NBER WORKING PAPER SERIES

A THEORY OF STOCK EXCHANGE COMPETITION AND INNOVATION: WILL THE MARKET FIX THE MARKET?

Eric Budish
Robin S. Lee
John J. Shim

Working Paper 25855
http://www.nber.org/papers/w25855

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2019

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 Cramton, Doug Diamond, David Easley, Alex Frankel, Joel Hasbrouck, Kate Ho, Anil Kashyap, Pete Kyle, Donald Mackenzie, Neale Mahoney, Paul Milgrom, Joshua Mollner, Ariel Pakes, Al Roth, Fiona Scott Morton, Sophie Shive, Andrei Shleifer, Jeremy Stein, Mike Whinston, Heidi Williams and Luigi Zingales for valuable discussions and 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

A Theory of Stock Exchange Competition and Innovation: Will the Market Fix the Market?
Eric Budish, Robin S. Lee, and John J. Shim
NBER Working Paper No. 25855
May 2019, Revised July 2020
JEL No. D02,D44,D47,D53,D82,G1,G2,G23,L1,L13,L5,L89

ABSTRACT

This paper builds a new model of financial exchange competition, tailored to the institutional details of the modern US stock market. In equilibrium, exchange trading fees are competitive but exchanges are able to earn economic profits from the sale of speed technology. We document stylized facts consistent with these results. We then use the model to analyze incentives for market design innovation. The novel tension between private and social innovation incentives is incumbents' rents from speed technology in the status quo. This creates a disincentive to adopt new market designs that eliminate latency arbitrage and the high-frequency trading arms race.

Eric Budish
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
eric.budish@chicagobooth.edu

John J. Shim
Booth School of Business
University of Chicago
5807 S Woodlawn Avenue
Chicago, IL 60637
jshim2@nd.edu

Robin S. Lee
Department of Economics
Harvard University
Littauer Center 120
Cambridge, MA 02138
and NBER
robinlee@fas.harvard.edu

A Data Appendix is available at http://www.nber.org/data-appendix/w25855

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 2020, all 13 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 more traditional economic one 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, eco-

nomically many aspects of trading activity behave as if there is just a single "synthesized" exchange.[1] 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).[2] 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. As long as trading firms multi-home and purchase speed technology from all exchanges (which they do in equilibrium), exchanges can charge positive fees for speed technology without incentives to undercut each other. 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

[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 the single synthesized exchange (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.

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.

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 single synthesized exchange. 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 arguably makes the single synthesized exchange 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 — or about 0.0001% per side for a \$100 stock. This is not zero, as the theory predicts, but is small.[3] 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 2018 proprietary data and co-location revenues are on the order of \$1 billion, or about five times regular-hours trading fee revenues.

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

[3]The \$0.0001 per share per side implies that across the approximately 1 trillion shares traded during regular hours each year, exchanges earn approximately \$200 million in trading fees. While not zero, \$200 million is small relative to both exchange operating expenses and overall exchange revenues (see Section 4.2).

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.[4] Formally, we assume that the alternative market design eliminates latency arbitrage but does not have any additional benefits or costs for the market.[5] All of this analysis applies equally to the asymmetric delay design with a very short delay interval.

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.[6] 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 unilater-

[4] 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 details regarding information policy which, following Budish, Cramton and Shim (2015), is analogous to information policy in the continuous market but with the same information (about trades, cancels, the state of the order book, etc.) disseminated in discrete time, at the end of each interval.

[5] See Budish, Cramton and Shim (2015) for discussion of potential benefits and costs of frequent batch auctions besides the elimination of latency arbitrage. In particular, if the batch interval is longer, this may introduce the kinds of benefits and costs that emerge in models such as Vayanos (1999) and Du and Zhu (2017) and that was discussed in earlier work on batch auctions by Schwartz (2001).

[6] This result may seem to contradict the result in Glosten (1994), Proposition 9, that finds that the electronic limit order book is in a certain sense "competition proof." 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.

ally 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 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.[7]

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.

We discuss two such pushes. First, reducing the costs of adoption, either by direct subsidy or by finding ways to lower the costs of launching a new stock exchange with a novel market design (e.g.,

[7]There have been two other exchange design proposals in the US that relate to latency arbitrage but outside of the displayed, Reg NMS protected market. The Investors' Exchange (IEX), whose design was approved by the SEC in June 2016, uses a *symmetric* speed bump (of 350 microseconds), and thus does not address latency arbitrage for displayed limit orders. For displayed limit orders, its market design is a standard continuous limit order book but with 350 microseconds of additional artificial distance from participants. IEX's market design does address latency arbitrage for non-displayed pegged orders and such orders are currently the source of most of IEX's trading volume (in the period Jan-June 2020, displayed orders are 13.4% IEX's trading volume and non-displayed orders are 86.6% of IEX's trading volume). Cboe's EDGA exchange proposed an asymmetric speed bump in 2019, but proposed for the exchange to not have Reg NMS quote protection, and proposed a delay amount (4 milliseconds) that was longer than the "de minimis" delay threshold of 1 millisecond adopted by the SEC in June 2016. EDGA's proposal was rejected by the SEC in Feb 2020, in part on grounds that the choice of 4 milliseconds was not adequately justified ("the Commission does not believe that the Exchange has supported its assertions and demonstrated that the [proposed] delay mechanism is appropriately tailored to address latency arbitrage and not permit unfair discrimination." U.S. Securities and Exchange Commission, 2020*b*)

reducing the costs of the regulatory approval process). Second, a market design exclusivity period for the innovator, roughly analogous to FDA exclusivity periods for non-patentable drugs. Back-of-envelope calculations suggest that the magnitude of either push could be modest relative to the stakes.

Contributions and Related Literature. Our paper makes three sets of contributions to the literature. First is our theoretical industrial organization model of the stock exchange industry, an industry which is both economically important per se and of symbolic importance. We depart from much of the previous literature on financial exchange competition in both our focus — the source of economic profits for U.S. stock exchanges and their incentives to adopt new market designs — and in our modeling approach. Most centrally, most other papers in this literature have some sort of single-homing, either by market participants choosing which one exchange to trade on (e.g., Pagano, 1989; Santos and Scheinkman, 2001; Ellison and Fudenberg, 2003; Pagnotta and Philippon, 2018; Baldauf and Mollner, 2019), or by financial instruments that are specific to a single exchange (as in Cantillon and Yin, 2008). This single-homing is often (though not always) accompanied by some meaningful differentiation across exchanges, either horizontally or vertically. By contrast in our model, motivated by the regulatory environment for modern electronic U.S. stock trading, stocks are fungible across exchanges, market participants can frictionlessly multi-home across exchanges, and exchanges are undifferentiated. This modeling approach also leads to economics of the status quo that are different from many platform or two-sided competition frameworks where, typically, platforms earn rents from platform-specific network effects by charging supra-competitive access or transaction fees (Caillaud and Jullien, 2003; Rochet and Tirole, 2003; Armstrong, 2006; Farrell and Klemperer, 2007). Here, since exchanges are modeled as undifferentiated and exchange-specific network effects are nullified due to frictionless search and access, trading fees are competitive — zero in our model, and very small in the data — and exchanges are only able to earn rents from the sale of an optional "add-on" service (speed technology). A related insight of our model that may be of interest to the platforms literature is that, while the market may appear to be fragmented across multiple exchanges, the market behaves in some respects as if there were a single "synthesized" exchange.

There are also two technical features of our theoretical analysis worth highlighting. First, we develop and motivate an equilibrium solution concept, *order-book equilibrium*, to address Nash equilibrium existence issues that arise in Glosten and Milgrom (1985), Budish, Cramton and Shim (2015) and related models. This solution concept is closely related to alternative solution concepts employed in the insurance market literature (e.g., Wilson, 1977; Riley, 1979), which also has to deal with existence issues arising due to adverse selection. Second, our model generates an interior split of latency arbitrage rents between exchanges and trading firms without relying on an explicit bargaining model; this arises as a result of exchanges being able to post prices for speed technology (which they do in reality), and trading firms being able to steer trading volume via the provision of liquidity (which they can in reality).

Our paper's second contribution is the seven stylized empirical facts. In particular, the facts on trading fees and speed technology revenue are directly applicable to current policy debates, which have also attracted attention from Jones (2018), Spatt (2019) and Glosten (2020). With respect to trading fees, while our results do not speak to the agency concerns at the heart of the policy debate (see Battalio, Corwin and Jennings, 2016), our results do show that, once one cuts through the complexity of modern fee schedules, the average fees are economically small. With respect to speed technology fees, in October 2018 for the first time in recent history the SEC rejected proposed data fee increases by NYSE and Nasdaq (Clayton, 2018), and in February 2020 the SEC proposed reforms to rules concerning proprietary market data. This paper was cited in a policy address on this topic by Commissioner Robert J. Jackson Jr. and in the market data reform proposal itself (U.S. Securities and Exchange Commission, 2020*a*; Jackson, 2020).

Last is our analysis of the incentives for market design innovation. To be precise, the theoretical contribution is characterizing unique aspects of equilibrium when there is a single adopter of a novel market design that eliminates latency arbitrage, and unique aspects of equilibrium when there are multiple such adopters. These analyses, paired with the earlier analysis of the status quo, enable us to fill in the cells of the payoff matrix corresponding to a market design adoption game. Once we understand that the adoption game payoffs constitute a prisoner's dilemma, as opposed to, e.g., a coordination game, the rest of the theoretical analysis follows standard ideas from the innovation and intellectual property literature. We then use this analysis to identify modest potential policy responses — a "push" as opposed to the "prescriptive regulation" of which the SEC Chair expressed wariness. We view this contribution as in the spirit of economic engineering (Roth, 2002), working with the real-world constraints of the specific market design setting, rather than assuming the ability to design institutions from scratch.[8]

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.

[8]In this spirit our work is related in approach, if not subject matter, to research in market design on topics such as spectrum auctions (Ausubel, Cramton and Milgrom, 2006; Levin and Skrzypacz, 2016; Milgrom and Segal, 2019), school choice (Abdulkadiroğlu and Sönmez, 2003; Abdulkadiroğlu, Agarwal and Pathak, 2017; Kapor, Neilson and Zimmerman, 2020), kidney exchange (Roth, Sönmez and Ünver, 2004; Agarwal et al., 2019; Akbarpour et al., 2019), course allocation (Sönmez and Ünver, 2010; Budish, 2011; Budish et al., 2017), online advertising (Edelman, Ostrovsky and Schwarz, 2007; Athey and Ellison, 2011), and transportation (Hall, 2018; Ostrovsky and Schwarz, 2018). There is also a burgeoning literature specifically on market design issues in financial markets. Recent examples include Allen et al. (2020), Antill and Duffie (2018), Asquith et al. (2013), Asquith, Covert and Pathak (2019), Bhattacharya, Illanes and Padi (2020), Brogaard, Hendershott and Riordan (2017), Bulow and Klemperer (2013), Bulow and Klemperer (2015), Du and Zhu (2017), Duffie and Dworczak (2018), Duffie and Zhu (2016), Hendershott and Madhavan (2015), Hortaçsu, Kastl and Zhang (2018), Kastl (2017), Kyle and Lee (2017), and Kyle, Obizhaeva and Wang (2018).

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).[9] 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.[10] 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.

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.[11] Sophisticated market participants can take on responsibility for compliance with

[9] Technically, 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.

[10] For surveys of modern electronic trading, focusing on a broader set of issues than stock exchanges per se, good starting points are Jones (2013), Fox, Glosten and Rauterberg (2015, 2019), O'Hara (2015) and Menkveld (2016).

[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

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)[12] 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.[13]

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. In Section 5 we extend the model and allow exchanges to be strategic with 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

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, 2016*b*).

[12]Investors and brokers who do not utilize proprietary data feeds from exchanges instead use a non-proprietary data feed called the SIP (Securities Information Processor). The SIP feed provides data on the best bid and offer across all exchanges, and is relatively cheap, with fees set by a regulatory process and revenues allocated across exchanges according to a regulatory formula. However, the SIP feed is slower than proprietary data feeds, primarily because of the time it takes to aggregate and disseminate data from geographically disparate exchanges. The SIP feed also lacks some additional data that is available from proprietary feeds, specifically data on depth beyond the best bid and offer, and data on trades of odd lots. One way to think about the SIP feed is that is appropriate for smaller, non-latency sensitive traders, but not latency-sensitive market participants. For the purpose of the model, we model the SIP as cheaper (modeled as free) but slower than proprietary feeds. We discuss exchange revenues from the SIP feed briefly in Section 4.3; we net these revenues out from our estimate of total exchange-specific speed technology revenues.

[13]Note that "dark pools," or Alternative Trading Systems, are not governed by Reg NMS. Instead, dark pools typically facilitate trade at prices that reference the best available quotes from exchanges (e.g., at the midpoint). This of course raises its own interesting economic issues, specifically that dark pools may "free ride" off of prices discovered by the exchanges. See, for instance, Hendershott and Mendelson (2000), Zhu (2014), and Antill and Duffie (2018).

