exchanges are Continuous: each exchange \( j \) earns (per trading game) economic profits of \( NF_j^* \) (Proposition 3.2).

- If there is a single Discrete exchange and all other exchanges are Continuous: the Discrete exchange earns economic profits denoted \( \Pi^D \), where \( \Pi^D \in (\frac{N-1}{N}\Pi_{\text{continuous}}^*, \Pi_{\text{continuous}}^*) \), and the Continuous exchanges earn zero economic profits (Proposition 5.2).

- If there are multiple Discrete exchanges: all exchanges earn zero economic profits (Proposition 5.3).

Proposition 3.3 places an upper bound on exchange ESST revenues in \{all Continuous\}, while Proposition 5.2 places a lower bound on the Discrete exchange’s profits in \{a single Discrete, the remainder Continuous\}. These bounds and some simple algebra (Lemma C.4 in the appendix) yields that \( \Pi^D > NF_j^* \) for all exchanges \( j \), for any equilibrium ESST revenues consistent with Proposition 3.3 and
for $\Pi^D$ as characterized in Proposition 5.2. Discrete is thus a dominant strategy, but all exchanges prefer \{all Continuous\}, where they earn economic profits from speed technology, to \{all Discrete\} where they do not. We summarize these results in the following Proposition.

**Proposition 5.4.** [Prisoner’s Dilemma] Add a Stage 0 to the exchange competition game in which each of $M$ exchanges simultaneously choose either to operate as a continuous limit order book exchange (Continuous) or as a frequent batch auction exchange (Discrete). After these market design decisions, Stages 1 through 3 of the exchange competition game are played as before, with equilibrium as characterized by either Proposition 3.2 for \{all Continuous\}, Proposition 5.2 for \{a single Discrete, the remainder Continuous\}, or Proposition 5.3 for \{multiple Discrete, the remainder Continuous\}. Exchange profits 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. In practice, there are a few reasons incumbent exchanges might strictly prefer positive share to zero share even at competitive trading fees; for example, there are the “Tape Plan” data revenues discussed in Section 4.3, which are roughly proportional to market share.\(^{51}\) For the purpose of our analysis of adoption incentives below we will assume that if there is an initial adoption of Discrete by some exchange, at least one incumbent will choose to imitate and adopt Discrete as well.

### 5.5 Adoption Incentives

Given the prisoner’s dilemma payoff structure of exchanges’ economic profits, the analysis of exchange adoption incentives is relatively standard.

Let $c_{\text{adopt}}$ denote the fixed costs of being the first adopter of Discrete. In practice, adoption costs would include the costs of winning regulatory approval from the SEC for a new market design, engineering costs, the costs of explaining the new design to market participants, etc. If the first adopter is a de novo entrant, we assume that the entrant also has to pay a cost $c_{\text{entry}}$ associated with setting up a new exchange company, being granted a new exchange license by the SEC, etc. If the first adopter is an incumbent they do not pay $c_{\text{entry}}$, since they already have entered, but instead pay opportunity costs of no longer being a Continuous exchange. Since our analysis is all on a per-trading-game basis, we will interpret $c_{\text{adopt}}$ and $c_{\text{entry}}$ as per-trading-game costs paid in perpetuity.

We assume that if there is an initial adopter, whether a de novo or an incumbent, then incumbents can imitate after $T$ iterations of the trading game. Rather than formally modeling a dynamic entry

\(^{51}\text{For NYSE and Nasdaq specifically, another reason to strictly prefer positive share to zero share even at competitive trading fees is the listings business. Listings are lucrative (both the listing fees per se and revenue from the opening and closing auctions, which are hosted by the listings exchange), and seem to be reasonably sticky, but presumably it would be difficult to maintain this business if regular-hours market share were too low.}\)
and adoption game, we directly assume that at least one incumbent does in fact imitate when able to do so. As discussed above, this assumption represents that incumbents strictly prefer positive share to zero share even at competitive trading fees.

**Adoption Incentives: A New Entrant Exchange.** If a de novo entrant starts a new Discrete exchange, the entrant earns revenues of $\Pi^D$ per-trading game until it is imitated. Let $\rho \equiv (\sum_{t=0}^T \delta^t) / (\sum_{t=0}^{\infty} \delta^t)$ denote the share of net present value represented by the first $T$ iterations of an infinitely repeated series of trading games, where $\delta < 1$ denotes the per-trading game discount factor. The condition for a de novo to find entry profitable is thus:

$$\rho \Pi^D \geq c_{\text{adopt}} + c_{\text{entry}},$$  \hspace{1cm} (5.2)

Profitable entry by a de novo thus depends not only on whether the profitability of a standalone Discrete exchange $\Pi^D$ is large relative to adoption and entry costs $c_{\text{adopt}} + c_{\text{entry}}$, but also on the term $\rho$ which captures how quickly the entrant is imitated. Clearly, if $\rho$ is small enough (5.2) will not obtain, even if the magnitude of latency arbitrage, and hence $\Pi^D$, is large.

