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, Revised January 2020

Project start date: April 2015. 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, 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, 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.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

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 January 2020
JEL No. D02,D44,D47,D53,D82,G1,G2,G23,L1,L13,L5,L89

ABSTRACT

This paper builds a model of stock exchange competition tailored to the institutional and regulatory details of the modern U.S. stock market. The model shows that under the status quo market design:(i) trading behavior across the seemingly fragmented exchanges is as if there is just a single synthesized exchange; (ii) as a result, trading fees are perfectly competitive; (iii) however, exchanges are able to capture and maintain economic rents from the sale of speed technology such as proprietary data feeds and co-location — arms for the high-frequency trading arms race. We document stylized empirical facts consistent with each of the three main results of the theory. We then use the model to examine the private and social incentives for exchanges to adopt new market designs, such as frequent batch auctions, that address the negative aspects of high-frequency trading. The robust conclusion is that private innovation incentives are much smaller than the social incentives, especially for incumbents who face the loss of speed technology rents. A policy insight that emerges from the analysis is that a regulatory "push," as opposed to a market design mandate, may be sufficient to tip the balance of incentives and encourage "the market to fix the market."

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

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 early 2019 there are 13 stock exchanges in the U.S., across which over 1 trillion shares ($50 trillion) are traded annually. All 13 exchanges use a market design called the continuous limit order book. A recent paper of Budish, Cramton and Shim (2015) showed that this market design has an important design flaw: the combination of (i) treating time as a continuous variable, and (ii) processing requests to trade serially, causes *latency arbitrage* — defined as arbitrage rents from symmetrically observed public information — to be a built-in equilibrium feature of the market design. Latency arbitrage causes markets to be less liquid than they could be, leads to a never-ending and socially-wasteful arms race for speed (now commonly measured in millionths and even billionths of seconds), and offends common economic intuitions about what constitutes an efficient market. Budish, Cramton and Shim (2015) showed that the fix is conceptually straightforward, requiring just two modifications to the continuous limit order book: (i) treat time as a discrete variable (analogously to prices, which come in discrete units); and (ii) in the event that multiple orders arrive at the same discrete time, process them in batch as opposed to serially, using a standard uniform-price auction. This idea of "frequent batch auctions" has received some high-profile attention, including from the SEC Chair (quoted above), the NY Attorney General, and the current Federal Reserve Chairman (then governor).[1] Yet, the main U.S. stock exchanges have not shown much interest in changing their market design, and indeed have vigorously opposed some more modest attempts to address the negative aspects of high-frequency trading.[2]

This context raises a pair of related questions. First, what are the economics of stock exchange competition in the modern high-frequency trading era — how do modern exchanges compete and make money? Second, what are exchanges' incentives to innovate on market design — how do the economics of modern stock exchange competition shape innovation incentives, and do private incentives align with social interests?

[1] See New York Attorney General Schneiderman (2014) and current Federal Reserve Chair (then Governor) Powell (2015), who in the context of the U.S. Treasury market remarked "Ideas such as these make me wonder whether it might collectively be possible to come to a compromise in which more trading is done directly on the public market, if at the same time the public market rules were adjusted to emphasize greater liquidity provision, and particularly more stable liquidity provision, over speed."

[2] Most notably, the exchange application of the Investors' Exchange (IEX) was opposed by all of the major incumbent exchanges and several major high-frequency trading firms (U.S. Securities and Exchange Commission, 2016*a*). The New York Stock Exchange wrote in an SEC comment letter "Like the 'non-fat yogurt' shop on Seinfeld, which actually serves tastier, full-fat yogurt to increase its sales, IEX advertises that it is 'A Fair, Simple, Transparent Market,' whereas it proposes rules that would make IEX an unfair, complex, and opaque exchange" (NYSE, 2015*b*). Please see Budish (2016*b*) for details of IEX's market design and its strengths and limitations. For the purpose of the present paper, the important limitation is that IEX's market design does not address latency arbitrage for conventional displayed limit orders, but only for non-displayed pegged orders. The other important example is the Chicago Stock Exchange (CHX), whose speed bump proposal was also widely opposed by incumbent exchanges and high-frequency trading firms (U.S. Securities and Exchange Commission, 2017*a*, 2018*a*). For example, Citadel wrote that it "unfairly structurally and systematically discriminates against market participants that are primarily liquidity takers." (Citadel, 2016) Please see Budish (2016*a*) for details of the proposal. CHX ultimately withdrew the proposal and was instead acquired by the New York Stock Exchange (Michaels and Osipovich, 2018).

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).

In this paper, we argue that incumbent exchanges' private incentives to innovate are misaligned with social interests *precisely because they earn rents from the arms race*. That is, even though there are many exchanges with little differentiation that appear to compete fiercely with one another for trading volume, they — alongside high-frequency trading firms and speed-technology providers — capture and maintain a significant share of the economic rents from speed-sensitive trading. We emphasize that our story is not one of 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 story 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, economically many aspects of trading activity behave as if there is just a single "synthesized" exchange.[3] Specifically: all

[3]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 a detailed discussion of the relationship between Glosten (1994) and this aspect of our analysis.

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).[4] 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.

As intuition for the first two results, consider a hypothetical world with buyers and sellers of a single good, and multiple platforms on which transactions can occur. A regulation corresponding to UTP would ensure that this good is perfectly homogeneous — e.g., no small differences between the types of drivers on Uber versus Lyft — and can be bought or sold on any platform. A regulation corresponding to Reg NMS ensures that searching for the best price across platforms and then potentially engaging in a transaction are literally frictionless — e.g., not 10 extra seconds to check a second ride-sharing app or 10 minutes to drive to a store, but no time at all. Given this, it is intuitive to see why: (i) aggregate economic activity will not depend on how sellers allocate their goods across platforms (as buyers will find sellers, regardless of where they are); and (ii) platform transaction fees will be Bertrand-competed down to the competitive level. There is a fundamental economic difference between an "almost" commodity and "cheap" search, and an identical commodity and zero-cost search (as in Diamond (1971)).

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 of fast trading firms and exchanges, our model suggests that exchanges in aggregate can extract at most 30% of the total

[4]These regulations are described in detail in Section 2.

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. A particularly extreme version of this move was announced recently as several large high-frequency trading firms and broker-dealers announced that they were exploring starting a new exchange, called MEMX, out of concern about rising co-location and proprietary data fees (Osipovich, 2019*b*).[5]

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.[6] 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 address our motivating question: will the market adopt new market designs, such as frequent batch auctions, that address the negative aspects of high-frequency trading? How do exchanges' private innovation incentives relate to social incentives? To conduct this analysis, we extend our theoretical model to allow for exchanges to operate one of two market designs: either

[5]The financial columnist Matt Levine wrote: "While the last new stock exchange to launch in the U.S., the Investors Exchange or IEX, was self-consciously about protecting long-term fundamental investors from the ravages of high-frequency trading, MEMX seems to be self-consciously about protecting high-frequency traders from the ravages of stock-exchange fees." (Levine, 2019)

[6]The \$0.0001 per share per side implies that across the approximately 1 trillion shares (\$50 trillion) 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). To put the 0.0001% and \$200 million figures in perspective, StubHub, the largest secondary-market venue for concert and sports tickets, has fees on the order of 30% and annual revenues exceeding \$1 billion (Budish, 2019); that is, a single secondary-market site for event tickets has revenues that are more than five times the revenues for all U.S. regular-hours equities trading.

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.[7]

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 wins 100% share. 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.[8] Please note that there are a variety of reasons, discussed in detail in Section 6.3 and an associated appendix, not to take the 100% aspect of this result literally. What we take seriously from the result is that a Discrete exchange will attract share if it enters and can earn economic rents, via non-zero trading fees, that are compensation for the 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 trading fees are competed down to zero, and trading volume is split among Discrete exchanges with zero fees. That is, we have the same Bertrand competition on trading fees as in the Continuous status quo, but now without the industry rents from the speed race.