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.[14] 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 but we include the case $M = 1$ in the model so that the BCS environment is a special case. Exchanges all use the continuous limit order book market design and are ex ante undifferentiated. The asset 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. This fungibility captures the economics of Unlisted Trading Privileges as discussed in Section 2. We assume that prices are continuous and that shares are perfectly divisible.

[14] 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.

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.

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 λ_{invest} in each trading game, and has an inelastic need to buy or sell one unit of x, with buying or selling equally likely. 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] 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).

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 λ_{public}, and the probability that there is a jump in y seen by an informed trader is $\lambda_{private}$. 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 $y - p$; similarly, his payoff if he sells a unit of x at price p 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 "slow" trading firms that do not possess such technology. Note, practically, that what we mean by a

[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 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.[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.[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.[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. In the case that only a single TF purchases ESST from a given exchange j, 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. 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.

The following objects are primitives of the game: (i) the arrival rates of investors (λ_{invest}), and of publicly (λ_{public}) and privately ($\lambda_{private}$) observed jumps in y; (ii) the jump size distribution; (iii) the number of fast TFs (N); and (iv) the number of exchanges (M).

[17] In practice exchanges often charge different fees for "making" liquidity as opposed to "taking" liquidity; see Section 4.2. 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 if allowed.

3.1.2 Timing

There are three *stages* to our game. In Stage One, exchanges simultaneously choose trading and ESST fees. In Stage Two, trading firms simultaneously decide which exchanges to purchase ESST from. Finally, in Stage Three, a *trading game* is repeated infinitely often. Formally:

1. Stage One *(Exchange Price Setting)*: All M exchanges simultaneously choose per-share trading fees $\mathbf{f} = (f_1, \ldots, f_M)$ and per-trading game ESST rental fees $\mathbf{F} = (F_1, \ldots, F_M)$.

2. Stage Two (*Speed Technology Adoption*): All N TFs with general speed technology simultaneously decide which exchanges to purchase ESST from.

3. Stage Three *(Infinitely Repeated Trading Game)*: At the beginning of each trading game, there is a publicly observed *state*, which consists of the current fundamental value of the security (y), and the current outstanding bids and asks in each exchange's limit order book ($\boldsymbol{\omega} = (\omega_1, \ldots, \omega_M)$, where ω_j is also referred to as the *state of exchange j's order book*). If it is the first play of the trading game, the initial fundamental value is y_0, and each exchange's order book is initially empty. Otherwise, the state is determined at the conclusion of the previous trading game and ω_j for each exchange j contains all limit orders that remain outstanding on that exchange. Each trading game is divided into two periods.

   (a) Period 1: Trading firms simultaneously submit orders to any subset of exchanges after observing the state $(y, \boldsymbol{\omega})$ at the beginning of the trading game. An *order* for TF i submitted to exchange j is a set of messages denoted by $o_{ij} \in \mathcal{O}$, where $\mathcal{O}$ is 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), where such an order 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 which case the TF simply maintains its existing limit orders in ω_j, if any exist. For each exchange j, all orders sent to exchange j in this period are serially processed by the exchange in a random sequence, with the speed of the TF sending the order serving as a tie-breaker: first, orders from fast TFs who have purchased ESST from the exchange are processed in a uniformly random sequence; then orders from fast TFs who have not purchased ESST from the exchange are processed in a uniformly random sequence; and last, orders from slow TFs are processed in a uniformly random sequence.[20]

[20] We assume that market participants can only send at most a single order (set of messages) to an exchange each

(b) Period 2: After period-1 orders have been processed by each exchange and incorporated into each exchange's order book, nature moves and selects one of four possibilities:

i. With probability λ_{invest}: an investor arrives, equally likely to need to buy or sell one unit of x. The investor has a single opportunity to send IOCs to all exchanges. The investor's activity may affect $\boldsymbol{\omega}$; y is unchanged.

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

iii. With probability λ_{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 $\boldsymbol{\omega}$.

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

The state $(y, \boldsymbol{\omega})$ at the end of the trading game remains the state for the beginning of the next trading game.

3.1.3 Discussion of Institutional Details

Unlisted Trading Privileges (UTP). As noted before, we incorporate UTP into the model by having the same asset trade on all exchanges, and by having the value of the asset be completely independent of the exchange on which it is bought or sold. We emphasize that the 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.

Regulation National Market System (Reg NMS). The Stage 3 trading game implicitly assumes that all market participants face, on an order-by-order basis, what we call *frictionless search and access*. More specifically, by frictionless search we refer to the fact that all market participants observe the current state of the order book on all exchanges, $\boldsymbol{\omega}$, at zero cost prior to taking any action in any period of a trading game. By frictionless access we refer to the fact 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.

period, and that exchanges process all messages within an order before processing any other order. This implies that market participants cannot improve the chances of their messages being processed faster by sending additional messages.

Synchronizing Trades Across Exchanges. The Stage 3 trading game implicitly assumes that investors and informed traders, upon arrival to the market, can synchronize their orders across exchanges such that they can execute trades across multiple exchanges before other market participants can react. That is, an investor or informed trader can send trades to exchanges j and j' such that their arrival times are sufficiently synchronized that it is not possible for a TF to observe the activity on exchange j and respond on exchange j', before the investor or informed trader's own order reaches j'. This is captured in the model by allowing the investor or informed trader to trade on all exchanges in Period 2 before TFs see the updated state and can react in Period 1 of the subsequent trading game.

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).

3.2 Equilibrium Analysis

3.2.1 Stage Three Trading Game in the BCS Environment

We begin our equilibrium analysis by analyzing the Stage 3 trading game in the subgame of our model that corresponds most closely to the environment in BCS: there is only a single exchange ($M = 1$), this exchange 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, and hence are equally fast. The level of the ESST fee F_1 is not important for this exercise, only the assumption that we are in the subgame where all N fast TFs are equally fast on the one exchange.

The analysis of this subgame will be a helpful input into the analysis of the full model 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 environment will be a helpful guide to the economics of the full model. In particular, in our analysis of the full model, we will confirm that there are equilibria in which all active exchanges set trading fees of zero and all N trading firms purchase ESST on all active exchanges, and that in these equilibria the effect of sniping on the market's cost of liquidity carries through from this special case.

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. In period 1 of a trading game, the state consists of the public value of y and the state of the order book $\boldsymbol{\omega}$ that carries over from the end of the previous trading game. In period 2 the state consists of the updated state of the order book $\boldsymbol{\omega}$ based on play in period 1, and the updated value of y if nature selected a public or private jump.

Even though the trading game is infinitely repeated, we will first analyze each trading game in isolation, thereby ignoring the possibility that actions in one trading game may affect continuation

payoffs in subsequent games. We then 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.

Period 2: Optimal Play. Working backwards, note that 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:

- Investor or Informed Trader Arrival. If either an investor arrives or an informed trader arrives in Period 2, they have essentially unique optimal strategies given the state. An investor sends an IOC order to trade up to one unit in their desired direction; additionally, if there are any remaining orders that are profitable to trade against based on the publicly observed state y 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. Any 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. 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 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 essentially unique optimal (i.e., dominant) 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. Consider a fast TF choosing to provide liquidity via non-marketable limit orders at the beginning of period 1. Since investors are equally likely to arrive needing to buy or sell one unit of x and the distribution of jumps in y is symmetric about zero, it is convenient to focus on the provision of liquidity via two limit orders: for a given

quantity q and fundamental value y, the TF submits an order to buy x at $y - s/2$, and an order to sell x at $y + s/2$ for some *bid-ask spread* $s \geq 0$. In traditional models of adverse selection (Copeland and Galai, 1983; Glosten and Milgrom, 1985), the benefit of offering to either buy or sell 1 unit of x at a spread of s (when there is no additional liquidity offered in the order book) is earning the bid-ask spread if an investor arrives and trades, which in a single play of our trading game yields benefit equal to $\lambda_{invest} \cdot \frac{s}{2}$ per-unit in expectation; and the cost of such liquidity provision is the cost of being adversely selected if the informed trader sees private information and trades, which in a single play of our trading game equals $\lambda_{private} \cdot L(s)$ per-unit, 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 upon arrival of a privately observed jump in y.

The continuous limit order book market design imposes an additional cost of liquidity provision, namely sniping: with probability $\lambda_{public} \cdot \frac{N-1}{N}$ a fast liquidity provider is sniped, and the loss if sniped is also $L(s)$ per-unit. For a fast TF to be indifferent between providing 1 unit of liquidity at some bid-ask spread and sniping a rival TF offering that same amount of liquidity at the same spread (succeeding with probability $\frac{1}{N}$), the spread $s^*_{continuous}$ must satisfy:[21]

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

Note that the cost of getting sniped on the right-hand side of (3.1), $\lambda_{public} \cdot L(s^*_{continuous})$, reflects both the $\frac{N-1}{N}$ probability that a fast liquidity provider loses the race to respond to public information, as well as a $\frac{1}{N}$ factor that captures a fast liquidity provider's opportunity cost of not themselves sniping. For this reason equation (3.1) also reflects the bid-ask spread at which a slow TF is indifferent between providing liquidity and not. 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.[22] Equation (3.1) has a unique solution since the left-hand side is strictly increasing and the right-hand size is strictly decreasing in $s^*_{continuous}$, 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. This existence issue has been known to the literature since Glosten and Milgrom (1985). The key intuition

[21] If public and private information had different jump distributions, denoted J_{public} and $J_{private}$, the right-hand side of (3.1) would be $\lambda_{public} \cdot \Pr(J_{public} > \frac{s}{2}) \cdot E(J_{public} - \frac{s}{2} | J_{public} > \frac{s}{2}) + \lambda_{private} \cdot \Pr(J_{private} > \frac{s}{2}) \cdot E(J_{private} - \frac{s}{2} | J_{private} > \frac{s}{2})$. Since assuming that public and private information have the same jump distribution simplifies the expression considerably without loss of economic meaning, we adopt that assumption, even though in practice the two distributions could of course be different.

[22] Evidence in Aquilina, Budish and O'Neill (2020) suggests that both cases, i.e., both fast and slow TFs providing liquidity that sometimes gets sniped, are empirically relevant.

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 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.[23]

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, in a single play of the Stage 3 trading game, strictly weakens Nash equilibrium). 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.

The first type of reaction allows for TFs to provide additional liquidity at a better price if profitable to do so (a "profitable price improvement"). This means that if a TF that is providing liquidity unilaterally deviates by widening their spread, another TF could undercut and provide additional liquidity in response; hence, TFs are able to discipline equilibrium price levels without having to put excess liquidity in the order book. This captures the spirit of competitive liquidity provision, as discussed and assumed in Glosten and Milgrom (1985), but in our setting where fast TFs earn strictly positive profits.

The second type of reaction allows for TFs to withdraw liquidity in response to deviations if profitable to do so. This reaction addresses a profitable deviation that we call "have your cake and eat it too," in which one TF adds liquidity at a slightly lower spread to both earn revenues from liquidity provision and earn rents from sniping the liquidity it just undercut. If a TF engaged in such a deviation, any rival TF whose quotes are undercut is able to withdraw if it would like to do so (e.g., if its liquidity would no longer be filled by an investor).

[23] Here is a short proof of the non-existence claim for a single play of the trading game. Consider any set of TF strategies where, at the end of period 1, exactly one unit of liquidity is offered, for instance at spread $s^*_{continuous}$ as defined in (3.1). This cannot be a Nash equilibrium because any TF that is providing liquidity strictly prefers to deviate and widen their spread: this strictly increases the TF's profits if an investor arrives, and strictly reduces the TF's expected adverse selection and latency arbitrage costs. If instead strictly greater than one unit of liquidity is offered, then any liquidity that would not be filled by an investor with certainty (either because there is at least one unit of liquidity that is more attractively priced, or because it is tied and would only get filled with some probability less than one) has a strictly profitable deviation as well, either to be withdrawn or to be offered at a slightly narrower price (jumping the queue if tied). Last, if there is strictly less than one unit of liquidity offered, there is a strictly profitable deviation to add the missing amount at a high spread, in case an investor arrives. Hence, there is no Nash equilibrium.

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 restore equilibrium existence, in Appendix B.2.

Similar restrictions on the set of allowable deviations have been employed by alternative solution concepts in insurance markets. Our particular concept is closest in spirit to and borrows inspiration from the *E2 equilibrium* in Wilson (1977) and the *reactive equilibrium* in Riley (1979) (see also discussion in Engers and Fernandez, 1987; Handel, Hendel and Whinston, 2015).[24] Our relation to this literature is not accidental: both financial and insurance markets feature adverse selection, and in both settings firms that are "undercut" by a rival (who offers a better price, or who offers a product that attracts less adversely selected consumers) may wish to withdraw from the market rather than face an adversely selected set of trading partners.