**Adoption Incentives: An Incumbent Exchange.** The condition for incumbent exchange $j$ to find it profitable to adopt Discrete, given that all other incumbents are still Continuous, is:

$$\rho \Pi^D \geq c_{\text{adopt}} + NF_j^\star_{\text{(status-quo rents)}}.$$  \hspace{1cm} (5.3)

The left-hand-side of (5.3) is the same as that for the de novo entrant, (5.2). The right-hand-side differs in that an incumbent does not incur entry costs, $c_{\text{entry}}$, but instead includes the incumbent’s opportunity cost of losing the exchange-specific speed technology rents it earns in the status quo, $NF_j^\star$. Our empirical results in Section 4.3 indicate that annual ESST revenues are on the order of $1$ billion per year, which, at a discount rate of between 5% and 10%, has a net present value of between $10$-$20$ billion. In contrast, anecdotal evidence suggests that $c_{\text{entry}}$ is on the order of $100$ million (see Section 6.1). Thus, the adoption condition for incumbent exchanges with ESST profits to protect is likely an order of magnitude more restrictive than the adoption condition for de novo entrants.

That incumbent exchanges continue to use the continuous limit order book market design is thus consistent with them maintaining the “cooperative” all-Continuous outcome of the prisoner’s dilemma summarized in Proposition 5.4. Does this sound reasonable? Consider the following quote from the Chief Economist of Nasdaq at a publicly recorded academic event in November 2013 when asked about adopting frequent batch auctions:

---

52Formally, the timing we consider is: there is one initial adopter of Discrete in Stage 0 (the other exchanges are Continuous); the exchanges and other players play Stage 1, then Stage 2, then $T$ iterations of Stage 3; then at least one other exchange adopts Discrete, and the exchanges and other players play Stage 1 again, then Stage 2 again, then infinitely repeat Stage 3.
“Technologically, we could do it. The big issue, one of the big issues for us, when I talked about cost, the cost we would bear, would be getting [the SEC] to approve it, which would take a lot of time and effort, and if we got it approved, it would immediately be copied by everybody else. ...So we would have essentially no first-mover advantage if we put it in there, we would have no incentive to go through the lift of creating [the new market design].”\(^{53}\) (Emphasis added.)

The quote suggests that industry participants believe that adoption costs are substantial, and — more importantly — if a new market design turns out to be successful, it would be swiftly imitated without much benefit to the first-mover. The quote does not underscore the additional disincentive for incumbents to adopt, namely the potential loss of rents from selling speed technology.

To capture formally that private and social incentives for innovation may diverge, add another parameter, \(DWL\), that captures the social deadweight loss from the arms race for speed (see discussion in Section 3.3). Social incentives for innovation are positive, but private incentives are negative — i.e., “the market will not fix the market” — if

\[
DWL > c_{\text{adopt}} + c_{\text{entry}} > \rho \Pi^D
\]

for new entrants and

\[
DWL > c_{\text{adopt}} \quad \text{and} \quad c_{\text{adopt}} + NF_j^* > \rho \Pi^D
\]

for all incumbents \(j\).

### 6 Discussion of Policy Implications

The basic question for policy is whether there will be a private-market solution to latency arbitrage and the arms race (i.e., “will the market fix the market”), or would some sort of regulatory intervention be required, and if so, of what form. On the one hand, the analysis in Section 5 suggests that private-market incentives may not be sufficient, even if the magnitude of latency arbitrage and associated deadweight loss is large (equations (5.2)-(5.3)). On the other hand, the analysis in Section 5 indicates that if there is an entrant, it will gain share in any equilibrium (Propositions 5.1-5.2). This suggests that, in the event that private incentives continue to be insufficient to motivate innovation, a potential option for policy is to provide a “push.” By push we mean any policy that tips the balance of incentives sufficiently to cause either condition (5.2) or (5.3) to obtain.\(^{54}\) Sections 6.1-6.2 discuss two

---

\(^{53}\)The event was a Workshop of The Program in the Law and Economics of Capital Markets at Columbia which featured a presentation of Budish, Cramton and Shim (2015) and an open discussion among the Program’s Fellows. The video is available via the internet wayback machine.

\(^{54}\)In our model, both a “push” of the sort described in this section and a market-design mandate would accomplish the same goal. Both would move the industry equilibrium from (all Continuous) to (all Discrete), and in doing so eliminate latency arbitrage and the associated arms race. With realistic frictions we would not expect a push to move the market to 100\% adoption; see Section 6.3 and Appendix F for discussion. A mandate, by definition, would move the market to
such potential pushes. Section 6.3 does rough back-of-envelope math to give a sense of magnitudes for the policy interventions and the overall costs and benefits for the market.