Together, these two results imply that the market design adoption game among incumbent exchanges can be interpreted as a prisoner's dilemma: while any one exchange has incentive to unilaterally "deviate" and adopt Discrete, all incumbents prefer the Continuous status quo, in which they share in latency arbitrage rents, to a world 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 "cooperate" and 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.

The prisoner's dilemma finding does not yield an unambiguous answer to our motivating question, "Will the market fix the market?" The robust conclusion is that private innovation incentives are much smaller than the social incentives, especially for incumbents who face the loss of speed technology rents, but whether these private incentives are strictly positive or strictly negative depends on parameters such as the cost of adoption, speed of imitation, and the magnitude of the latency arbitrage prize.

At the same time, the prisoner's dilemma finding does yield a clear, and perhaps surprising, insight

[7] 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. See Section 2.2 for additional discussion of Reg NMS. 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.

[8] 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.

about the potential role of 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, which is a regulatory "push": any policy that tips the balance of incentives sufficiently to get a de novo exchange to enter or an incumbent to adopt. Such an initial entrant will gain share, which would not necessarily be the case in a coordination game environment, and, taking the model literally, helps move the market from the (all Continuous) cell of the adoption game matrix to the (all Discrete) cell.

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., 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.

Our paper makes several 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 innovative 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 leads to economics of the status quo that are also fundamentally different from those that would emerge under standard platform or two-sided competition frameworks where, typically, platforms earn rents from platform-specific network effects by charging supra-competitive access and transaction fees (cf. 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. 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. The market microstructure literature has in the past been puzzled by fragmentation (see Madhavan (2000) and what he terms the "Network Externality Puzzle"). Here, we provide a theoretical rationale for why fragmentation per se may not necessarily lead to trading inefficiencies — this aspect of our analysis builds on a prescient result of Glosten (1994) and aligns with empirical evidence in O'Hara and Ye (2011).

There are also two technical features of our theoretical analysis worth highlighting. First, we develop and motivate an equilibrium solution concept which we refer to as 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, we generate a strictly interior split of latency arbitrage rents between exchanges and trading firms without relying on an explicit bargaining model; we show that 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 on speed technology fees may be of direct use for current policy debates. The SEC recently announced a pilot study on transaction fees, focusing on the controversial practice of "maker-taker" fee-and-rebate pricing models (U.S. Securities and Exchange Commission, 2018*c*). While our results do not speak to the agency concerns at the heart of the controversy (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). In a speech around that time Commissioner Robert J. Jackson Jr. called for "greater transparency about how exchanges make their money. . . and a clear and uniform approach to disclosing revenues across exchanges and over time." He described that he and his staff "tried and failed to use public disclosures to meaningfully examine exchanges' businesses. . . [and] attempted to look into the revenues that exchanges generate from selling market data and connectivity services. We expected that such numbers would be available. . . but found. . . it nearly impossible" (Jackson Jr., 2018). Our estimate of total exchange speed-technology revenues — which, as the reader will see, triangulates from numerous data sources in lieu of obvious, transparent numbers from exchange filings — is surely not perfect, but it provides a magnitude that market policy makers currently lack.

Last is our analysis of the motivating question, "Will the market fix the market?" More precisely, the intellectual contribution is in using the model to fill in the cells of the market design adoption game payoff matrix. Once we understand that the adoption game payoffs can be interpreted as a prisoner's dilemma, as opposed to, e.g., being in a coordination-game environment, the rest of the analysis follows standard ideas from the innovation literature. We 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 particular 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.[9]

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.

[9]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 (2019)), 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 Antill and Duffie (2018), Brogaard, Hendershott and Riordan (2017), Bulow and Klemperer (2013, 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,[10] a large majority of its volume transacted on the Nasdaq exchange (86% in 1993).[11] 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.[12] 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 quite complex and in many ways opaque (see Chao, Yao and Ye (2019)), are ultimately quite small, as we will document rigorously 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. The first set, 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. The second set, Regulation National Market System (Reg NMS), was implemented in 2007 and requires that information about trading opportunities (i.e., quotes) be automatically disseminated across the whole market (including both other exchanges and entities such as brokers), and also requires, roughly, that the whole market pay attention to such information and direct trades to the most attractive prices across the whole system. As we will see in our formal model in Section 3, this effectively nullifies any exchange-specific network effects.[13]

In this institutional background section we describe each of these sets of regulations; our goal is to provide a level of detail that is sufficient to justify our modeling choices.

We note that while our discussion focuses on the United States, there are economically similar regulations for stock exchanges in Canada and somewhat similar regulations in Europe.[14] Regulations for futures exchanges, on the other hand, are quite different from those for stock exchanges, both in the U.S. and abroad.

[10] 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.

[11] 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.

[12] 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).

[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). A good topic for future research would be to incorporate latency arbitrage into a model with competition between exchanges and dark pools.

[14] In Canada's version of the Order Protection Rule (which goes by the same name), the key difference is that the rule applies to the full depth of the order book, not just the first level (Canadian Securities Administrators, 2009). In Europe, instead of the (prescriptive) Order Protection Rule there are (principles-based) best execution regulations (Petrella, 2010). Note however that principles-based best execution requirements leave some ambiguity with regard to whether market participants have to "pay attention" to quotes from small exchanges, which could affect innovation incentives; whereas under the Order Protection Rule there is no such ambiguity. This seems a good topic for future research.

In particular, there is no analogue of UTP in futures markets because each contract is proprietary to a particular exchange. Similarly, there are differences between the regulation of stock exchanges and the regulation of financial exchanges for other financial instruments like government bonds, corporate bonds, foreign currency, etc.; in particular, the information dissemination provisions of Reg NMS are often economically different in these asset classes. As we emphasize in the conclusion, there are many open directions for future research.

2.1 Unlisted Trading Privileges (UTP)

Section 12(f) of the 1934 Exchange Act (15 U.S.C. 78a, 1934), passed by Congress, directed the Securities and Exchange Commission to "make a study of trading in unlisted securities upon exchanges and to report the results of its study and its recommendations to Congress." Since that time, the right of one exchange to facilitate trading in securities that are listed on other exchanges has undergone several evolutions. In its current form, passed by Congress in the Unlisted Trading Privileges Act of 1994 (H.R. 4535, U.S. Congress, 1994) and clarified by the SEC in a Final Rule effective November 2000 (U.S. Securities and Exchange Commission, 2000), one exchange may extend unlisted trading privileges (UTP) to a security listed on another exchange immediately upon the security's initial public offering on the listing exchange, without any formal application or approval process through the SEC.[15]

For the purposes of our theoretical model, we will 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, and its value is the same regardless of where it is traded.

2.2 Regulation National Market System (Reg NMS)

Regulation National Market System ("Reg NMS," U.S. Securities and Exchange Commission, 2005) passed in June 2005 and implemented beginning in October 2007, is a long and complex piece of regulation, with roots tracing to the Securities Exchange Act Amendments of 1975 and the SEC's "Order Handling Rules" promulgated in 1996.[16] For the purpose of the present paper, however, there are two core features to highlight.[17]

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. A quote on a particular exchange is "protected" if it is (i) at that exchange's current best bid or offer; and (ii) "immediately and automatically accessible" by other exchanges. Reg NMS does not provide a

[15] Prior to 1994, exchanges had to formally apply to the SEC for the right to extend UTP to a particular security; such approval was "virtually automatic" following a delay of about 30-45 days (Hasbrouck, Sofianos and Sosebee, 1993). Between the passage of the UTP Act of 1994 and the Final Rule in 2000, extension of UTP was automatic but only after an initially two-day, and then one-day, delay period after the security first began trading on its listing exchange (U.S. Securities and Exchange Commission, 2000). For further historical discussion of UTP, please see the background section of the 2000 Final Rule document, and also Amihud and Mendelson (1996).