Equilibrium in the BCS Environment. We can now formally characterize equilibrium of the Stage 3 trading game in the BCS environment. 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 with all N fast trading firms having purchased exchange-specific speed technology from this exchange. Any equilibrium has the following properties. In period 1 of each trading game: a single unit of liquidity is provided at bid-ask spread $s^*_{continuous}$ (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, upon arrival, immediately transacts one unit at the best bid or offer if their privately-observed jump in y exceeds $\frac{s^*_{continuous}}{2}$; and if there is a publicly-observed jump in y that exceeds $\frac{s^*_{continuous}}{2}$, 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 in appendix.) The proof confirms that repeated play of any order book equilibrium in which period-1 TF strategies condition only on the publicly observable state $(y, \boldsymbol{\omega})$ comprises an order book equilibrium of the infinitely repeated trading game. Since the bid-ask spread is $s^*_{continuous}$ in any

[24] Both Wilson (1977) and Riley (1979) examine equilibria among firms providing insurance policies, and introduce solution concepts that admit dynamic responses to deviations in order to address related equilibrium existence issues. A set of policies comprises an *E2 equilibrium* (Wilson, 1977) if there are no strictly profitable unilateral deviations that remain profitable even if policies, rendered unprofitable by the deviation, are withdrawn. A set of policies comprises a *reactive equilibrium* (Riley, 1979) if there are no strictly profitable unilateral deviations that remain profitable even if a rival reacted by offering additional policies, and such a reaction would not generate losses for the rival even if additional policies were offered. To counter profitable deviations, our order book equilibrium solution concept allows for two types of reactions: the withdrawal of unprofitable liquidity (similar to Wilson), and the addition of liquidity that must remain profitable even if others' liquidity could then be withdrawn (similar to Riley).

order book equilibrium of an individual trading game, beginning at any publicly observable state, each fast TF earns the same amount in expectation, $\frac{1}{N}\lambda_{public} \cdot L(s^*_{continuous})$, whether it provides liquidity or snipes stale quotes, and slow TFs earn zero in expectation whether they provide liquidity or do nothing. As a result, as long as other TFs play strategies prescribed by an order book equilibrium, no individual TF has a profitable deviation even when future play is accounted for.

3.2.2 Equilibrium of the Full Exchange Competition Game

We now turn to equilibrium of the full exchange competition game. Recall that in Stage 1, the M exchanges simultaneously choose trading fees and exchange-specific speed technology (ESST) fees; in Stage 2, the N fast TFs make their ESST purchase decisions; in Stage 3, the infinitely repeated trading game is played, only now with multiple exchanges. 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 states 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 meaning that 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 $s^*_{continuous}$ as in the BCS environment studied above, but now this liquidity is spread across multiple exchanges according to a vector of market shares, denoted $\boldsymbol{\sigma}^*$. Investors and informed traders, when they arrive, split their orders across exchanges according to $\boldsymbol{\sigma}^*$. 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 to "synthesize" a single exchange from the M parallel exchanges, and then act economically the same way as in the single exchange case. The main difference here is that exchanges and fast TFs now split the rents generated from sniping.

Proposition 3.2. *Consider the full exchange competition game with $M \geq 2$ exchanges. For any vector of market shares $\boldsymbol{\sigma^*} = (\sigma^*_1, \ldots, \sigma^*_M : \sum_j \sigma^*_j = 1)$, and for any vector of exchange-specific speed technology (ESST) fees $\boldsymbol{F}^* = (F^*_1, \ldots, F^*_M)$ 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 $(y, \boldsymbol{\omega})$. At the end of period 1, σ^*_j quantity of liquidity is provided on each exchange j at spread $s^*_{continuous}$ (defined in (3.1)) around y. In period 2: an investor, upon arrival, immediately transacts σ^*_j at the best bid or offer on each exchange j; an informed trader, upon arrival, immediately transacts σ^*_j at the best bid or offer on each exchange j if their privately-observed jump in y exceeds $\frac{s^*_{continuous}}{2}$; and if there is a

publicly-observed jump that exceeds $\frac{s^*_{continuous}}{2}$*, a sniping race occurs on all exchanges, in which all fast trading firms attempt to trade against all stale quotes provided by trading firms other than themselves, and all fast trading firms providing any liquidity on any exchange attempt to cancel their stale quotes.*

The condition on ESST fees is:

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

where $\Pi^*_{continuous} \equiv \lambda_{public} \cdot L(s^*_{continuous})$ *denotes the total* "sniping prize," *and* $\pi_N^{lone-wolf}$ *is a constant discussed below and defined in Appendix B.3.2, equation (B.3).*

The proof of this result is constructive. We first examine behavior in the multi-exchange version of our trading game in Stage 3, and show that if all N fast trading firms purchase ESST from every exchange 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 B.1). In the equilibria that we construct, investors' *routing table strategies*, i.e., how they break ties if indifferent across exchanges, serve to coordinate TFs' liquidity-provision decisions with investors' trade-routing decisions. 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. How trading activity is ultimately split across exchanges, however, is not pinned down: indeed, for any arbitrary split of market shares $\boldsymbol{\sigma^*} = (\sigma^*_1, \ldots, \sigma^*_M)$, there is an equilibrium in which each exchange j's share of depth and volume are each exactly σ^*_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. 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, $\frac{\Pi^*_{continuous}}{N}$. We analyze a specific deviation for fast TFs, the *lone-wolf deviation*, and show that it is the most attractive deviation to consider, hence ruling it out is sufficient. In the lone-wolf deviation, instead of purchasing ESST from all exchanges, a fast TF deviates and purchases ESST from just a single exchange in Stage 2, and then provides a single unit of liquidity in each Stage 3 trading game on this single exchange at a spread that is strictly narrower than $s^*_{continuous}$ (which we prove to be an equilibrium of the Stage 3 subgame; Lemma B.2). In doing so, the deviating TF attracts all trading volume to the one exchange on which they are fast. Since the deviating TF provides liquidity at a spread narrower than $s^*_{continuous}$ they earn an amount per trading game of strictly less than $\frac{\Pi^*_{continuous}}{N}$; Appendix B.3.2, equation (B.3) derives the amount earned explicitly, and shows that $\pi_N^{lone-wolf} \in (\frac{N-2}{N-1} \times \frac{\Pi^*_{continuous}}{N}, \frac{\Pi^*_{continuous}}{N})$. 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 $\frac{\Pi^*_{continuous}}{N}$ per trading game than purchasing ESST from just a single exchange and earning $\pi_N^{lone-wolf}$ per trading game.

Proposition 3.3 in Section 3.3 will use condition (3.2) to characterize an upper bound on exchanges' total rents from ESST. Intuitively, the bound requires that exchanges leave enough rent for fast TFs so as not to tempt them to deviate. It is worth emphasizing that if there were only a single exchange, the fast TFs could not leverage the lone-wolf deviation to play exchanges off against each other, and the single exchange would be able to extract the entire sniping prize $\Pi^*_{continuous}$ via ESST fees.

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, any exchange that raises its trading fee from zero gets zero share.

3.3 Discussion

We now discuss the three main features of the equilibria described in Proposition 3.2. In Section 4, we show that these features are consistent with patterns that we observe in the data.

Single Synthesized Exchange. Regulatory features of the U.S. equities market, specifically Reg NMS and UTP, support an environment where market participants can "stitch" together multiple exchanges into what we refer to as a *single synthesized exchange*. 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 $s^*_{continuous}$, 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. Indeed, as proven in Proposition 3.2, there exists a continuum of equilibria that supports *any* arbitrary vector of market shares.

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. Liquidity on an exchange with 20% volume share and 20% depth share enjoys 20% of the market's total benefit from providing liquidity to investors on the left-hand-side of (3.1), while incurring 20% of the market's total adverse selection and sniping cost on the right-hand-side of (3.1). An exchange with 10% volume and depth share enjoys 10% of the total benefit and 10% of the total cost. 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.

These results are closely related to Glosten (1994) and Ellison and Fudenberg (2003). Glosten (1994) considers a model with 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 (2003) 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.[25]

Similar to these other models, our model does not yield much insight into the determination of equilibrium exchange market shares. That said, our model 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 routing table strategies (see Appendix B.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 relatively stable over time 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 B.1 in Appendix B.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 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

[25]The "plateau" refers to an interval of market shares that can be sustained in equilibrium among platforms with the same seller-buyer ratio; outside of this interval, the only equilibria are those with complete tipping. This difference versus our model derives from the single-homing assumption in the Ellison and Fudenberg (2003) model (versus multi-homing in ours) and the way their model deals with integer issues (versus perfectly divisible shares in our model).

unlimited profits — i.e., a money-pump.[26] In Appendix B.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* $\frac{M}{(M-1)(N-1)}\Pi^*_{continuous}$.

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.[27] 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.[28] A strength of our approach is that it highlights that even if 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

[26] 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.

[27] 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.

[28] 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).

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 U.S. stock exchange competition. These facts relate to each of the three main results of Section 3's model of the status quo. Section 4.1 presents facts that relate to the model's equilibrium characterization of the Stage 3 trading game. Section 4.2 presents facts that relate to the model's equilibrium characterization of exchange trading fees. Section 4.3 presents facts that relate to the model's equilibrium characterization of exchange-specific speed technology fees. Section 4.4 provides discussion of the stylized facts taken in total, with reference both to our model which focuses on modern U.S. stock exchanges and to other previous models of financial exchange competition.

4.1 Evidence on the Stage 3 Trading Game

There are three main features of the multi-exchange trading game equilibria, characterized in Proposition 3.2 and 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, these exchange depth and volume shares can be interior and stable, i.e., there need not be tipping.

Before proceeding, we wish to acknowledge that none of the results in this section will be particularly surprising to a researcher familiar with modern U.S. equity market microstructure. However, we think they are useful to document carefully both because they provide empirical support for our stylized model of trading and because they rule out some other potential models of financial exchange competition.

Data. We use the Daily NYSE Trade and Quote ("TAQ") dataset accessed via Wharton Research Data Services. 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.[29] 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 together constitute 15% of trading volume, all use the "inverted" (or "taker-maker") pricing model in which the taker is paid a rebate and the maker pays a fee, and this difference in fee structure relative to the larger exchanges raises some subtleties for the analysis that we discuss in Appendix C.[30] Third, we restrict attention to the 100 most heavily traded stocks and ETFs.[31] In 2015, there were 9,175 symbols that traded at least once; however, most trade relatively infrequently. 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. These 100 symbols constitute about one-third of daily volume.[32]

Stylized Fact #1: Many Exchanges Simultaneously at the Best Bid and Best Offer. For each symbol i, exchange j, millisecond k, and date t, we compute the exchange's best bid and best offer (ask), denoted BB_{ijkt} and BO_{ijkt}. In case there are multiple quote updates in the symbol-exchange-millisecond, we use the last one. We then compute, for each symbol-millisecond-date, the number of exchanges at the overall best bid and best offer, i.e., we compute:

$$N_{ikt}^{bid} = \sum_j 1\{BB_{ijkt} = \max_{j' \in J} BB_{ij'kt}\} \qquad \text{and} \qquad N_{ikt}^{offer} = \sum_j 1\{BO_{ijkt} = \min_{j' \in J} BO_{ij'kt}\},$$

where J is the set of all exchanges.

As one might expect, the distributions of N_{ikt}^{bid} and N_{ikt}^{offer} are virtually identical, so we combine the data into a single distribution and present it as Figure 4.1. We present the results separately for NYSE-listed symbols and non-NYSE listed symbols. The reason for this difference is that non-NYSE listed symbols do not trade on NYSE (but do trade everywhere else), whereas NYSE listed symbols trade everywhere.[33] 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 (typically, listed on Nasdaq) the maximum is 4. As can be seen, the modal answer to the question "how many exchanges are at the best price?" is "all of them." For NYSE-listed symbols, all Top 5 exchanges

[29] 2015 was the most-recently available full year of data when we began presenting early versions of this research publicly. 2015 is also the best year in terms of data availability for the analysis of ESST revenues as will be described in Section 4.3.

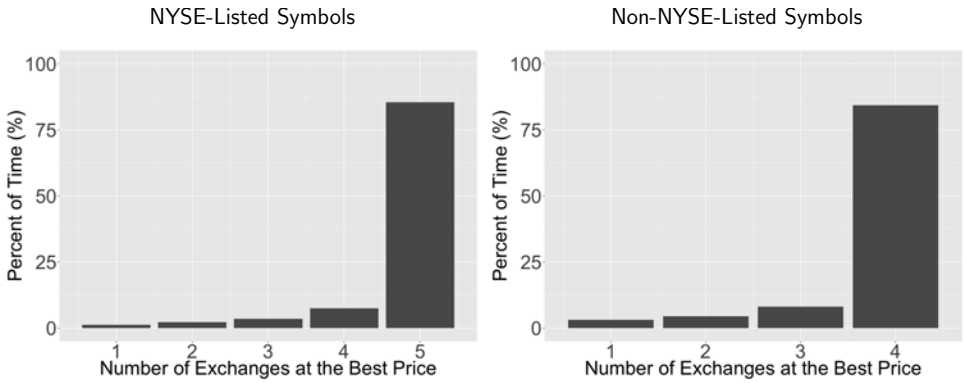
[30] The remaining 4 exchanges active during 2015, sometimes called the "regional" exchanges, together had about 2% market share. Anecdotally, industry participants regard them as vestiges of an earlier era of stock exchange competition.