6.1 Policy Response 1: Reduce entry and adoption costs.

Examining equation (5.2), it is immediate that if policy could sufficiently lower entry and adoption costs (i.e., \( c_{\text{entry}} + c_{\text{adopt}} \)), it could ensure that a de novo entrant would have incentive to enter, because the left-hand side of (5.2) is strictly positive.

The cost of starting a new stock exchange is significant, and the risk of a new stock exchange design not getting approved is substantial as well. As evidence of the significant costs of entry, 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 (Crunchbase, 2018); this figure would combine what we call \( c_{\text{adopt}} \) and \( c_{\text{entry}} \). The Chicago Stock Exchange was purchased by NYSE for, reportedly, $70M, and many industry observers speculated that the sole reason NYSE bought CHX was to acquire its exchange license,\(^{55}\) that is, costs that are part of what we call \( c_{\text{entry}} \). As evidence of the significant risk of a new stock exchange design not getting approved, again consider IEX and CHX. IEX went through a protracted fight over its exchange design, and ultimately made significant concessions to gain approval. CHX, too, went through a protracted regulatory process over its proposed exchange design and ultimately withdrew its proposal after being acquired by the NYSE Group (Michaels and Osipovich, 2018).

One specific way the SEC could lower the risk-adjusted costs of entering as a new exchange with a new market design would be to proactively clarify what kinds of exchange designs are and are not allowed within the boundaries of Reg NMS (see Budish, 2016c). Such proactive clarification would certainly reduce risk, and would likely also reduce costs per se (e.g., legal costs).

In principle, if the social returns to a new market design are large but the private returns are negative, this would also justify a direct entry subsidy. The subsidy could be provided either by the government (with all the usual caveats) or by investors if they could find a way to act collectively. The subsidy would need to be large enough to get inequality (5.2) to obtain.


Examining equations (5.2) and (5.3), a key parameter that determines whether the innovator has sufficient incentive is \( \rho \), which captures the speed with which the innovator is imitated. The quote by the Nasdaq executive, “it would be immediately copied by everyone else,” is consistent with \( \rho \) being

\(^{55}\)The Wall Street Journal reported, of the merger, “Analysts say CHX’s most valuable asset is its license to run a national securities exchange. Applying for a new exchange license from the SEC can take years” (Michaels and Osipovich, 2018). At an industry conference attended by one of the authors around that time, numerous industry participants referred to CHX’s value to NYSE as coming entirely from its “medallion,” i.e., its license to run a stock exchange.
small in practice. The speed with which IEX’s symmetric speed bump was imitated by an exchange controlled by NYSE also speaks to ρ being small in practice.\textsuperscript{56}

Our impression is that the reasons ρ might be small in practice are that the “hard” parts of starting an exchange with a novel market design (given that the design itself has already been invented) are getting regulatory approval and educating the market as to how the novel exchange design works, whereas the actual programming and implementing of an exchange with a novel design is relatively cheap and fast. Therefore, once a first-mover has done the hard work of getting regulatory approval and educating the market, a second-mover can rapidly and cheaply imitate if they would like.

This economic issue — that a potential innovator would not have incentives to invest if their innovation will be quickly imitated — is of course a familiar one. In many other contexts, the problem is solved by patents or other legal forms of market exclusivity (see Williams, 2017). Such policies explicitly trade off the static inefficiency of monopoly for the dynamic efficiency of eliciting useful innovations.

Here, patents do not seem to be a viable way to create market exclusivity for at least two reasons. 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 (Hosman et al., 2017) in Jan 2016 for a market design idea that a close reader will recognize as, in essence, a form of batch auction, without using the word “auction” a single time.\textsuperscript{57}

A potential alternative way to create market exclusivity would be to have the SEC grant a modest period of exclusivity to the innovator, during which time other exchanges would not be allowed to imitate the design (either identically or with designs judged to be essentially similar). This idea is somewhat analogous to a practice of the Food and Drug Administration, wherein it grants a period of market exclusivity for certain kinds of drugs that, for various reasons, are not patentable (Food and Drug Administration, 2015).

\textsuperscript{56}IEX’s exchange application was approved in June 2016. In Jan 2017 NYSE MKT LLC, subsequently renamed NYSE American, filed for approval to incorporate an analogous speed bump into its exchange (“... the proposed Delay Mechanism would function similarly to the intentional delay mechanism of IEX ...”) and explicitly cited IEX’s approval as legal precedent for its approval (“The proposed rule text is based on Supplementary Material ... to IEX Rule 11.510 without any substantive differences”).