[16] The goal of the National Market System is described by the SEC as follows: "The NMS is premised on promoting fair competition among individual markets, while at the same time assuring that all of these markets are linked together, through facilities and rules, in a unified system that promotes interaction among the orders of buyers and sellers in a particular NMS stock. The NMS thereby incorporates two distinct types of competition—competition among individual markets and competition among individual orders—that together contribute to efficient markets." U.S. Securities and Exchange Commission (2005, pg 12)

[17] For an overview of Reg NMS, a good source is the introductory section of the SEC's final ruling itself (U.S. Securities and Exchange Commission, 2005). For an overview of the National Market System prior to Reg NMS, good sources are O'Hara and Macey (1997) and the SEC's "Market 2000" study (U.S. Securities and Exchange Commission, 1994).

precise definition of "immediately and automatically accessible," but the phrase certainly included automated electronic continuous limit order book markets and certainly excluded the NYSE floor system with human brokers. A June 2016 rules clarification issued by the SEC indicated that exchanges can use market designs that impose delays on the processing of orders and still qualify as "immediate and automatic" so long as (i) the delay is of a *de minimis* level of less than 1 millisecond, 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*). This rules clarification suggests that quotes in a frequent batch auction exchange would be protected under Rule 611 so long as the batch interval satisfies the de minimis delay standard; however, this specific market design has not yet been put before the SEC for explicit approval.[18]

An additional detail about the Order Protection Rule that bears emphasis is that, in practice, sophisticated market participants can take on responsibility for compliance with the Order Protection Rule themselves, absolving exchanges of the responsibility. They do so using what are known as intermarket sweep orders, or ISOs. If an exchange receives an order that is not marked as ISO, then it is the exchange's responsibility to ensure that it handles the order in a manner compliant with the Order Protection Rule (e.g., it cannot execute a trade that trades through a protected quote elsewhere). If an exchange receives an order that is marked as ISO, then the exchange may presume that the sender of the order has ensured compliance with the Order Protection Rule (e.g., by also sending orders to other exchanges to attempt to trade with any relevant protected quotes) and the exchange need not check quotes elsewhere before processing the order.

The second key provision to highlight 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. As the SEC writes (pg. 26), "... protecting the best displayed prices against trade-throughs would be futile if broker-dealers and trading centers were unable to access those prices fairly and efficiently."

The Access Rule has three sets of provisions that together are aimed at ensuring such efficient access—or what we will sometimes call "search and access," to highlight that economically the Access Rule (and related rules that affect information provision, such as those governing slower, non-proprietary market data feeds)[19] enables market participants to both search available quotes and then "access" them, i.e., trade against them. First, Rule 610(c) limits the trading fee that any exchange can charge to 0.3 pennies, which, importantly, is less than the minimum tick size of 1 penny. This ensures that if one exchange has a strictly better displayed price than another exchange, the price is economically better after accounting for fees.

[18] To date, the one market design that has been approved by the SEC that imposes a de minimis delay is IEX's (see footnote 2). The SEC issued its rules interpretation of "immediate and automatic" simultaneously with its approval of IEX's exchange application on June 17, 2016 (see U.S. Securities and Exchange Commission (2016*b*) and the related materials referenced therein). Subsequent to IEX's approval, the Chicago Stock Exchange (CHX) applied for approval of an asymmetric delay market design, in which marketable limit orders are slightly delayed, to give liquidity providing quotes a small head start against snipers in the event of a sniping race. This market design has not been approved. See U.S. Securities and Exchange Commission (2017*a*, 2018*a*) for the history and public comments regarding CHX's two versions of the proposal, the latter of which the SEC officially stayed on Oct 24, 2017, and CHX officially withdrew on July 25, 2018 when it was acquired by the New York Stock Exchange. The main substantive argument against the CHX proposal expressed in public comment letters was that the *asymmetry* of the delay is inconsistent with the fairness provisions of the Exchange Act.

[19] 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.

Second, Rule 610(d) has provisions that together ensure that prices across markets do not become "locked" or "crossed" — specifically, each exchange is required to monitor data from all other exchanges and to ensure that it does not display a quote that creates a market that is locked (i.e., bid on one exchange equal to ask on another exchange) or crossed (i.e., bid on one exchange strictly greater than an ask on another exchange). Together, then, rules 610(c) and 610(d) ensure that there is a well-defined "national best bid and offer" (NBBO) across all exchanges (at least ignoring the complexities that arise due to latency, see Section 4 of Budish (2016*b*)). Third, Rule 610(a) prevents exchanges from charging discriminatory per-share trading fees based on whether the trader in question does or does not have a direct relationship with the exchange. In our model, the notion of a direct relationship with the exchange is captured by the decision of whether to buy exchange-specific speed technology, which represents exchange products like proprietary data feeds, co-location, and connectivity. What Rule 610(a) ensures is that market participants face the same trading fee schedule, whether or not they have such a direct relationship.

To summarize, any time any market participant submits an order, it is required under Rule 611 that either the market participant themself (if using ISOs) or the exchange they submit their order to checks quotes on all exchanges. Rule 610 then ensures that this mandatory search is feasible, and that the only marginal costs of accessing a particular quote on a particular exchange are the exchange's per-share trading fees. For our theoretical model, therefore, 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. The choice of zero (as opposed to epsilon) is appropriate both because the marginal costs in practice really are negligible, and because compliance with Rule 611 is mandatory, and zero captures that it is cheaper to comply with the rule than not to.

3 Theory of the Status Quo

We now introduce our model of stock exchange competition to better understand the status quo of the market. In this Section, we restrict all exchanges to employ the continuous limit order book market design; later, in Section 5, we extend our model and allow exchanges to be strategic with respect to their market design choice.

3.1 Overview of the 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, whereas BCS examined trading on a single non-strategic exchange, our model has multiple exchanges competing in an environment shaped by the key institutional details reviewed in Section 2. Second, rather than working with a continuous-time model in which certain events — including the arrival of investors or changes in the fundamental value of a security — 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.[20] This approach will retain the economic interpretability of the continuous-time

[20]Even for the highest activity symbol in all of US equity markets, SPY, on its highest-volume day of 2018 (Feb 6th), 95.2%

Poisson model used in BCS while providing tractability when modeling trading behavior across multiple exchanges. Third, 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. Last, we develop and employ an alternative equilibrium solution concept for our trading game that we refer to as 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 in our trading game by restricting the set of deviations that an equilibrium must be robust to, in a manner similar to alternative equilibrium notions developed in insurance markets (Wilson, 1977; Riley, 1979). These restrictions attempt to capture the idea of competitive liquidity provision as in Glosten and Milgrom (1985), and immediate responses by rivals as in BCS. In the single-exchange case, our approach reaches the same economics as BCS. In the multi-exchange case, the added formalism of our concept is useful for clarity and precision.

3.1.1 Setup

There is a single security, x, and a signal of the value of the security, y. Following BCS, 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. Exchanges all use the continuous limit order book market design and are ex ante undifferentiated. The asset x is completely fungible across exchanges, i.e., its value does not depend on the exchange on which it is traded, which captures the economics of Unlisted Trading Privileges as discussed in Section 2. We assume that prices are continuous and that shares are perfectly divisible. Assuming continuous prices allows us to abstract from the queueing dynamics that are present in markets with binding tick-size constraints; we discuss tick-size frictions further in Section 4.4 and 6.3. 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.

In our model, we focus on the actions of 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 and there is no discounting.

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 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.[21] Formally, if an investor arrives to market needing to buy one unit of x, buys a unit

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.

[21] Alternatively, we could model investors as preferring to transact sooner rather than later all else equal (e.g., they possess a small cost of delay per unit time). Since the bid-ask spread will be stationary in equilibrium, and y is a martingale, this

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 $v + (p - y)$.[22] Note that what we call investors could also be termed "noise traders" since they are essentially mechanical. 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 mechanically trade a single time upon arrival, but the model does not accommodate strategic trading over time as in Kyle (1985) and related models.

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 $p - y$.[23]

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 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. In this Section when all exchanges employ the same market design, there will be no role in equilibrium for slow trading firms; hence, we will use TFs to refer to the N fast TFs unless explicitly noted otherwise.