[31] We use the phrase "symbols" to include both stocks and ETFs. When clear from the context we will sometimes use the phrase "stocks" to mean both stocks and ETFs.

[32] We have also conducted robustness tests in which we look at the top 1000 symbols by share volume that satisfy these filters except for the $10 million average daily volume filter, which constitutes roughly three-quarters of total volume. Results are qualitatively similar but with more noise.

[33] NYSE changed this practice as of April 2018 and began allowing non-NYSE listed stocks and ETFs to trade on NYSE. That is, NYSE recently exercised its right to extend Unlisted Trading Privileges to non-NYSE listed stocks.

Figure 4.1: Multiple Exchanges at the Same Best Price



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 C for details. Sample is 100 highest volume symbols that satisfy data-cleaning filters (see text for description) on all dates in 2015.

are at the best bid (similarly, best 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.

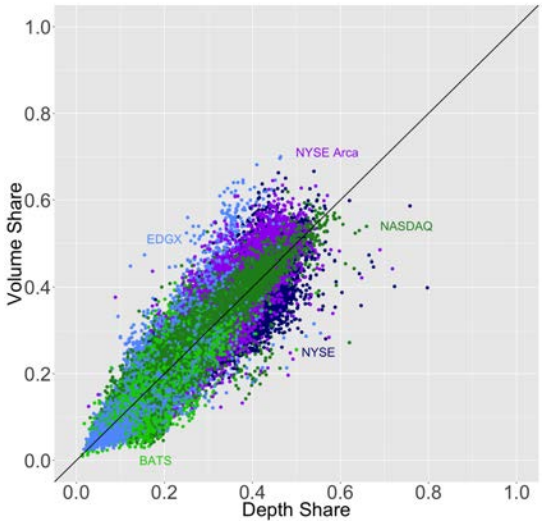
Stylized Fact 1. *At any given moment in time, for highly traded stocks and ETFs, the modal number of exchanges at the best bid and best offer is "all of them." Of the Top 5 exchanges, in about 85% of milliseconds all exchanges are at the best bid (similarly, best offer). It is rare (about 1% of milliseconds for NYSE-listed symbols and 3% for non-NYSE) for there to be just one exchange at the best bid or best offer.*

Stylized Fact #2: Depth Equals Volume. For each symbol i, exchange j, and date t, we compute the exchange's "depth share" and "volume share" for regular-hours trading in that symbol on that date. Volume share, $VolumeShare_{ijt}$, is calculated as the regular-hours volume in shares for symbol i on exchange j on date t divided by total regular-hours volume in symbol i on date t. We calculate depth share, $DepthShare_{ijt}$, by first computing depth for symbol i on exchange j at each millisecond k within the regular-hours trading period of day t, defined as

$$Depth_{ijtk} = \frac{q^{bid}_{ijtk} \cdot 1\{BB_{ijtk} = \max_{j' \in J} BB_{ij'kt}\} + q^{offer}_{ijtk} \cdot 1\{BO_{ijtk} = \min_{j' \in J} BO_{ij'kt}\}}{2},$$

where q^{bid}_{ijtk} and q^{offer}_{ijtk} denote the quantity at exchange j's best bid and offer for symbol i at millisecond k, and the indicator function requires that j's best bid or offer equals the national best at that mil-

Figure 4.2: 2015 Daily Volume Share vs. Depth Share



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. Since both depth and volume shares turn out to be relatively stable over time and across symbols (see Stylized Fact #3), we color code by exchange and label each exchange's cluster of dots. For details of share calculations and details of data-cleaning filters, see the text.

lisecond.[34] We then compute the average depth for each symbol-exchange-date by averaging $Depth_{ijtk}$ over all milliseconds, then calculate $DepthShare_{ijt}$ as this average depth divided by the sum of the average depth for each symbol-exchange-date across exchanges. Figure 4.2 presents a scatterplot of $VolumeShare_{ijt}$ against $DepthShare_{ijt}$, wherein each dot represents a symbol-exchange-date tuple. We color code by exchange and label each exchange's cluster of dots.

The figure shows that the depth-volume data falls along the 45 degree line for the Top 5 exchanges. The slope of a regression of volume share on depth share is 0.991 (s.e. 0.020), and the R^2 of the relationship is 0.865.[35] In robustness tests, we found that the depth-volume relationship along the 45 degree line obtains at significantly higher frequencies than a day, such as 5 minutes (albeit with more noise), but that at frequencies such as 1 second or 1 millisecond the relationship is not meaningful.[36] This stems from the fact that, at the level of an individual trade, exchange volume shares are often 0% or 100%, so the depth-volume relationship is only meaningful with some aggregation.[37]

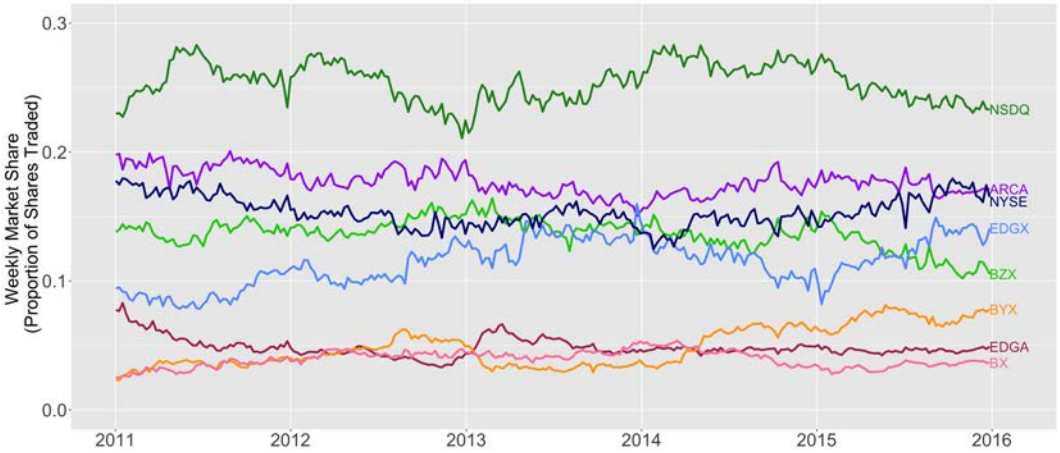
[34] We use all milliseconds between a symbol's first quote at or after 9:30 and 16:00 (13:00 on half-days), dropping any milliseconds where the NBBO is locked or crossed.

[35] As a robustness test we looked at the depth-volume relationship for each symbol in our data, running 100 regressions of daily exchange market shares on daily exchange depth shares, one for each symbol. The regression coefficients are very close to one (mean 0.991, st. dev. 0.026) and the R^2 of the relationship is high (mean 0.840, st. dev. 0.136), suggesting that the depth-volume relationship holds at the individual symbol level.

[36] The R^2 of the regression of volume share on depth share is 0.531 at 5 minutes, 0.635 at 10 minutes, 0.745 at 30 minutes, and 0.788 at 1 hour. The regression coefficients are 0.951, 0.957, 0.963 and 0.965 (each statistically indistinguishable from 1).

[37] Our model assumes all investors demand exactly "1" unit of perfectly-divisible liquidity and in equilibrium exactly 1

Figure 4.3: Weekly 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.

Stylized Fact 2. *Among the Top 5 exchanges, all of which use the same maker-taker fee structure, there is a one-for-one relationship between depth share and volume share at the daily level. The coefficient from regressing volume share on depth share is 0.99 (statistically indistinguishable from 1) and the* R^2 *is 0.87.*

Stylized Fact #3: Exchange Market Shares are Interior and Relatively Stable. Figure 4.3 presents aggregate weekly exchange market shares from January 2011 to December 2015 for the Top 8 exchanges. We start the time period in 2011 since that is the first full year of data after the original BATS exchanges (BZX and BYX) and Direct Edge exchanges (EDGX and EDGA) were approved as exchanges (prior to that they operated Alternative Trading Systems, or ATS's). Figure C.3 in Appendix C presents exchange market shares from October 2007, the start of the Reg NMS era, through the end of 2015.

As can be seen in the figure, aggregate exchange market shares are certainly interior, with no exchange's market share ever rising above 30%. Aggregate exchange market shares are also relatively stable in the sense that in the 2011-2015 period, if we regress s_{jt}, the market share of exchange j on date t, on a set of exchange fixed effects but nothing else, the R^2 is 0.967. Appendix C (see Figure C.4) presents related results at the individual-symbol level; as at the exchange level, shares are certainly

unit of liquidity is offered across exchanges so investors must spread their demand across multiple exchanges. In reality, investors of course 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, which we know from Stylized Fact #1 will not be consistent with depth shares. 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.

interior and are relatively stable.[38]

Stylized Fact 3. *Exchange market shares are interior at both the aggregate level and the individual-symbol level. Exchange market shares are also relatively stable in the sense that simple exchange fixed effects explain about 97% of the aggregate-level variation and about 76% of the individual-symbol level variation.*

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.

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. All fee schedules are from 2015 for consistency with the other analyses; the specific months range from Feb to Sept depending on the Internet Archive's coverage.[39]

Second, we use exchange company financial filings that cover 2015; specifically 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). It is important to clarify that exchange companies each control several exchanges, and while the fee schedules mentioned above are at the exchange level, most of the financial data in the annual report is at the exchange company level. For example, in 2015 the exchange company BATS, Inc., controlled four exchanges, two maker-taker exchanges (BZX and EDGX) and two taker-maker exchanges (BYX and EDGA).

Stylized Fact #4: Trading Fees are Economically Small. 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" for a given exchange. For example, for BATS's maker-taker exchange BZX, takers of liquidity (i.e., the submitter of an order that trades against a resting bid or offer) pay a fee of \$0.0030, while makers of liquidity (i.e., the resting bid or offer) earn a rebate of between \$0.0020 – \$0.0032, where the exact rebate is determined by the maker's volume on BZX (higher volume participants receive larger

[38] Among all symbols in our sample, the single highest average exchange market share in 2015 is less than 40%. Additionally, average daily symbol-exchange market shares are also relatively stable in the sense that if we regress s_{ijt}, the market-share of symbol i on exchange j on date t, on a set of exchange fixed effects and control for whether or not the symbol is listed on NYSE but nothing else, the R^2 is 0.76. Overall, the results at the individual stock level tell the same story of interior and relatively stable shares as the aggregate results, just with more noise.

[39] Typically, current exchange fee schedules are posted on exchanges' websites, while changes to exchange fee schedules are filed with the SEC and accessible via the SEC's website. Since the fee schedules can be so complicated, it can be difficult to build out the full fee schedule from the fee-change filings posted permanently on the SEC website; therefore we use fee schedules accessed directly from exchange websites via the Internet Archive.

rebates). BZX's net fee per-share per-side therefore ranges, based on the participants in a trade, from -$0.0001 to +$0.0005, or "-1 to +5 mills" in the industry jargon (1 mill = $0.0001); this can be thought of as the observed range for what our model calls f.[40]

Table 4.1 Panel A presents the observed range for per-share per-side trading fees for the top 8 exchanges. The table focuses on each exchange's most representative fee schedule, with additional details for special fee programs like the NYSE Supplemental Liquidity Provider program presented in Appendix Table D.1. As can be seen, many of the exchanges have minimum fees on a per-share per-side basis that are actually slightly negative. The complete table in the Appendix shows that 7 of the 8 exchanges have a minimum per-share per-side fee that is negative, with 4 having a negative minimum fee based on a pure volume threshold and an additional 3 with negative minimum fees based on participation in a special fee program. The maximum fee per-side is always strictly positive and typically about 5 mills, though it is noticeably lower for the two BATS taker-maker exchanges (1 mill for BYX, 1.5 mills for EDGA) and higher for the Nasdaq taker-maker exchange (8 mills).

To get a more precise estimate for average per-share per-side trading fees — that is, average f — we use the major exchange families' annual financial filings. The advantage of using annual financial filings is that we can estimate the average regular-hours trading fee, not just the potential range. There are two disadvantages. First, we have to conduct this analysis at the level of the exchange family, not the individual exchange. Second, we have to make some assumptions about fees from non-regular trading (e.g., opening and closing auctions, routed volume) to get to an estimate for regular hours per-share per-side trading fees. These disadvantages in mind, the results are presented in Table 4.1 Panel B; supporting details are provided in Appendix D.2.