\textsuperscript{57}Here is an excerpt of the text from the abstract of the CME patent application (emphasis added): “The disclosed embodiments may mitigate such [latency] disparities by buffering or otherwise grouping temporally proximate competing transactions together upon receipt, e.g. into a group, collection, set, bucket, etc., and subsequently arbitrating among those grouped competing transactions, in a manner other than solely based on the order in which the competing transactions in the group were received, to determine the order in which those competing transactions will be processed, thereby equalizing priority of transactions received from participants having varying abilities to rapidly submit transactions or otherwise capitalize on transactional opportunities” (Hosman et al., 2017).
6.3 Rough Magnitudes

Recent empirical evidence in Aquilina, Budish and O’Neill (2020) finds that latency arbitrage profits as a proportion of trading volume is about 0.4 basis points in UK equity markets (0.004%). This number would imply annual latency arbitrage profits in U.S. equity markets on the order of $2 billion per year, based on the approximately $50 trillion of annual regular-hours on-exchange trading volume, or about $0.0020 per share, based on approximately 1 trillion shares traded. Exchange ESST revenues combined with the bound from the theory (Proposition 3.3) also point to annual latency arbitrage profits of roughly this magnitude; see the discussion in Section 4.3. A discount rate of between 5% and 10% applied to the $2 billion per year figure implies that the net present value of latency arbitrage profits in U.S. equity markets is on the order of $20 to $40 billion. While admittedly rough, and based on extrapolation from UK equities which may not be comparable to U.S. equities, this gives a sense of magnitudes for the benefits of addressing latency arbitrage in the U.S. stock market.

This magnitude suggests that an entry subsidy easily passes a cost-benefit test. Getting a sense of magnitudes for the exclusivity period is more involved. In Appendix G, we provide details for a back-of-the-envelope calculation that suggests an exclusivity period on the order of 1-2 years might be sufficient.

7 Conclusion

In the quotation at the beginning of this paper, the SEC Chair asked “whether trading venues have sufficient opportunity and flexibility to innovate successfully with initiatives that seek to deemphasize speed as a key to trading success...” We have put forth a theoretical model of stock exchange competition that clarifies why, even if allowed, exchanges may not want to innovate: they profit from the speed race generated by the existing market design. Our story is not about new markets failing to gain traction if introduced (as may be the case in other settings with stronger network externalities and potential for coordination failure), 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 proposals, which borrow simple economic insights from the innovation and intellectual property literature, attempt to alter the incentives for private innovation to better align private incentives with social interests, to encourage “the market to fix the market.”

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 (U.S. Securities and Exchange Commission, 2019a). Suspending UTP is a way of creating exclusivity for the innovator, analogous to our ideas in Section 6.2. In February 2020, the SEC issued a proposed reform to the market for exchange data (U.S. Securities
and Exchange Commission, 2020a). The proposed rule cited our theoretical finding that each exchange has market power in the sale of proprietary market data and related speed technology (pg. 366), as well as our empirical finding that exchanges earn significant revenue from selling these products (pg. 365). 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.” (Jackson, 2020)

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 U.S. stock exchange industry. One natural direction for future research would be to use the model as a starting point for analysis of the entry and merger incentives of stock exchanges.58 Another natural direction for future research is to extend this style of analysis — at the intersection of market design, IO and finance, with theory and empirical work guided by institutional and regulatory details — to other asset classes and geographies with different regulatory frameworks. As emphasized in the text, futures markets would be of particular interest, since the seemingly small difference that futures contracts are not fungible across exchanges leads to large differences in industry structure.59 The U.S. Treasury secondary market would be another natural subject, given both its size and importance per se, and recent scrutiny regarding market design issues (Powell, 2015; Joint Staff Report, 2015).

References


58 A recently announced entrant exchange, the Members’ Exchange (MEMX), is owned by a consortium of nine major trading firms and broker-dealers. From reports to date, it appears that MEMX is not innovating on market design, but rather seems motivated by concern over rising fees for proprietary data, co-location and connectivity, as we documented in Stylized Fact #7 (Osipovich, 2019b; Levine, 2019). In this regard, MEMX’s entry might be interpretable as a combination of business-stealing in the sense of Mankiw and Whinston (1986) and an attempt to gain bargaining leverage, rather than innovation on welfare-enhancing dimensions.

59 Interestingly, in early 2019, one of the world’s largest futures exchange operators, the Intercontinental Exchange, filed for approval for a market design that would address latency arbitrage (Osipovich, 2019a), even while its subsidiary, the New York Stock Exchange, has in the past opposed such innovations in equity markets.


Food and Drug Administration. 2015. “Patents and Exclusivity.”