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.[24] 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.[25] In reality, such technology allows trading firms

modeling convention would also lead investors to trade immediately in the period they arrive.

[22] 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 will always transact exactly one unit.

[23] 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 activity (e.g., trading slowly over time, as in Kyle (1985) and a large literature thereafter).

[24] 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. This point is made in Chao, Yao and Ye (2017, 2019) who then show that when prices are discrete (i.e., there is a minimum tick size), asymmetric fee structures can be used to "fill in the ticks". We discuss these issues further in Section 4.2 and in Appendix B.

[25] In practice the dividing line between exchange-specific technology and general-purpose technology is not sharp — for

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 (2018)), 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 TFs with both exchange-specific and general-purpose speed technology; next are messages from TFs with only general-purpose speed technology; and last are messages from market participants with neither. We assume that the processing order on an exchange is uniformly random among firms that have the same set of speed technologies. 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, gets their money back, 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 TFs, any TF would be able to unilaterally deny usage of ESST on any exchange to the other TF by not purchasing it.

The following objects are primitives of our 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).

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

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.

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 $\boldsymbol{\omega}_j$, 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: i.e., 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 other market participants (including slow trading firms) are processed in a uniformly random sequence.[26]

(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, UTP is incorporated 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

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

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 particular exchange is the per-share trading fee. As discussed in detail in Section 2, these assumptions capture the key provisions of Reg NMS, namely the Order Protection Rule (Rule 611) which mandates that market participants (or broker-dealers or exchanges operating on their behalf) search across all exchanges on an order-by-order basis, and the Access Rule (Rule 610).

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 (2018).

3.2 Equilibrium Analysis

3.2.1 Stage Three Trading Game Behavior: A Single Exchange

To establish intuition for equilibrium play in our overall game, it is helpful to initially focus on behavior in Stage 3 of our model when there is only a single exchange ($M = 1$). In this simplified setting, it is easiest to understand the behavior of market participants when the single exchange has zero trading fees, and all N fast TFs have access to ESST from this exchange. This case is analogous to the setting analyzed in BCS, and we show that the key economics introduced in BCS are still present.

In Stage 3 (both here and later with multiple exchanges), we restrict attention to pure Markov strategies: market participants are only able to condition their pure strategies on the publicly observable state, and not on the history of play in previous trading games. Our focus on Markov strategies implies that in period 1 of any trading game, when sending orders to the exchange, all TFs condition their actions only on the state at the beginning of the trading game (which includes outstanding orders on each exchange's order book); in period 2, all market participants condition their actions only on the updated state, which accounts for

actions taken by all market participants in period 1 and by nature.[27]

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:

- Upon arrival, an investor and informed trader have essentially unique optimal strategies: an investor trades against all available liquidity, up to one unit, at the best price(s) possible, and, additionally, trades against any remaining profitable orders based on the publicly observed y (i.e., those willing to buy at more or sell at less than y); and an informed trader immediately trades against any profitable orders based on his privately observed y. Note that after the informed trader has traded and any private information is publicly revealed, there no longer exist strictly profitable trading opportunities.

- 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 will occur. Any TF providing liquidity that wishes to replace an order will be indifferent between cancelling 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 (i.e., y increases to a price higher than the best ask or decreases to a price lower than the best bid), there will be a sniping race as described in BCS: those TFs that are providing such liquidity at unprofitable prices will send cancellation messages to the exchange to try to cancel these stale quotes, while at the same time all other TFs will send IOCs to the exchange to try to trade against ("snipe") these stale quotes. Note that firms may simultaneously try to cancel their own quotes and snipe others' quotes. If there are N TFs that are all equally fast, the probability that any one liquidity provider is sniped in response to public information is $(N-1)/N$, since unless his request to cancel is first (with probability $1/N$) he will get sniped. In equilibrium, whether a quote is sniped or cancelled will be randomly determined, and the winner of the sniping race will be one of the fast TFs.

Thus, as market participants have essentially unique optimal strategies in period 2 (conditional on a stochastic decision by nature), the analysis of each trading game simplifies to understanding TF behavior in period 1.

Consider now a 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 for clarity of exposition to focus on the provision of liquidity via two limit orders: for a given quantity q and fundamental value y, a 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.

[27] Even though trading games are infinitely repeated, we will first analyze each trading game in isolation of others (thereby ignoring the possibility that actions in one trading game may affect continuation payoffs in subsequent games). Later, we check and show that repeated play of the equilibrium that we construct for a single trading game remains an equilibrium for the infinitely repeated trading game when such interactions are accounted for.

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

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

Such a bid-ask spread equalizes liquidity providers' expected benefits to the expected costs from both traditional adverse selection from private information as well as sniping from symmetrically observed public information.[28] Here, as in BCS, 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 liquidity provider loses the race to respond to public information, as well as a $\frac{1}{N}$ factor that captures a liquidity provider's opportunity cost of not sniping. 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 our restriction to Markov strategies, a natural solution concept for our repeated trading game is pure-strategy Markov-perfect equilibrium (MPE). However, a pure-strategy MPE does not exist for our trading game.[29] This non-existence result arises because of adverse selection. In a standard model of undifferentiated Bertrand competition without adverse selection, an equilibrium exists with marginal-cost pricing — i.e., any firm is willing to sell as much as the market demands at its marginal cost. In a sense, "excess liquidity provision" in this standard environment is riskless and constrains the price that any particular firm 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. Without this discipline, the TFs who are providing (non-excess) liquidity will have an incentive to widen their spreads.

To address this non-existence issue in our trading game, we introduce an alternative equilibrium solution concept that we refer to as *order book equilibrium.* In contrast to MPE, which requires that no strictly profitable unilateral deviations exist, our concept strictly weakens MPE and allows for certain strictly profitable deviations to exist as long as they are rendered unprofitable by specific reactions from rivals. Our concept allows for two types of reactions. The first type of reaction allows for TFs to provide additional liquidity at a better price. This means that if a liquidity-providing TF deviates by widening its spread, another TF could undercut and provide additional liquidity if it would wish to do so. This allows TFs to discipline equilibrium

[28] 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.

[29] To see why, consider any set of TF strategies where, following period 1 and heading into period 2, exactly one unit of liquidity is offered, for instance at spread $s^*_{continuous}$ as defined in (3.1). This cannot be an MPE 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 provided, 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 provided, there is a strictly profitable deviation to add the missing amount at a high spread, in case an investor arrives. Hence, there is no MPE.

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 TFs earn strictly positive profits. The second type of reaction allows for TFs to withdraw liquidity in response to deviations. 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). By using anticipated reactions to counter otherwise profitable deviations, our concept 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.[30] We provide a formal definition of our order book equilibrium concept, and also intuition for why it helps to restore equilibrium existence, in Appendix A.1.

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)).[31] 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.

Single-Exchange Trading Game Equilibrium. We now characterize equilibrium behavior in Stage 3:

Proposition 3.1. *Any order book equilibrium with a single exchange charging zero trading fees and all fast TFs having purchased ESST has the following properties. A single unit of liquidity is provided at bid-ask spread $s^*_{continuous}$ (defined in (3.1)) around y following Period 1 of each trading game. In period 2: an investor, upon arrival, immediately purchases or sells one unit of security x at the best price; an informed trader, upon arrival, trades immediately against any profitable orders; and if a publicly observable jump in y occurs, a sniping race occurs whereby all trading firms attempt to trade against existing quotes if profitable, and all trading firms providing liquidity will attempt to cancel their orders that are no longer profitable to offer. Such an equilibrium exists.*

(All proofs in appendix.) As in BCS, in any equilibrium, the N fast TFs endogenously engage in both liquidity provision and sniping; at the equilibrium spread $s^*_{continuous}$, TFs are indifferent between these two activities. This implies that all TFs, including those that are liquidity providers, earn rents by splitting the surplus generated by latency arbitrage activities. We refer to this surplus as the total *sniping prize*, defined as $\Pi^*_{continuous} \equiv \lambda_{public} \cdot L(s^*_{continuous})$. Note that, although the equilibrium bid-ask spread $s^*_{continuous}$ reflects

[30] These features are shared with the equilibria described in BCS, which examines trading activity in continuous time and assumes that any profitable unilateral deviation (e.g., widening spreads or undercutting) can be met with "immediate" reactions from rivals. Our solution concept adapts the same reasoning — that rivals can react to deviations — to our discrete time environment.