As can be seen, the average trading fee across the 3 major exchange families is about $0.0001 per-share per-side, or 1 mill. 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 4. *Exchange trading fees are economically small. While there is no single number for what our model calls f, 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%.*

[40] In addition to these fees for standard regular-hours trading, there are also dozens of other fees for orders that are routed to other exchanges, executed in the opening or closing auctions, etc. Both NYSE and Nasdaq have fee schedules that differ slightly based on whether the stock being traded is listed on NYSE or Nasdaq.

Table 4.1: U.S. Equity Exchange Trading Fees ("f")

Panel A: Range of Fees Per Share

		Taker Fee		Maker Fee		Total fee per share per side	
Exchange	Fee Type	Min	Max	Min	Max	Min	Max
NASDAQ	Maker-Taker	0.00300	0.00300	-0.00325	-0.00150	-0.00013	0.00075
BATS BZX	Maker-Taker	0.00300	0.00300	-0.00320	-0.00200	-0.00010	0.00050
EDGX	Maker-Taker	0.00300	0.00300	-0.00320	-0.00200	-0.00010	0.00050
NYSE	Maker-Taker	0.00270	0.00270	-0.00220	-0.00140	0.00025	0.00065
NYSE Arca	Maker-Taker	0.00280	0.00300	-0.00270	-0.00200	0.00005	0.00050
BATS BYX	Taker-Maker	-0.00160	-0.00160	0.00140	0.00180	-0.00010	0.00010
EDGA	Taker-Maker	-0.00020	-0.00020	0.00030	0.00050	0.00005	0.00015
NASDAQ BX	Taker-Maker	-0.00150	-0.00040	0.00165	0.00200	0.00008	0.00080

Panel B: Estimate of Average Trading Fees

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

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 D.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 D.2 and the associated spreadsheet for supporting details for these calculations.

Stylized Fact #5: Money-Pump Constraint Binds. 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. 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. Because the SEC fee is assessed based on dollar volume, the sum of SEC and FINRA fees on a per-share basis increases with the nominal share price.

For the purpose of calculating the money-pump boundary, we should 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. For a \$5 stock this would be 1.14 mills, i.e., per-share per-side fees could go to -1.14 mills without creating a money pump. This may help explain why exchange trading fees are able to go slightly negative, as exhibited in Table 4.1 and in Appendix Table D.1, without creating a money pump.

Stylized Fact 5. *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, with the lowest observed fee being -$0.00015 per-share per-side (Nasdaq, BZX, EDGX, BYX). For another 3 of the 8 exchanges, 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 (NYSE, NYSE Arca, Nasdaq BX). 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 do not get negative enough to create a money pump once we account for SEC + FINRA fees, with the possible exception of very-low priced stocks.*

4.3 Evidence on Exchange-Specific Speed Technology Revenue

The last series of stylized facts is related 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 one's own servers next to Nasdaq's servers. Notably, our model does not pin down the exact level of ESST, but does indicate that, in aggregate across exchanges and trading firms, 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 the aggregate tape revenues (revenues that come from a data feed not used by latency sensitive traders), which we subtract for our main estimate of ESST revenues in Stylized Fact #6. We discuss specific details of the data in the text below and in Appendix E.

Stylized Fact #6: Exchanges Earn Significant Revenues from Co-Location/Connectivity and Proprietary Market Data. For estimating the overall magnitude of ESST revenues we focus on 2015. In its April 2016 S-1 filing (i.e., IPO prospectus), BATS directly reports financials for its U.S. equities business as a separate financial reporting segment, and within that reporting segment separately breaks out revenue from market data and co-location/connectivity. BATS was acquired by CBOE later in 2016 and following that acquisition no longer reported U.S. equities revenue with such granularity. Neither Nasdaq nor NYSE report U.S. equities revenue with the granularity of BATS in 2015, so for those exchanges we make some assumptions (described below) and we report a range.

BATS's 2015 market data revenues were $114.1M and its co-location/connectivity revenues were $64.3M, for a total of $178.4M. For context, its net transactions revenues were $81.0M and its operating expenses were $110.2M.[41] This means that the BATS U.S. Equities business is profitable with market

[41] Net transactions revenues are computed as Transaction Fees ($938.8M) less Liquidity Payments (i.e., rebates, $814.1M) less Routing and Clearing Fees ($43.7M).

data and co-location/connectivity revenues (profits before tax of $149.2M) but loss-making on the basis of trading revenues alone (loss of $29.2M).

For both Nasdaq and NYSE our exercise is less straightforward because neither firm breaks out its U.S. equities business as its own reporting segment. For NYSE a further complication is its Nov 2013 acquisition by Intercontinental Exchange (ICE). Our approach for Nasdaq utilizes market data and connectivity revenue figures for its global securities business, and information on the proportion of U.S. to global revenue, and information about the breakdown of U.S. revenue between equities and options. Our approach for NYSE utilizes information about NYSE's market data and connectivity revenue contained in ICE's 2014 10-K filing (i.e., the first fiscal year after the acquisition closed) plus additional information from ICE's 2015 filing. We provide a detailed description of our calculations for Nasdaq and NYSE in Appendix E.

Table 4.2 summarizes our analysis. Across all three major exchange families, we estimate $555.4-$623.0M for market data revenue and $436.8-$484.8M 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 (see footnote 12). 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. The Consolidated Tape Authority reports that in the 12 month period through March 2014, total tape revenues across all U.S. equities exchanges were $317M.[42] If we subtract this $317M from the total we have proprietary market data revenue of $238.4-306.0M, and total ESST revenue of $675.2-790.8M.[43]

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 6. *Exchanges earn significant revenue from exchange-specific speed technology. While the data reported in exchange parent company financial filings are not perfect, sensible assumptions applied to that data suggest that in 2015 total ESST revenue was between $675-790M. This is*

[42] The prices for consolidated feed data are set by a consortium, and then the revenues are allocated to exchanges based on a formula that relates to their volume share and NBBO depth share. Whereas proprietary data revenues appear to have grown significantly in the past decade or so, tape revenue growth appears to be much flatter. For example, Nasdaq's revenue from proprietary data increased more than 100% from 2006-2012 (the last year they reported it separately), whereas its tape revenues declined by about 10% during this same period. For this reason, we are comfortable using the March 2014 tape revenue number as part of our 2015 ESST revenue analysis. See the appendix for more details.

[43] 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. For example, if an exchange's average take fee is 30 mills and its average make rebate is 28 mills, we think of the fee revenue per share as 30-28=2 mills whereas Jones (2018)'s analysis of exchange revenues treats the fee per share as 30 mills, and implicitly treats the 28 mills rebate as a cost. Under this latter interpretation, revenues from trading fees are considerably larger than revenues from data, co-location, and connectivity. Additionally, revenues from data, co-location, and connectivity appear small as a fraction of total exchange revenues.

Table 4.2: Estimated Market Data and Co-Location Revenues for U.S. Equities Market in 2015 (Millions of Dollars)

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

Notes: 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, because the 2014 filing had more granular information on the contribution of the NYSE business to ICE's overall business, following its acquisition in Nov 2013. BATS directly reports a U.S. equities revenue breakdown including market data and co-location/connectivity revenue. 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 please consult the text and Appendix E.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. Refer to SEC Release No. 34-73278.

several times larger than regular-hours trading revenues.

Stylized Fact #7: Exchange Revenue from Co-Location/Connectivity and Proprietary Market Data Appears to have Grown Significantly in the Reg NMS Era. While exchange companies do not directly report U.S. equities ESST revenue (as evident from the work involved in Stylized Fact #6), we can get a sense of magnitudes for U.S. equities ESST revenue growth over time by looking at revenue growth in the financial reporting categories that contain U.S. equities ESST. We are able to build meaningful time-series for Nasdaq co-location and connectivity revenues and proprietary market data revenues from 2006 to 2017 and for BATS co-location and connectivity revenues from 2010 to 2017, using various financial filings which are outlined in detail in Appendix E.2.[44] BATS only began charging for the proprietary market data that we think of as part of ESST relatively recently (Q3 2014) and we discuss the limited data that is available in the appendix.[45] NYSE's financial reporting segments unfortunately changed too frequently in the Reg NMS era for the exercise to be instructive.[46]

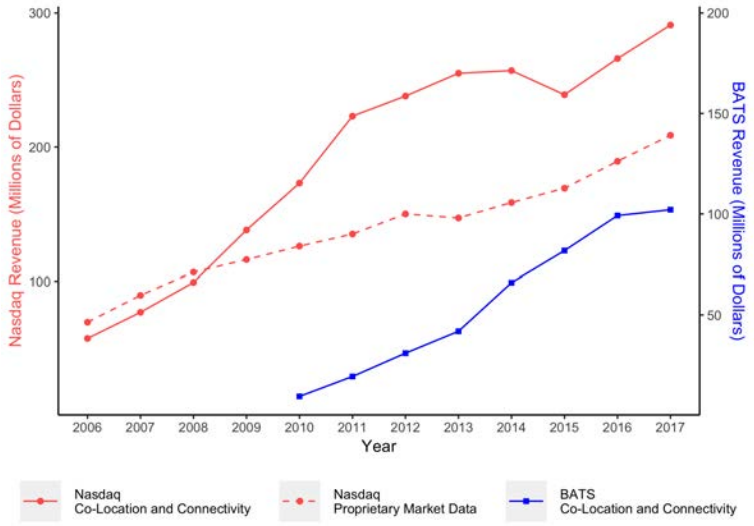
Figure 4.4 presents growth for the three series we construct. For Nasdaq co-location and connec-

[44]For Nasdaq, 2006 was both the year before Reg NMS was implemented and was the first year the word "co-location" appears in a Nasdaq annual financial filing (it has appeared every year since). For BATS, 2010 was the first full year that BATS started charging for co-location/connectivity.

[45]BATS reports that it only began charging for proprietary market data on the two original BATS exchanges, BZX and BYX, in the 3rd quarter of 2014. Triangulating between the various data sources, it appears that BATS proprietary market data revenue has been growing rapidly since it began charging for it, but that the overall revenues are still much smaller than BATS's co-location/connectivity revenue, which has been growing rapidly for five additional years.

[46]This was due to NYSE's merger with Euronext in 2007, its acquisition by ICE in 2013, and some financial segment reporting changes in the period in between. This caveat in mind, the category "Technology Services Revenues" which includes co-location/connectivity grew from $137M to $341M over the period 2006-2012, and the category "Market Data Revenues" grew from $223M to $348M over this time period.

Figure 4.4: Co-Location/Connectivity and Proprietary Market Data Revenue: 2006-2017



Notes: Nasdaq data come from 2006-2017 10-K filings. BATS co-location/connectivity revenue data come from the 2012 S-1 filing (years 2010–2011), the 2016 S-1 filing (2012–2015), the 2016 CBOE/BATS Merger Proxy and the CBOE 2017 10-K. We omit BATS proprietary market data revenue from the figure because BATS only started charging for proprietary data in Q3 2014. Please see the text and Appendix E.2 for further discussion of all of these data.

tivity, revenue in the reporting category quadrupled in the 2006-12 period (growth of 26.7% per year), was roughly flat in the 2012-2015 period which contained some minor reporting category changes, and then in 2015-2017 growth was 10.3% per year. For Nasdaq proprietary market data, revenue growth was 13.7% annually for the period 2006-2012. Starting in 2013, Nasdaq made some changes to the reporting category containing market data: U.S. proprietary data revenue was directly reported from 2006-2012, but was then combined with U.S. tape plan revenue starting in 2013, then combined with U.S. tape revenue and global data revenue starting in 2014. We deduct estimates of U.S. tape plan revenue and global data revenue to make the revenue estimates from the 2013-2017 comparable with the figures in filings from 2006-2012. We estimate that Nasdaq proprietary market data growth was 6.8% for the period 2012-2017, with the caveat that this 2012-2017 growth rate is based on some assumptions whereas the 2006-2012 growth rate is based off of numbers directly in Nasdaq filings (see Appendix E.2 for details). For BATS co-location and connectivity, revenue more than quadrupled (growth of 64.0% per year) from 2010 through 2013, the last full year before BATS's acquisition of Direct Edge (which combined EDGX and EDGA under the same exchange company as BZX and BYX). Revenue then doubled from 2013 to 2015, but likely in large part due to the Direct Edge acquisition in 2014, and then grew 11.7% per year from 2015 to 2017. Please see Appendix E.2 for full details on the data and growth figures.

Overall the data, while imperfect, are suggestive of exchanges "discovering a new pot of gold" in the Reg NMS era — that is, discovering that they could charge significant money 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. If, as in Stylized Fact #6 above, we take the lone-wolf bound from the theory seriously and assume that exchanges extract at most 30% of latency arbitrage rents, this range suggests a lower bound on the total size of the latency-arbitrage pie for U.S. equities of $3.0 billion in 2018.

Stylized Fact 7. *Exchanges' revenues from exchange-specific speed technology appear to have grown significantly in the Reg NMS era. With the caveat that the data are imperfect, we compute overall annual growth rates of: 15.9% for Nasdaq co-location/connectivity (2006-2017), 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 the 2015 ESST figures reported in Stylized Fact 6, this implies annual ESST revenue in 2018 on the order of $1 billion per year.*

4.4 Discussion of Model Fit and Alternative Models

We now discuss other potential models of exchange competition which 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 exchange competition and not specifically the modern U.S. stock market.

The first class of models are those in which some market participants "single home," thereby generating exchange-specific network effects. One example is the classic model by Pagano (1989), who when motivating his single-homing model, insightfully noted that if traders could frictionlessly multi-home and arbitrage across markets, "the two markets would collapse into a single one, and the choice between the two would be vacuous" (pg. 260). A modern example of a single-homing model is that of Pagnotta and Philippon (2018), who allow for exchanges to compete on the overall technological sophistication of their exchange — modeled as the frequency of trading opportunities, which they call "competing on speed" — in an effort to attract traders to single-home on their exchange as opposed to the competition. In contrast, the speed in our model enables some market participants to be faster than other market participants on the same exchange.[47] Relatedly, Cantillon and Yin (2008) consider a model in which participants can multi-home but the financial instruments (in their case, futures contracts) are specific to a single exchange — i.e., assets are not fungible across exchanges — which also generates exchange-specific network effects. 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. For example, in Pagano (1989) tipping is the only equilibrium if transactions fees are the same across exchanges, and Cantillon and Yin (2008) are motivated by the "Battle of the Bund," a famous real-world example of market tipping.

[47] See also Cespa and Vives (2019) who model speed in a similar fashion to Pagnotta and Philippon (2018) by allowing exchanges to sell technology which allows market participants to trade in both periods of a two-period Walrasian trading game as opposed to just one. Cespa and Vives (2019) then study the Cournot equilibria of a game among exchanges in which they strategically choose their technological capacity for such two-period participants.

Another class of models are those in which exchanges are meaningfully differentiated. This includes Pagnotta and Philippon (2018), discussed above, in which exchanges are vertically differentiated, as well as Baldauf and Mollner (2019), in which exchanges are horizontally differentiated. Baldauf and Mollner (2019) consider a model in which exchanges are located on a Salop circle (to capture horizontal differentiation), the size of the latency arbitrage pie increases with the number of exchanges, and the social planner trades off the benefits of increased competition from more exchanges (i.e., lower trading fees) against the cost of increased latency arbitrage.[48] Such differentiation allows exchanges to charge supra-competitive trading fees, which is inconsistent with Stylized Facts #4-#5. Also, such models suggest segmentation of market participants and securities across venues, which is at odds with Stylized Facts #1-#3.

Third, Chao, Yao and Ye (2019) provide a model in which tick-size frictions are central to understanding exchange fragmentation and competition. Their key point is that exchanges can use differential fee structures to enable trading firms to provide liquidity at slightly different net-of-fee prices across exchanges, which both "fills in the penny" for the market (i.e., makes tick-size constraints less binding) and gives exchanges market power. However, while the model helps explain the coexistence of the maker-taker fee model and the taker-maker fee model, the 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 showed in Stylized Fact #4, Table 4.1 — in the Chao, Yao and Ye (2019) model, the way to maximize economic profits is to have as different a fee structure as possible from all other exchanges, and exchanges do not have a source of economic profits beyond trading fees.

5 Incentives for 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 (i.e., interior as opposed to tipping), 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.

[48] In our model, the size of the latency arbitrage pie does not grow with the number of exchanges but in principle it could if either (i) aggregate market depth increases with the number of exchanges (as in Baldauf and Mollner, 2019); or (ii) some investors are unable to synchronize their trading across exchanges, allowing for the possibility that high-frequency trading firms, detecting an investor's trade on one exchange, may be able to "front run" on other exchanges (as in Baldauf and Mollner, 2020). If the latency-arbitrage pie grows with the number of exchanges, that should only strengthen the arguments we make in Section 5.

In this section, we use our theoretical model to examine exchanges' incentives for market design innovation. Our discussion will focus on frequent batch auctions 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.[49] Formally, for our theoretical analysis, we assume that the alternative market design eliminates latency arbitrage 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 frequent batch auction exchange and one or more continuous limit order book exchanges. Section 5.3 analyzes equilibrium if there are multiple frequent batch auction 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. We discuss the policy implications of our analysis in Section 6.

5.1 Model Details

We first briefly describe the frequent batch auction (FBA) market design proposed and analyzed in Budish, Cramton and Shim (2015), and then discuss how we incorporate it formally into our model of exchange competition introduced in Section 3.

Brief Description of Frequent Batch Auctions. The FBA market design 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

[49] 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. Please see Section VIII.C-D of Budish, Cramton and Shim (2015) for a discussion of asymmetric speed bumps, and please see Baldauf and Mollner (2020) for a detailed theoretical analysis.

a uniform-price auction, rather than being serially processed upon receipt as in a limit order book. 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.[50] 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. 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 modify the Stage 3 trading game of our model of exchange competition introduced in Section 3 as follows. We assume that an FBA exchange, which we call "Discrete" in the formal analysis, first processes all cancelations 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 the Continuous case, 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. In particular, in Period 1 of the trading game, TFs have an opportunity to submit limit orders, IOC orders, and cancel messages to all exchanges; after these orders are processed, the resulting order book on each exchange is displayed publicly as part of the state $\boldsymbol{\omega}$. In Period 2, if an investor or informed trader arrives, they have a single opportunity to send IOC orders to all exchanges. If in Period 2 there is public information, TFs can respond as before by sending IOC orders and cancel messages to all exchanges, but because of the market design there will not be any latency arbitrage rents.

The only additional modification we make to our model is that Discrete exchanges do not sell exchange-specific speed technology in Stage 1. 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 prices commensurate with their role, on continuous exchanges, in extracting sniping rents.[51]

To summarize, in Stage 1, Continuous exchanges set both trading and ESST fees, while Discrete exchanges only set trading fees; in Stage 2, fast TFs make ESST adoption decisions for Continuous exchanges; and in Stage 3, the infinitely repeated trading game is played with the modifications

[50] 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.

[51] 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, 2015). 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.

described above. As before, our equilibrium concept is subgame perfect Nash equilibrium for Stages 1 and 2, and order book equilibrium for Stage 3 period 1. In Stage 3 period 2, investors and informed traders have essentially unique optimal strategies, as before. In the event of public information in Stage 3 period 2, TFs no longer have weakly dominant strategies as they did in a sniping race. We prove that in any Nash equilibrium of the subgame that arises when there is new public information, either all stale quotes are canceled or any trade that does occur is at the new value of y, i.e., the price reflects the new public information. Hence, there are no latency arbitrage rents. We assume that investors and informed traders play their optimal strategies and that TFs, in the event of public information, play Nash equilibrium.

Before proceeding, we emphasize that the way in which we model competition between FBA exchanges and the continuous market is only appropriate for sufficiently fast batch intervals, and would become more practically strained with a longer batch interval.[52] One reason is that for an FBA exchange to have protected quotes under Reg NMS, the batch interval must be less than 1 millisecond in order to satisfy the SEC's *de minimis* delay standard (U.S. Securities and Exchange Commission, 2016*c*; see Appendix A for further discussion). With a longer batch interval, the frictionless search and access assumption in our trading game may thus be inappropriate. A second reason is that with a longer batch interval, our assumption that at most one event happens per trading game (i.e., an investor arrival, informed trader arrival, or public news) becomes less realistic.[53]

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.[54] 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 $s^*_{continuous}$ given by (3.1): $\lambda_{invest}\frac{s^*_{continuous}}{2} = (\lambda_{public} + \lambda_{private}) \cdot L(s^*_{continuous})$. In contrast, if there was only a single Discrete exchange also charging zero trading fees,

[52] Practically, we think of the batch interval in this model as long enough for an exchange computer to effectively batch process in the event that there is public news and multiple TFs respond, but then otherwise essentially as fast as possible. Some industry participants have suggested to us that as little as 50 microseconds (i.e., 0.000050 seconds) would be sufficient for this purpose, and our sense from aggregating many similar conversations is that 500 microseconds to 1 millisecond would be more than sufficient. Recent empirical evidence in Aquilina, Budish and O'Neill (2020) is consistent with these magnitudes being enough to eliminate most latency arbitrage.

[53] As noted above in footnote 14, even for the highest activity symbols in the market, nearly all milliseconds have neither any trade nor change in the national best bid or offer. It is thus empirically reasonable for an investor or informed trader to assume that nothing else will happen in the millisecond that they are trading, unless their trading is itself responding to public news. Even at intervals like 1 second, which will sound very fast to an economist relative to the speed at which fundamentals evolve, this is no longer the case. Whereas at 1 millisecond, more than 97.6% of milliseconds have no trades or best bid or offer changes for SPY and more than 99.6% of milliseconds have no such activity for GOOG, at 1 second the figures are just 3.5% for SPY and 37.1% for GOOG. These numbers are based on 12 randomly selected trading days in 2018.

[54] 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 single synthesized exchange. 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 single synthesized exchange.

arguments developed in BCS imply that a single unit of liquidity would be provided in equilibrium at the spread $s^*_{discrete}$ which solves:

$$\lambda_{invest}\frac{s^*_{discrete}}{2} = \lambda_{private} \cdot L(s^*_{discrete}) \ . \tag{5.1}$$

The difference between (5.1) and (3.1) is the $\lambda_{public}L(s^*)$ term — this reflects that Discrete eliminates latency arbitrage rents, and hence the associated cost for liquidity providers. For this reason, $s^*_{discrete} < s^*_{continuous}$. Intuitively, there are two reasons why Discrete eliminates latency arbitrage. First, the discrete-time interval gives liquidity providers an opportunity to cancel stale orders in response to public news before new orders are processed at the end of the batch interval. Second, competition among trading firms in the batch auction ensures that transaction prices reflect public news, even if there are stale quotes that are not canceled by the end of the batch interval.

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 exist 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 equilibrium has the following properties. In period 1 of each trading game: exactly one unit of liquidity is provided on Discrete at bid-ask spread $s^*_{discrete}$ (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 $\frac{s^*_{discrete}}{2}$; if there is a publicly-observed jump in y that exceeds $\frac{s^*_{discrete}}{2}$, either all TFs with stale quotes cancel their stale quotes, or if the auction results in trade the auction price is the new value of y. Such an equilibrium of the trading game exists.*

To understand why liquidity cannot be offered on the Continuous exchange in equilibrium, first note that a liquidity provider must charge at least the "zero-variable profit spread" on Continuous, denoted $\bar{s}_{continuous}$ and given by the solution to $\lambda_{invest}\frac{\bar{s}_{continuous}}{2} - (\frac{N-1}{N}\lambda_{public} + \lambda_{private}) \cdot L(\bar{s}_{continuous}) = 0$.[55] This spread is strictly greater than $s^*_{discrete}$. As a result, since investor demand is perfectly elastic with respect to the bid-ask spread, if any liquidity provider on Continuous were (weakly) profitably offering liquidity at some spread $s \geq \bar{s}_{continuous}$, that provider could be (strictly) profitably undercut on Discrete at a strictly smaller spread $s' \in (s^*_{discrete}, s)$. Furthermore, any liquidity cannot be offered at any spread other than $s^*_{discrete}$ in equilibrium on Discrete: any greater, and it could be profitably

[55] This spread is smaller than $s^*_{continuous}$ given by (3.1) since it does not account for the opportunity cost of not sniping.

undercut by another TF; any lower, and the liquidity provider would be losing money and be better off withdrawing. We show that these 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 $f > 0$.[56]

Given these results, we establish the following:

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* $\frac{N-1}{N}\Pi^*_{continuous}$. *Such an equilibrium exists.*

In essence, Discrete is compensated for the elimination of the tax that latency arbitrage imposes on trading. As long as Discrete charges a 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 Proposition 9 of Glosten (1994), who 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 one derived in Proposition 3.2 with multiple Continuous exchanges:

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* $s^*_{discrete}$ *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 single synthesized exchange: a single unit of liquidity is

[56] Let $\bar{s}_{discrete}(f)$ (defined formally in the Appendix in (B.5)) denote the zero-variable-profit spread for a liquidity provider on Discrete when Discrete charges trading fee f, so that $s^*_{discrete} = \bar{s}_{discrete}(0)$. The proof of Proposition 5.1 establishes that if $\bar{s}_{discrete}(f)/2 + f < \bar{s}_{continuous}/2$ (so that an investor would prefer trading on Discrete at spread $\bar{s}_{discrete}(f)$ and paying a trading fee f to trading on Continuous at spread $\bar{s}_{continuous}$), any profitable provision of liquidity on Continuous could always be profitably undercut by liquidity provision on Discrete, and hence cannot occur in equilibrium.