[31] 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).

both the cost of sniping and the cost of traditional adverse selection $(\lambda_{public} + \lambda_{private}) \cdot L(s^*_{continuous})$, the sniping prize depends only on the magnitude of sniping opportunities $\lambda_{public} \cdot L(s^*_{continuous})$.[32]

3.2.2 Equilibrium of the Full (Multi-Exchange) Game

We now examine our full game, where there are multiple strategic exchanges ($M \geq 2$) across which the security x can be traded. Recall that in Stage 1, exchanges simultaneously choose trading and ESST fees; in Stage 2, TFs choose which exchanges to purchase ESST from; and in Stage 3, our infinitely repeated trading game is played. In the first two stages of our game, which are played once, we assume exchanges and TFs play a subgame perfect Nash equilibrium given anticipated order-book equilibrium behavior in Stage 3.

The main result of this Section is Proposition 3.2 (below), which states that there exist equilibria of our 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 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 TFs. And last, in each trading game, outcomes of the single-exchange case are essentially replicated, but now across multiple exchanges according to some vector of market shares, denoted $\boldsymbol{\sigma}^*$: i.e., TFs provide exactly one unit of liquidity at spread $s^*_{continuous}$ across all exchanges according to $\boldsymbol{\sigma}^*$; investors and informed traders, upon arrival, trade wherever profitable, splitting their orders across exchanges according to $\boldsymbol{\sigma}^*$; and in the event of a sniping race, the sniping race plays out in parallel across all M exchanges, with all N 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 TFs now split the rents generated from latency arbitrage activity.

Proposition 3.2. *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 of the multiple-exchange game where:*

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

(Stage 2): All N 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 purchases or sells one unit of x at the best price, transacting σ^*_j of volume on each exchange j; an informed trader, upon arrival, trades immediately against any profitable orders on all exchanges; and if a publicly observable jump in y occurs, a sniping race occurs whereby all trading firms attempt to trade against existing quotes if profitable, and all trading firms providing liquidity will attempt to cancel their orders that are no longer profitable to offer.

[32] It is straightforward to demonstrate that repeated play of any order book equilibrium strategies that condition only on the state $(y, \boldsymbol{\omega})$ comprises an equilibrium of the infinitely repeated trading game. This need not have been true, since orders resting in the order book at the end of a trading game have priority that carries over to subsequent trading games, potentially generating a queueing motive. However, this concern is not an issue here as TFs are indifferent between liquidity provision and sniping in each trading game. Nevertheless, as discussed in BCS, if prices are restricted to lie on a discrete grid of points (e.g., on the penny), liquidity provision will be strictly preferred to latency arbitrage in the equilibrium that we construct. In such a setting, the benefits from liquidity provision across multiple trading games must be accounted for, and analyzing individual trading games in isolation as we have done here is no longer appropriate.

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 $\pi_N^{lone-wolf}$ *is a constant discussed below and defined in Appendix A.2.2, equation (A.3).*

The proof of this result is constructive. We first examine behavior in the multi-exchange version of our trading game (Stage 3). We show that if all N 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 has exactly a single unit of liquidity provided at spread $s^*_{continuous}$: i.e., the outcome is the same as the single exchange trading game — TFs engage in both liquidity provision and sniping, and split the sniping prize $\Pi^*_{continuous}$ — with the only difference being that trading activity now may be split across multiple exchanges (Lemma A.1). In the equilibria that we construct, what we refer to as 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 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 $\boldsymbol{\sigma^*} = (\sigma^*_1, \ldots, \sigma^*_M)$ such that $\sum_j \sigma^*_j = 1$, there is an equilibrium in which each exchange j's share of depth and volume are each exactly σ^*_j.

Next, we 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 TFs to purchase ESST from all exchanges as long as condition (3.2) is satisfied. If all TFs purchase ESST from all exchanges, in any order book equilibrium of the subsequent trading game, each TF obtains (in expectation, gross of ESST fees) their share of the sniping prize, $\Pi^*_{continuous}/N$. We next examine what we refer to as a *lone-wolf deviation* for any TF i at Stage 2 given equilibrium strategies: to purchase ESST from a single exchange charging the lowest ESST fee, and subsequently provide a single unit of liquidity at a spread that is strictly narrower than $s^*_{continuous}$ in each subsequent trading game (which we prove to be an equilibrium of the Stage 3 subgame; Lemma A.2). In doing so, TF i is guaranteed to earn in expectation an amount $\pi_N^{lone-wolf}$, where $\pi_N^{lone-wolf} \in (\frac{N-2}{N-1} \times \frac{\Pi^*_{continuous}}{N}, \frac{\Pi^*_{continuous}}{N})$ per trading game and $\pi_N^{lone-wolf}$ is explicitly derived in Appendix A.2.2, equation (A.3). Condition (3.2) ensures that such a lone-wolf deviation would be unprofitable for all TFs, as each TF would earn more in expectation by purchasing ESST from all exchanges and earning $\Pi^*_{continuous}/N$ per-trading game than purchasing ESST from only a single exchange and earning $\pi_N^{lone-wolf}$. It is worth emphasizing that if there were only a single exchange, TFs could not leverage such a lone-wolf deviation to play exchanges off against each other, and an exchange would be able to extract the entire amount $\Pi^*_{continuous}$ via ESST fees.

Finally, we show that there is an equilibrium of our overall game in which exchanges all charge zero trading fees and levy any arbitrary vector of ESST fees that satisfy condition (3.2).

3.3 Discussion

We now discuss the three main features of the equilibria described in Proposition 3.2. Later, 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, with both equal to σ^*_j for each exchange j. 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, i.e., any vector of σ^*_j's.

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) which we discuss in turn. 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 (Proposition 8) that if there is an equilibrium price schedule $R(q)$ for a single limit order book exchange (where q denotes the quantity traded and $R(q)$ denotes the total price paid by the investor) there also exists an equilibrium in which there are two exchanges whose price schedules sum up to $R(q)$. That is, multiple exchanges can co exist in equilibrium if their liquidity schedules add up to what would have been provided on a single exchange. Relative to our Stage-3 trading game, Glosten (1994)'s model is more general in that investors may have multi-unit demands and risk-aversion — hence, an equilibrium price schedule, rather than equilibrium bid-ask spread for a single unit, as in our model. That said, our simpler treatment of investor preferences allows us to model exchanges as strategic players who set prices, and allows us to incorporate latency arbitrage into the analysis. Another difference is that the trading firms in our model both provide liquidity and snipe stale quotes, and make strategic decisions about speed technology, whereas in the Glosten (1994) model liquidity is provided by an infinite number of small liquidity providers.

Ellison and Fudenberg (2003) study competition among platforms for single-homing buyers and sellers. Their setting encompasses elements of the Pagano (1989) exchange competition model as well as platform competition in other settings. They 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.[33] Our notion of the depth-volume ratio is inspired by the seller-buyer ratio in the Ellison and Fudenberg (2003)

[33]The "plateau" refers to an inverval 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).

model.

Similarly to these other models, our model does not yield much insight into the determination of equilibrium exchange market shares, i.e., the σ_j^*'s. 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 A.2.3). Such strategies, in turn, coordinate where TFs provide liquidity. This implies the following. First, if in reality investors (or broker-dealers acting on their behalf) and TFs prefer there to be multiple active exchanges — for example, to mitigate market power that any single exchange family can wield in some manner outside of our model — they may jointly wish to spread their trading activity across exchanges via interior routing tables and fragmented liquidity provision. Second, although there exist equilibria in which routing table strategies are not stationary (indeed, they could be arbitrarily chaotic), one may expect that routing tables are relatively stable over time in practice, because it makes the coordination described above more plausible.[34]

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 A.1 in Appendix A.2).