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 $s^*_{discrete}$, not $s^*_{continuous}$, which is better for investors and informed traders because $s^*_{discrete} < s^*_{continuous}$. Second, there are no longer latency arbitrage rents for exchanges or trading firms.

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). It is straightforward to see 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 Π^D, where $\Pi^D \in (\frac{N-1}{N}\Pi^*_{continuous}, \Pi^*_{continuous})$, 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 B.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 Π^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 (all Continuous), Proposition 5.2 (a single Discrete, the remainder Continuous), or Proposition 5.3 (multiple Discrete). 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.[57] For the purpose of our analysis of adoption incentives below we will assume that if there is an initial adoption at least one incumbent will imitate.

5.5 Adoption Incentives

Given the prisoner's dilemma payoff structure as summarized in Section 5.4, the analysis of exchange adoption incentives is relatively standard.

Let c_{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_{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_{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_{adopt} and c_{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 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.[58]

Adoption Incentives: A New Entrant Exchange. If a de novo entrant starts a new Discrete exchange, the entrant earns revenues of Π^D per-trading game until it is imitated. Once imitated, which occurs after T periods, the entrant earns zero profits. 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.[59] The condition for a de novo to

[57] 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.

[58] In case helpful, this assumption can be formalized as follows. Let R denote the total revenue from non-latency-sensitive data and assume R is split proportional to market share. Let $c_{imitate}$ denote the fixed cost of imitation. Then for at least one incumbent j, their anticipated market share if they adopt Discrete, denoted σ_j^D, satisfies $\sigma_j^D R - c_{imitate} > 0$.

[59] In the terminology of Budish, Roin and Williams (2015) ρ is the ratio of an invention's expected monopoly life to its expected total life, or $\frac{EML}{ETL}$.

find entry profitable is thus:

$$\rho\Pi^D \geq c_{adopt} + c_{entry}. \tag{5.2}$$

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

$$\rho\Pi^D \geq c_{adopt} + \underbrace{NF_j^*}_{\text{(status-quo rents)}}. \tag{5.3}$$

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_{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^*. 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_{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:

> "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].*"[60]

[60]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 was available for 5 years at https://capital-markets.law.columbia.edu/content/fellow-workshops. A copy of the video is available via the internet wayback machine at https://web.archive.org/web/20170418174002/https://www.law.columbia.edu/capital-markets/previous-workshops/2013 (accessed on Jan 8, 2019).

(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 in the continuous market.

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 is large (equations (5.2)-(5.3)). On the other hand, the analysis in Section 5 indicates that <u>if</u> 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.[61] Sections 6.1-6.2 discuss two 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., $c_{entry}+c_{adopt}$), 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_{adopt} and c_{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;[62] that is, costs that are part of what we call c_{entry}. As evidence of the significant

[61] 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 100% adoption but raises other issues that are difficult to model. Understanding the full tradeoffs between these two kinds of policy responses is beyond the scope of this paper.

[62] 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,

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 (CHX, 2017, 2018; U.S. Securities and Exchange Commission, 2017*c*), 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, 2016*c*). 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.

We can bound the maximum necessary subsidy by $c_{entry}+c_{adopt}$. The anecdotal evidence described above suggests this is on the order of $100 million.

6.2 Policy Response 2: Modest exclusivity period.

Examining equations (5.2) and (5.3), a key parameter that determines whether the innovator has sufficient incentive is ρ. The parameter ρ 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 ρ being 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.[63]

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

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.

[63] 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"). The SEC approved the proposal in May 2017, citing the similarity to IEX ("... the Commission does not find any legal basis to distinguish the Exchange's proposed Delay Mechanism from the IEX access delay.") Please see NYSE MKT LLC (2017) and U.S. Securities and Exchange Commission (2017*b*), respectively for details. Please see footnote 7 and Budish (2016*b*) regarding the strengths and limitations of IEX's speed bump design.

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.[64]

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 by the SEC 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 (see 21 CFR § 314.108 2018 and Food and Drug Administration, 2015). The purpose of the FDA policy is to induce drug companies to go through the effort and expense of the FDA clinical trials necessary to bring a new drug to market. Analogously, the purpose of the SEC exclusivity period would be to induce an exchange company to go through the effort and expense of the SEC approval process, and the other costs associated with developing and implementing a new market design.

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.

[64] 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).

This magnitude suggests that an entry subsidy almost surely passes a cost-benefit test. Getting a sense of magnitudes for the exclusivity period is more involved. In Appendix F, 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. For this calculation, we use estimates of the magnitude of latency arbitrage from Aquilina, Budish and O'Neill (2020) and, in a stylized manner, take account of frictions that are outside of the analysis in Section 5. Specifically, we consider tick-size constraints, agency frictions between investors and brokers trading on their behalf, and investors and liquidity providers not being perfectly elastic with respect to prices net of fees. Though a richer model and a more direct estimate of the magnitude of latency arbitrage in the U.S. stock market would be needed to get to a number with more confidence, we hope this exercise at least gives a sense of magnitudes that a relatively modest exclusivity period could be sufficient to induce entry.

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 in ways that 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 (see U.S. Securities and Exchange Commission, 2019*a*,*b*). Suspending UTP is a way of creating exclusivity for the innovator, analogous to our ideas in Section 6.2.[65] In February 2020, the SEC issued a proposed reform to the market for exchange data (U.S.

[65]Since suspending UTP would make trading exclusive to a single exchange per stock (the listing exchange) it would be important to think carefully about the resulting trading fees. In our analysis of equilibrium with a single Discrete exchange (Proposition 5.2), trading fees on the Discrete exchange are disciplined by competition from the Continuous exchange. Without UTP, the discipline on trading fees could come either from competition among listing exchanges (i.e., if a listing exchange raises its trading fee too much, listed companies can switch exchanges), competition from off-exchange trading venues, or explicit fee caps.

Securities and Exchange Commission, 2020*a*). 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."[66] (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.[67] 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.[68] 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).

[66]The speech also specifically cited our theoretical finding that exchanges have market power in the sale of speed technology: "Important recent research shows that, even when the market for trading is perfectly competitive, exchanges can extract supra-competitive rents from selling speed technology in the form of proprietary data feeds."

[67]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, 2019*b*; 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. One industry analyst remarked "we've seen this playbook before—it's BATS 2.0" (Stafford, 2019). BATS and Direct Edge, like MEMX, also entered as continuous exchanges owned by consortia of market participants. In our model of the status quo, since exchanges earn positive profits from exchange-specific speed technology, there is an incentive to enter as another continuous exchange if the entrant has a way of obtaining market share. To analyze business-stealing entry more fully, a researcher would have to extend our model to study the determination of equilibrium exchange market shares.

[68]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, 2019*a*), even while its subsidiary, the New York Stock Exchange, has in the past opposed such innovations in equity markets.

References

21 CFR § 314.108. 2018. New Drug Product Exclusivity.

Abdulkadiroğlu, Atila, and Tayfun Sönmez. 2003. "School Choice: A Mechanism Design Approach." *American Economic Review*, 93(3): 729–747.

Abdulkadiroğlu, Atila, Nikhil Agarwal, and Parag A. Pathak. 2017. "The Welfare Effects of Coordinated Assignment: Evidence from the New York City High School Match." *American Economic Review*, 107(12): 3635–3689.

Agarwal, Nikhil, Itai Ashlagi, Eduardo Azevedo, Clayton R. Featherstone, and Ömer Karaduman. 2019. "Market Failure in Kidney Exchange." *American Economic Review*, 109(11): 4026–4070.

Akbarpour, Mohammad, Julien Combe, Yinghua He, Victor Hiller, Robert Shimer, and Olivier Tercieux. 2019. "Unpaired Kidney Exchange: Overcoming Double Coincidence of Wants without Money." Working Paper.

Allen, Jason, Robert Clark, Brent Hickman, and Eric Richert. 2020. "Resolving Failed Banks: Uncertainty, Multiple Bidding & Auction Design." Bank of Canada Staff Working Paper 2019-30.

Antill, Samuel, and Darrell Duffie. 2018. "Augmenting Markets with Mechanisms." NBER Working Paper No. 24146.

Aquilina, Matteo, Eric Budish, and Peter O'Neill. 2020. "Quantifying the High-Frequency Trading 'Arms Race': A New Methodology and Estimates." Financial Conduct Authority Occasional Paper No. 50.

Armstrong, Mark. 2006. "Competition in Two-Sided Markets." *The RAND Journal of Economics*, 37(3): 668–691.

Arrow, Kenneth. 1962. "Economic Welfare and the Allocation of Resources to Invention." In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, ed. A Conference of the Universities – National Bureau Committee for Economic Research and Committee on Economic Growth of the Social Science Research Council, 609-626. Princeton University Press.

Asquith, Paul, Andrea S. Au, Thomas Covert, and Parag A. Pathak. 2013. "The Market for Borrowing Corporate Bonds." *Journal of Financial Economics*, 107(1): 155–182.

Asquith, Paul, Thomas Covert, and Parag A. Pathak. 2019. "The Effects of Mandatory Transparency in Financial Market Design: Evidence from the Corporate Bond Market." NBER Working Paper No. 19417.

Athey, Susan, and Glenn Ellison. 2011. "Position Auctions with Consumer Search." *The Quarterly Journal of Economics*, 126(3): 1213–1270.

Ausubel, Lawrence M., Peter Cramton, and Paul R. Milgrom. 2006. "The Clock-Proxy Auction: A Practical Combinatorial Auction Design." In *Combinatorial Auctions.* , ed. Peter Cramton, Yoav Shoham and Richard Steinberg, Chapter 5, 115–138. MIT Press.

Baldauf, Markus, and Joshua Mollner. 2019. "Trading in Fragmented Markets." *Journal of Financial and Quantitative Analysis.* forthcoming.

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

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

Bhattacharya, Vivek, Gaston Illanes, and Manisha Padi. 2020. "Fiduciary Duty and the Market for Financial Advice." NBER Working Paper No. 25861.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan. 2017. "High Frequency Trading and the 2008 Short-Sale Ban." *Journal of Financial Economics*, 124(1): 22–42.

Budish, Eric. 2011. "The Combinatorial Assignment Problem: Approximate Competitive Equilibrium from Equal Incomes." *Journal of Political Economy*, 119(6): 1061–1103.

Budish, Eric. 2016*b*. "Re: Investors' Exchange LLC Form 1 Application (Release No. 34-75925; File No. 10-222)." Retrieved December 22, 2018 from `https://www.sec.gov/comments/10-222/10222-371.pdf`.

Budish, Eric. 2016*c*. "Re: Proposed Commission Interpretation Regarding Automated Quotations Under Regulation NMS (Release No. 34-77407; File No. S7-03-16)." Retrieved January 9, 2019 from `https://www.sec.gov/comments/s7-03-16/s70316-12.pdf`.

Budish, Eric, Benjamin N. Roin, and Heidi L. Williams. 2015. "Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials." *American Economic Review*, 105(7): 2044–85.

Budish, Eric, Gérard P. Cachon, Judd B. Kessler, and Abraham Othman. 2017. "Course Match: A Large-Scale Implementation of Approximate Competitive Equilibrium from Equal Incomes for Combinatorial Allocation." *Operations Research*, 65(2): 314–336.

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

Bulow, Jeremy, and Paul Klemperer. 2013. "Market-Based Bank Capital Regulation." Working Paper.

Bulow, Jeremy, and Paul Klemperer. 2015. "Equity Recourse Notes: Creating Countercyclical Bank Capital." *The Economic Journal*, 125(586): 131–157.

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

Cantillon, Estelle, and Pai-Ling Yin. 2008. "Competition between Exchanges: Lessons from the Battle of the Bund." CEPR Discussion Papers No. 6923.

Cantillon, Estelle, and Pai-Ling Yin. 2011. "Competition between Exchanges: A Research Agenda." *International Journal of Industrial Organization*, 29(3): 329–336.

Cespa, Giovanni, and Xavier Vives. 2019. "Exchange Competition, Entry, and Welfare." Working Paper.

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

CHX. 2017. "Re: File No. SR-CHX-2016-16; Self-Regulatory Organizations; Chicago Stock Exchange, Inc.; Notice of Filing of Proposed Rule Change to Adopt the CHX Liquidity Taking Access Delay (Release No. 34-78860; File No. SR-CHX-2016-16)." Retrieved February 25, 2019 from `https://www.sec.gov/comments/sr-chx-2016-16/chx201616-1559194-131519.pdf`.

CHX. 2018. "Re: Release No. 34-82034; In the Matter of the Chicago Stock Exchange, Inc. ("Exchange") - For an Order Granting the Approval of Proposed Rule Change to Adopt the CHX Liquidity Enhancing Access Delay ("LEAD") on a Pilot Basis (File No. SR-CHX-2017-04)." Retrieved February 25, 2019 from `https://www.sec.gov/comments/sr-chx-2017-04/chx201704-4118079-171622.pdf`.