ESST Fees and the Division of Latency Arbitrage Rents. 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 what we refer to as a *binding money-pump constraint*. Trading fees are zero across all exchanges. Any dissipation of ESST rents via trading fees in order to attract trading volume would require such fees to be negative, which in turn would create an incentive for market participants to execute an unlimited number of trades and make unlimited profits — i.e., a money-pump.[35] This constraint is critical: if market participants perceived transactions to be sufficiently costly (e.g., due to clearing or other transaction costs) so that a money-pump would never exist even if trading fees were negative, then exchanges would not be able to earn positive rents in any equilibria where TFs only purchase ESST from exchanges with the lowest trading fees.[36]

[34] Clearly, our assumption that all investors trade a single unit of the security and can arbitrarily split such orders across exchanges is a modeling convention that abstracts away from many realistic details. An alternative interpretation of these routing table strategies is the probability that a broker-dealer sends any order to a given exchange when indifferent. Under this interpretation, market shares may be highly variable across small time intervals representing individual trades (high-frequency data), but these shares are more likely to be stable across longer intervals (e.g., minutes, hours or days). See a related discussion of the depth-volume empirics in Section 4.1.

[35] Although exchanges theoretically could dissipate rents via fixed payments to investors or broker-dealers for trading volume, such payments are not observed nor, to our understanding, legal.

[36] To see why, consider the setting with two exchanges, A and B , and a candidate equilibrium where exchange A charges fees f_A^* and $F_A^* > 0$. If A earns positive rents (implying that TFs purchase ESST from A, and trading fees, even if negative, do not

We now turn to the determination of ESST fees. First, though exchanges are able to "post prices" and make take-it-or-leave-it offers to TFs, they cannot capture all latency arbitrage rents: each TF maintains some degree of bargaining leverage with any given exchange through its ability to steer trading volume via liquidity provision on rival exchanges (referred to above as a lone-wolf deviation). This gives rise to the condition on ESST fees given by condition (3.2). Though Proposition 3.2 also establishes that there are equilibria where condition (3.2) does not bind and exchanges charge lower ESST fees (including zero in total), such equilibria require TFs to coordinate with one another and not purchase ESST from any exchange that raises its ESST fee above some arbitrary threshold. Such coordination may be difficult to maintain, as there also exist subgame equilibria (beginning in Stage 2) where all TFs continue to purchase from an exchange that slightly raises its ESST fee. Motivated by this observation, we believe a reasonable refinement to be one that rules out the possibility that an exchange could increase its ESST fee, and all TFs purchasing ESST from that exchange still comprises an equilibrium. Indeed, the only equilibria where TFs purchase ESST from all exchanges that also satisfy this additional requirement are those in which condition (3.2) is binding. The following proposition summarizes these results:

Proposition 3.3. *In any equilibrium in which all trading firms purchase ESST from all exchanges and trading fees are zero for all exchanges, ESST fees* $\boldsymbol{F^*}$ *must satisfy (3.2). Furthermore, among these equilibria, if there does not exist (i) a vector of ESST fees* $\boldsymbol{F'}$ *such that* $F'_j \geq F^*_j$ *for all exchanges* j *and* $F'_k > F^*_k$ *for at least one exchange* k*, and (ii) a subgame equilibrium beginning in Stage 2 where all trading firms purchase ESST from all exchanges at fees* $\boldsymbol{F'}$*, then (3.2) must bind.*

With this additional refinement, our theory delivers a rather striking prediction: there is a strictly interior division of latency arbitrage rents between TFs and exchanges pinned down by (3.2). On the one hand, exchanges are able to extract a share of latency arbitrage rents as price-setters for ESST, and do not dissipate these rents via competition due to the money-pump constraint. On the other hand, TFs are able to maintain a significant portion of the rents generated from latency arbitrage activity, even as price takers, because they are able to affect equilibrium trading volume: TFs choose where to provide liquidity (which, in the equilibria constructed, has a linear relationship to realized trading volume), and can discipline any exchange demanding higher ESST by only purchasing ESST from rival exchanges and providing liquidity there.[37]

The proportion of the latency arbitrage rents that TFs maintain is economically meaningful:

Proposition 3.4. *In any equilibrium in which all* N *trading firms purchase exchange-specific speed technology (ESST) from all exchanges and ESST fees satisfy condition (3.2), exchanges' total rents from ESST fees,* $N \times \sum_j F^*_j$*, 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 (2019) 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.

completely offset ESST fees), B could undercut A with trading fees $f'_B = f^*_A - \varepsilon$ and levy higher ESST fees; such a deviation would induce all TFs to only purchase ESST from B (as any exchange that does not have the lowest trading fees cannot sustain positive trading volume in equilibrium), which would be strictly profitable for some $F'_B > F^*_B$ and sufficiently small $\varepsilon > 0$.

[37] In contrast, consider an alternative model in which each exchange j is able to process no more than σ_j units of volume in any period, where $\sum_j \sigma_j = 1$: in such a model, each exchange has essentially a monopoly over its share of trading volume, and TFs cannot restrict ESST fees by providing additional liquidity elsewhere. In this alternative model, it is straightforward to show that exchanges would be able to extract all latency arbitrage rents.

equities.[38] Proposition 3.4 implies that if $M \geq 3$ and $N \geq 6$, then exchanges in total are able to extract at most 30% of latency arbitrage rents.

We emphasize that while this particular interior division of latency arbitrage rents is specific to our model, what will ultimately matter for the main questions posed by our paper is simply that exchanges are able to *capture and maintain* some positive share of rents generated from latency arbitrage activity in the status quo. 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.[39] 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. This is because TFs have power to re-direct trading volume if exchanges charge too much for ESST. Yet, we by no means think this is the only useful way to model this rent-division game, nor would we want readers of the paper to take the specific formula given by (3.2) or the specific off-path "lone-wolf" threats too literally.

Sources of Deadweight Loss. In our model, there are N trading firms exogenously endowed with general-purpose speed technology, and M exchanges exogenously present in the market and able to sell exchange-specific speed technology to TFs. TFs' payments to the exchanges for this speed technology, represented by the F_j^*'s in our model, are transfers as opposed to deadweight loss.

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

In addition, the harm to liquidity caused by latency arbitrage could itself have efficiency consequences, such as by making markets less informationally efficient (as in Baldauf and Mollner (2018)), or by making it more difficult for investors to spread and rebalance risk (as in models like Vayanos (1999) and Sannikov and Skrzypacz (2016)).

Last, we note that standard excess entry and business stealing incentives (Mankiw and Whinston, 1986) 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 Empirical Validation

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, and in particular the notion of a synthesized single exchange. Section 4.2 presents facts that relate to the model's equilibrium characterization of exchange trading fees, i.e., f. Section 4.3 presents facts that relate to the model's equilibrium characterization of exchange-specific speed technology fees, i.e., F. Section 4.4 will

[38] 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.

[39] Roth and Wilson (2018) 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).

provide 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.

This analysis utilizes a combination of exchange-labeled trades-and-quotes data, exchange fee schedules, and exchange financial documents such as 10-K's and S-1's. Each section describes the specific data utilized. We emphasize that the goal of this section is not to show that our model of the status quo is "correct" (no model is), but rather to suggest that our model is sensible and does a reasonable job of organizing data about a complicated industry.

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 at this common "best bid and offer" (i.e., its share of liquidity) 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. We will discuss these three predictions in turn after describing the data utilized.

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 admittedly-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 publicly-listed stocks and exchange-traded funds in the U.S., 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.[40]

For the results presented in this section, we make the following sample restrictions:

- Time Period. We use data from all trading days in 2015. 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.