Clayton, Jay. 2018. "Statement on Market Data Fees and Market Structure." October 16. Public Statement. Retrieved January 4, 2019 from `https://www.sec.gov/news/public-statement/statement-chairman-clayton-2018-10-16`.

Collard-Wexler, Allan, Gautam Gowrisankaran, and Robin S. Lee. 2019. "'Nash-in-Nash' Bargaining: A Microfoundation for Applied Work." *Journal of Political Economy*, 127(1): 163–195.

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

Crunchbase. 2018. "IEX Group." Retrieved December 22, 2018 from `https://www.crunchbase.com/organization/iex`.

Duffie, Darrell, and Haoxiang Zhu. 2016. "Size Discovery." *The Review of Financial Studies*, 30(4): 1095–1150.

Duffie, Darrell, and Piotr Dworczak. 2018. "Robust Benchmark Design." NBER Working Paper No. 20540.

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

Edelman, Benjamin, Michael Ostrovsky, and Michael Schwarz. 2007. "Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords." *American Economic Review*, 97(1): 242–259.

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

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

Engers, Maxim, and Luis Fernandez. 1987. "Market Equilibrium with Hidden Knowledge and Self-Selection." *Econometrica*, 55(2): 425–439.

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

Farrell, Joseph, and Paul Klemperer. 2007. "Coordination and Lock-In: Competition with Switching Costs and Network Effects." In *Handbook of Industrial Organization*, vol. 3, ed. Mark Armstrong and Robert Porter. Elsevier B.V.

Food and Drug Administration. 2015. "Patents and Exclusivity." Retrieved January 9, 2019 from `https://www.fda.gov/downloads/drugs/developmentapprovalprocess/smallbusinessassistance/ucm447307.pdf`.

Fox, Merritt B., Lawrence R. Glosten, and Gabriel V. Rauterberg. 2015. "The New Stock Market: Sense and Nonsense." *Duke Law Journal*, 65(2): 191–277.

Fox, Merritt B., Lawrence R. Glosten, and Gabriel V. Rauterberg. 2019. *The New Stock Market.* Columbia University Press.

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

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

Glosten, Lawrence R. 2020. "Economics of the Stock Exchange Business: Proprietary Market Data." *SSRN*. Available at SSRN: `https://ssrn.com/abstract=3533525`.

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

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

Hall, Jonathan D. 2018. "Pareto Improvements from Lexus Lanes: The Effects of Pricing a Portion of the Lanes on Congested Highways." *Journal of Public Economics*, 158: 113–125.

Handel, Benjamin, Igal Hendel, and Michael D. Whinston. 2015. "Equilibria in Health Exchanges: Adverse Selection Versus Reclassification Risk." *Econometrica*, 83(4): 1261–1313.

Hendershott, Terrence, and Ananth Madhavan. 2015. "Click or Call? Auction versus Search in the Over-the-Counter Market." *The Journal of Finance*, 70(1): 419–447.

Hendershott, Terrence, and Haim Mendelson. 2000. "Crossing Networks and Dealer Markets: Competition and Performance." *Journal of Finance*, 55(5): 2071–2115.

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

Hortaçsu, Ali, Jakub Kastl, and Allen Zhang. 2018. "Bid Shading and Bidder Surplus in the US Treasury Auction System." *American Economic Review*, 108(1): 147–169.

Hosman, Bernard, Sean Castette, Fred Malabre, Pearce Peck-Walden, and Ari Studnitzer. 2017. "Mitigation of Latency Disparity in a Transaction Processing System." US Patent Application No. 14991654. Publication No. 20170046783A1.

IEX. 2015. "Re: Investors' Exchange LLC Form 1 Application (Release No. 34-75925; File No. 10-222)." Retrieved January 5, 2019 from `https://www.sec.gov/comments/10-222/10222-26.pdf`.

Jackson, Robert J., Jr. 2020. "Statement on Reforming Stock Exchange Governance." January 8. Public statement. Retrieved June 30, 2020 from `https://www.sec.gov/news/public-statement/statement-jackson-open-meeting-2020-01-08`.

Joint Staff Report. 2015. "The U.S. Treasury Market on October 15, 2014." *U.S. Department of the Treasury, Board of Governors of the Federal Reserve System, Federal Reserve Bank of New York, U.S. Securities and Exchange Commission, and U.S. Commodity Futures Trading Commission*. Retrieved on May 6, 2019 from `https://www.treasury.gov/press-center/press-releases/Documents/Joint_Staff_Report_Treasury_10-15-2015.pdf`.

Jones, Charles M. 2013. "What Do We Know About High-Frequency Trading?" Columbia Business School Research Paper No. 13-11.

Jones, Charles M. 2018. "Understanding the Market for U.S. Equity Market Data." Working Paper.

Kapor, Adam J., Christopher A. Neilson, and Seth D. Zimmerman. 2020. "Heterogeneous Beliefs and School Choice Mechanisms." *American Economic Review*, 110(5): 1274–1315.

Kastl, Jakub. 2017. "Recent Advances in Empirical Analysis of Financial Markets: Industrial Organization Meets Finance." In *Advances in Economics and Econometrics: Eleventh World Congress*, vol. 2, ed. Bo Honoré and Ariel Pakes and Monika Piazzesi and Larry Samuelson, 231-270. Cambridge University Press.

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

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

Kyle, Albert S., and Jeongmin Lee. 2017. "Toward a Fully Continuous Exchange." *Oxford Review of Economic Policy*, 33(4): 650–675.

Kyle, Albert S., Anna A. Obizhaeva, and Yajun Wang. 2018. "Smooth Trading with Overconfidence and Market Power." *The Review of Economic Studies*, 85(1): 611–662.

Levine, Matt. 2019. "Traders Want Their Own Exchange Too." *Bloomberg Opinion*, January 7. Retrieved December 17, 2019 from `https://www.bloomberg.com/opinion/articles/2019-01-07/traders-want-their-own-exchange-too`.

Levin, Jonathan, and Andrzej Skrzypacz. 2016. "Properties of the Combinatorial Clock Auction." *American Economic Review*, 106(9): 2528–51.

Lewis, Michael. 2014. *Flash Boys.* W. W. Norton and Company.

Madhavan, Ananth. 2000. "Market Microstructure: A Survey." *Journal of Financial Markets*, 3(3): 205–208.

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

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

Michaels, Dave, and Alexander Osipovich. 2018. "NYSE in Talks to Buy Chicago Stock Exchange." *Wall Street Journal*, March 30. Retrieved December 17, 2019 from `https://www.wsj.com/articles/nyse-in-talks-to-buy-chicago-stock-exchange-1522429813`.

Milgrom, Paul R., and Ilya Segal. 2019. "Clock Auctions and Radio Spectrum Reallocation." *Journal of Political Economy*. forthcoming.

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

NYSE MKT LLC. 2017. "Proposal to amend Rules 7.29E and 1.1E to provide for a Delay Mechanism." Retrieved December 16, 2019 from `https://www.nyse.com/publicdocs/nyse/markets/nyse-american/rule-filings/filings/2017/NYSEMKT-2017-05.pdf`.

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

Osipovich, Alexander. 2019*a*. "ICE Wants to Bring First 'Speed Bump' to Futures Markets." *The Wall Street Journal*, February 15. Retrieved December 17, 2019 from `https://www.wsj.com/articles/ice-wants-to-bring-first-speed-bump-to-futures-markets-11550228400`.

Osipovich, Alexander. 2019*b*. "Wall Street Firms Plan New Exchange to Challenge NYSE, Nasdaq." *The Wall Street Journal*, January 7. Retrieved December 17, 2019 from `https://www.wsj.com/articles/wall-street-firms-plan-new-exchange-to-challenge-nyse-nasdaq-11546866121?mod=searchresults&page=1&pos=1`.

Ostrovsky, Michael, and Michael Schwarz. 2018. "Carpooling and the Economics of Self-Driving Cars." NBER Working Paper No. 24349.

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

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

Powell, Jerome H. 2015. "The Evolving Structure of U.S. Treasury Markets." October 20. Speech at the Federal Reserve Bank of New York. Retrieved February 20, 2019 from `https://www.federalreserve.gov/newsevents/speech/powell20151020a.htm`.

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

Rochet, Jean-Charles, and Jean Tirole. 2003. "Platform Competition in Two-Sided Markets." *Journal of the European Economic Association*, 1(4): 990–1029.

Roth, Alvin E. 2002. "The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics." *Econometrica*, 70(4): 1341–1378.

Roth, Alvin E., and Robert B. Wilson. 2019. "How Market Design Emerged from Game Theory: A Mutual Interview." *Journal of Economic Perspectives*, 33(3): 118–43.

Roth, Alvin E., Tayfun Sönmez, and M. Utku Ünver. 2004. "Kidney Exchange." *The Quarterly Journal of Economics*, 119(2): 457–488.

Rothschild, Michael, and Joseph Stiglitz. 1976. "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information." *The Quarterly Journal of Economics*, 90(4): 629–649.

Santos, Tano, and Jose A. Scheinkman. 2001. "Competition among Exchanges." *The Quarterly Journal of Economics*, 116(3): 1027–1061.

Schwartz, Robert A., ed. 2001. "The Electronic Call Auction: Market Mechanism and Trading: Building a Better Stock Market." Springer.

Sönmez, Tayfun, and M. Utku Ünver. 2010. "Course Bidding at Business Schools." *International Economic Review*, 51(1): 99–123.

Spatt, Chester S. 2019. "Is Equity Market Exchange Structure Anti-Competitive?" Working Paper.

Stafford, Philip. 2019. "MEMX turns up the heat on US stock exchanges." *Financial Times*, January 9. Retrieved December 17, 2019 from `https://www.ft.com/content/4908c8b0-1418-11e9-a581-4ff78404524e`.

U.S. Securities and Exchange Commission. 1994. "Market 2000: An Examination of Current Equity Market Developments." Retrieved November 9, 2018 from `https://www.sec.gov/divisions/marketreg/market2000.pdf`.

U.S. Securities and Exchange Commission. 2016*b*. "SEC Approves IEX Proposal to Launch National Exchange, Issues Interpretation on Automated Securities Prices." Retrieved November 9, 2018 from `https://www.sec.gov/news/pressrelease/2016-123.html`.

U.S. Securities and Exchange Commission. 2016*c*. "Staff Guidance on Automated Quotations under Regulation NMS." Retrieved November 4, 2019 from `https://www.sec.gov/divisions/marketreg/automated-quotations-under-regulation-nms.htm`.

U.S. Securities and Exchange Commission. 2017*b*. "Order Approving Proposed Rule Change Amending Rules 7.29E and 1.1E to Provide for a Delay Mechanism." Retrieved December 16, 2019 from `https://www.sec.gov/rules/sro/nysemkt/2017/34-80700.pdf`.

U.S. Securities and Exchange Commission. 2017*c*. "Re: Notice of Filing of Amendments No. 1 and No. 2 and Order Granting Accelerated Approval of a Proposed Rule Change, as Modified by Amendments No. 1 and No. 2, to Adopt the CHX Liquidity Enhancing Access Delay on a Pilot Basis, Securities Exchange Act of 1934, Release No. 34-81913 (October 19, 2017)." Retrieved December 22, 2018 from `https://www.sec.gov/rules/sro/chx/2017/34-81913-letter-from-secretary.pdf`.

U.S. Securities and Exchange Commission. 2019*a*. "Commission Statement on Market Structure Innovation for Thinly Traded Securities, Release No. 34-87327." Retrieved December 6, 2019 from `https://www.sec.gov/rules/policy/2019/34-87327.pdf`.

U.S. Securities and Exchange Commission. 2019*b*. "Division of Trading and Markets: Background Paper on the Market Structure for Thinly Traded Securities." Retrieved December 6, 2019 from `https://www.sec.gov/rules/policy/2019/thinly-traded-securities-tm-background-paper.pdf`.

U.S. Securities and Exchange Commission. 2020*a*. "Market Data Infrastructure." Proposed Rule. Retrieved June 30, 2020 from `https://www.sec.gov/rules/proposed/2020/34-88216.pdf`.

U.S. Securities and Exchange Commission. 2020*b*. "Order Disapproving Proposed Rule Change to Introduce a Liquidity Provider Protection Delay Mechanism on EDGA." Retrieved June 30, 2020 from `https://www.sec.gov/rules/sro/cboeedga/2020/34-88261.pdf`.

Vayanos, Dimitri. 1999. "Strategic Trading and Welfare in a Dynamic Market." *Review of Economic Studies*, 66(2): 219–254.

White, Mary Jo. 2014. "Enhancing Our Equity Market Structure." June 5. Speech to Sandler O'Neill and Partners, L.P. Global Exchange and Brokerage Conference, New York, N.Y. Retrieved January 4, 2019 from `https://www.sec.gov/news/speech/2014-spch060514mjw`.

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

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

Zhu, Haoxiang. 2014. "Do Dark Pools Harm Price Discovery?" *The Review of Financial Studies*, 27(3): 747–789.