- Exchanges. In 2015, the top 5 exchanges by market share all used what is commonly referred to as the "maker-taker" pricing model, in which the taker of liquidity (i.e., the submitter of an order that trades against a resting bid or offer) is charged a fee, and the provider of liquidity (i.e., the resting bid or offer) is paid a rebate. These 5 exchanges together constituted 83% of total trading volume in 2015. The next 3 exchanges by market share all used the "taker-maker" (or "inverted") pricing model, in which the taker of liquidity gets the rebate and the maker pays the fee. These 3 taker-maker exchanges together constituted 15% of total trading volume in 2015. The remaining 4 exchanges active during 2015, sometimes called the "regional" exchanges, together had about 2% market share.[41] For Stylized Facts #1 and #2 we report results for the Top 5 maker-taker exchanges in the main text and report

[40] A disadvantage of this dataset is that it only provides top-of-book information; that is, it does not record non-trade activity (adds or cancels) away from the best bid or best offer on a particular exchange. Unfortunately, direct-feed data, which does provide complete depth-of-book information, are not available for academic research from at least one major exchange family. Since comprehensiveness across exchanges is critical for our purposes, TAQ data was the obvious choice.

[41] Anecdotally, industry participants regard them as vestiges of an earlier era of stock exchange competition.

results for the Top 8 exchanges (i.e., also the taker-maker exchanges) in Appendix B. The Appendix contains additional discussion about the taker-maker exchanges that is helpful for interpreting the Top 8 results. For Stylized Fact #3 we report results for the Top 8 exchanges.

- Symbols. In 2015, there were 9,175 stocks or exchange traded funds that traded at least once; however, most stocks and ETFs trade relatively infrequently.[42] For our main results, we focus on the 100 highest-volume stocks or ETFs that also satisfy a set of data-cleaning filters: trading continuously throughout the year under the same ticker symbol, having a share price of at least $1, not having an exchange listing change, and having at least $10 million in average daily trading volume. These 100 symbols together constitute about one-third of daily volume. 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.

Stylized Fact #1: Many Exchanges Simultaneously at the Best Bid and Best Offer. The first feature of our Stage 3 trading game equilibria that we explore is that all exchanges that have liquidity posted for a given stock do so at the same equilibrium bid and ask.

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:

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

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.[43] 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 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.

We conclude:

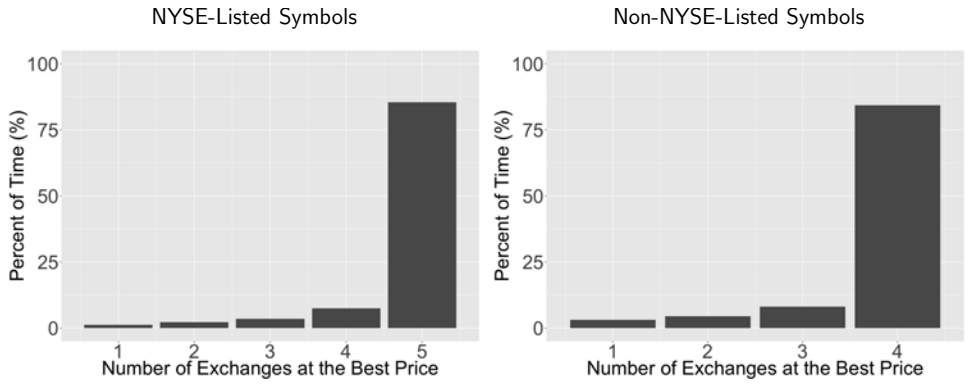
Stylized Fact 1. *At any given moment in time, for highly traded stocks and ETFs, the modal number of exchanges at the best bid and best offer is "all of them." Of the Top 5 exchanges, in about 85% of milliseconds all exchanges are at the best bid (similarly, best offer). It is rare (about 1% of milliseconds 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. A second feature of our Stage 3 equilibria is that "volume follows depth." More precisely, while our analysis is silent as to what determines exchange j's share of

[42] When clear from the context, we will sometimes use the phrase "stocks" to mean both stocks and ETFs. We will also use the phrase "symbol."

[43] NYSE recently (April 2018) changed this practice 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 B for details. Sample is 100 highest volume symbols that satisfy data-cleaning filters (see text for description) on all dates in 2015.

displayed liquidity, it does state that whatever is exchange j's share of displayed liquidity will also be exchange j's share of routed volume. In the notation of Section 3, both are equal to σ_j^*.

Before proceeding, we caveat that at the level of an individual trade this prediction does not hold, nor would we expect it to. Our model, which is deliberately stylized, assumes that all investors demand exactly "1" unit of perfectly-divisible liquidity; this then leads to an equilibrium in which exactly 1 unit of liquidity is offered across all exchanges, so that investors, upon arrival, have no choice but to spread their order across exchanges in order to satisfy their trading demand. In reality, investors of course demand varying amounts of liquidity, and investors who only need to trade a small amount (e.g., 100 shares) often do so with a single small trade on a single exchange. Moreover, investors who need to trade a large amount often break their total desired quantity into smaller individual trades, and these trades are not identified in the data any differently than the trades from investors who only want to trade a small amount. So, volume shares at the trade-by-trade level are often 100% for a single exchange and 0% for all others, which, as 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 (i.e., across many such trades), volume shares should match depth shares — else, the marginal unit of liquidity will be too adversely selected on some exchanges and will be too favorable on others. Also, taking the model a bit less literally, one could interpret the volume share on exchange j as corresponding to the equilibrium probability that an investor routes a small trade to exchange j, if otherwise indifferent. This, too, would lead to the depth-volume relationship obtaining at a higher level of aggregation. We thus explore the depth-volume relationship aggregating all trades in a particular symbol over the course of each trading day in our data. For robustness, we also explore the relationship at higher frequencies than a day, though as noted we would expect that at high-enough frequency the relationship is not meaningful.

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 is calculated straightforwardly as

$$VolumeShare_{ijt} = \frac{Volume_{ijt}}{\sum_{j'} Volume_{ij't}},$$

with $Volume_{ijt}$ the number of shares in symbol i traded on exchange j on date t. We calculate depth share by first computing depth for symbol i on exchange j at each millisecond k within 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 millisecond. We then compute the average depth during the day and the depth share, respectively, as

$$Depth_{ijt} = \frac{1}{T_{it}} \sum_k Depth_{ijtk} \qquad \text{and} \qquad DepthShare_{ijt} = \frac{Depth_{ijt}}{\sum_{j'} Depth_{ij't}},$$

with T_{it} denoting the number of milliseconds for symbol i on date t between the symbol's first quote at or after 9:30 on that date and 16:00 (13:00 on half-days), dropping any milliseconds where the NBBO is locked or crossed. Figure 4.2 presents a scatterplot of $VolumeShare_{ijt}$ against $DepthShare_{ijt}$, wherein each dot in the figure represents a symbol-exchange-date tuple. 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.

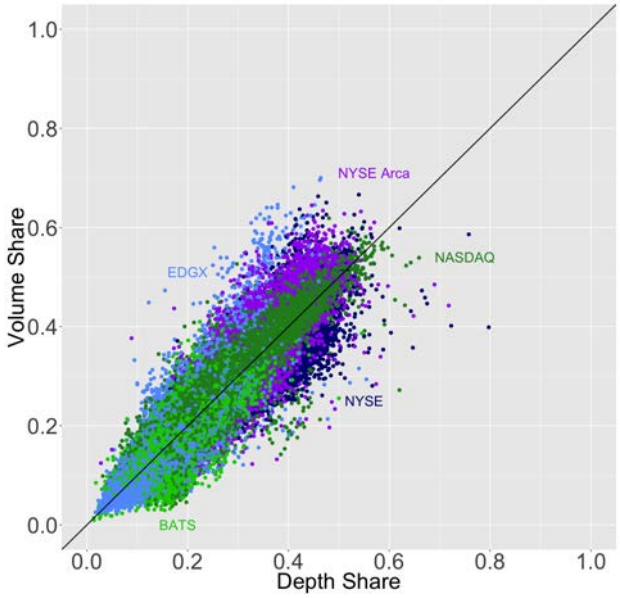
The figure shows that the depth-volume data falls along the 45 degree line for the Top 5 exchanges. The slope of a regression of volume share on depth share is 0.991 (s.e. 0.020), and the R^2 of the relationship is 0.865. In robustness tests, we found that the depth-volume relationship along the 45 degree 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.[44] As emphasized above, 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.

As another 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 level of the individual symbol as well, as should be the case given the theory.

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. This depth-volume relationship does obtain at higher frequencies than a day (e.g., 5 minutes), but breaks down at high-enough frequency (e.g., 1 second). The depth-volume relationship at the daily level holds at both the aggregate and individual-symbol level.*

[44] For our main sample described in the data section above, 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). Focusing on just SPY, the highest-volume symbol in our data, the R^2 is already 0.518 at 30 seconds and is 0.892 at 30 minutes. However, even for SPY, the relationship is extremely noisy at 1 second (R^2 of 0.057). These results are all based on a sample of 12 randomly selected days in 2015.

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.

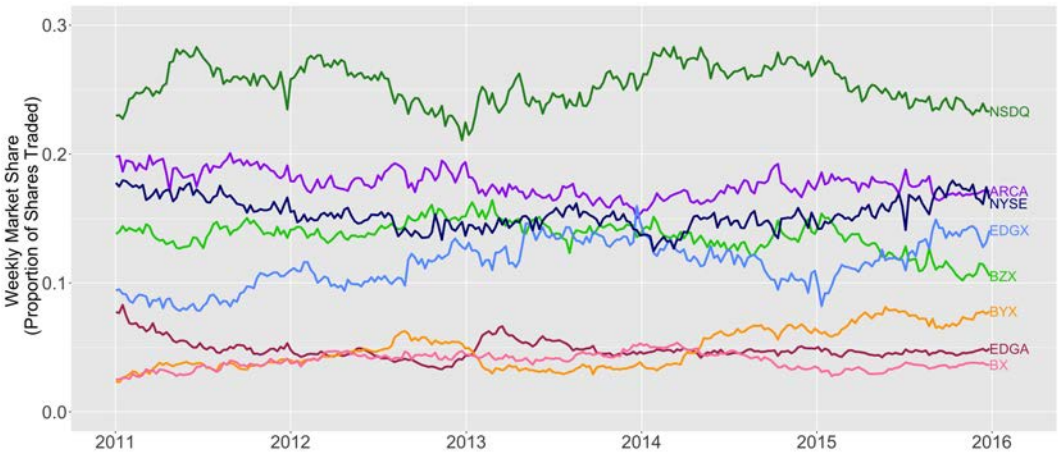
Stylized Fact #3: Exchange Market Shares are Interior and Relatively Stable. The third feature of the trading game equilibria that we explore is that market shares can be interior, i.e., there need not be tipping to one exchange. Furthermore, as discussed in Section 3.3, if investors (or broker-dealers acting on their behalf) use what we called stationary routing table strategies, then these exchange market shares will be stable over time. To be clear, stable market shares are not a prediction of our model. In principle, investors and TFs could coordinate on arbitrarily chaotic market shares. However, since stationary routing table strategies seem both natural in the model and plausible as a description of reality, we think it makes sense to empirically explore both whether exchange market shares are interior and whether they are 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 BATS and Direct Edge were approved as exchanges (prior to that they operated Alternative Trading Systems, or ATS's). Figure B.3 in Appendix B 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 Figure B.4 presents data on average daily exchange market shares for individual stocks. As at the symbol-exchange level, shares are certainly interior. 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

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.

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.

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 (i.e., f)

We now examine the two predictions of our theoretical model regarding exchange trading fees, i.e., f in the model. First, f is 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.[45]

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

[45]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.

at the exchange company level. For example, the exchange company BATS, Inc., controls 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, i.e., 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 (Bats BZX), takers of liquidity during regular-hours trading pay a fee of \$0.0030, while makers of liquidity during regular-hours trading earn a rebate of between \$0.0020 – \$0.0032, where the exact rebate is determined by the maker's volume on Bats BZX (higher volume participants receive larger rebates). Bats BZX's net fee per-trade 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."[46]

Table 4.1 Panel A presents the observed range for "f" 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 or Designated Market Maker program presented in Appendix Table C.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 (Nasdaq, Bats BZX, EDGX, Bats BYX) and an additional 3 with negative minimum fees based on participation in a special fee program (NYSE, NYSE Arca, Nasdaq BX). The maximum observed 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 f, we use the major exchange families' annual financial filings. The advantage of this exercise 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 f. These disadvantages in mind, the results are presented in Table 4.1 Panel B; supporting details are provided in Appendix C.2 and in an associated spreadsheet available in the online appendix.

As can be seen, the average f 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' 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 U.S. equities and options business in 2012, its last full-year of operation before the ICE acquisition, were \$718M.[47] 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 (Table 4.1 Panel A) is, on a per-*

[46] 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.

[47] Source: 2012 NYSE 10-K, page 45, Operating Expenses for the "Cash Trading and Listings" business segment. Nasdaq does not break out its operating expenses by business unit.

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 (BATS Global Markets, Inc., 2015*a*,*b*,*c*,*d*; Nasdaq, Inc., 2015*a*,*b*; NYSE, 2015*a*; NYSE Arca Equities, Inc., 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 C.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 C.2 and the associated spreadsheet for supporting details for these calculations.

share per-side basis, -\$0.00015 to +\$0.00080 or -1.5 mills to +8 mills for regular hours trading on the Top 8 exchanges. The average per-share per-side fee paid for regular hours trading (Table 4.1 *Panel B) is about +\$0.0001. For a \$100 share of stock, the fee in percentage terms is 0.0001%.*

Stylized Fact #5: Money-Pump Constraint Binds. Exchanges have incentive to cut their trading fees even below the perfectly competitive (i.e., zero profit) level in order to win market share and increase revenues from market data and co-location/connectivity. In the language of our model, exchanges are in principle willing to lose money on f in order to make more money from F. However, trading fees are bounded below by a money-pump constraint. In the model, if $f < 0$ there is a money pump: trading firms would engage in infinite volume in order to extract infinite dollars from the exchange with $f < 0$. In practice, the money-pump boundary is below zero, because of SEC Section 31 fees and, for firms that are FINRA members, FINRA fees. At the time of our data, the SEC Section 31 fee was \$21.80 per \$1M traded and the FINRA Trading Activity fee was \$0.000119 per share traded; both fees are assessed on sales but not purchases, i.e., they are assessed on just 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 a $5 stock, the total of the two fees is 2.28 mills to the seller.[48] For a $100 stock, the total of the two fees is 22.99 mills to the seller.

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, where SEC + FINRA fees would be relatively small, 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, as exhibited in Table 4.1, are able to go slightly negative without creating a money pump. Note as well that purposefully exploiting a money-pump with self-trading (e.g., for a very low priced stock) would likely run afoul of securities laws.

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 or -1.5 mills per-share per-side (Nasdaq, BATS BZX, EDGX, BATS BYX; see Table* 4.1*). 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 or -4 mills per-share per-side (NYSE, NYSE Arca, Nasdaq BX; see Appendix Table* C.1*). 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 (i.e., F)

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 3.4).

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. For more details on the data used throughout this section, see Appendix D.

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 in such granularity. Neither Nasdaq nor NYSE

[48] 1.09 mills (5×0.218 mills) of SEC fees + 1.19 mills of FINRA fees.

report U.S. equities revenue with the granularity of BATS in 2015, so for those exchanges we have to 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.[49] This means that the BATS U.S. Equities business is profitable with market 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, information on the proportion of global revenue that comes from the U.S., and information about the proportion of U.S. revenue that comes from equities as opposed to 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 D.

Table 4.2 summarizes our analysis of 2015 U.S. equities market data and co-location/connectivity revenues. 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 consortium data products or the SIP feed (see footnote 19). Proprietary data feeds are utilized by latency-sensitive market participants, whereas the market-wide consortium data products are 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.[50] 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.[51]

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 for discussion), 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*

[49] 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).

[50] 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.

[51] 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 D.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 (U.S. Securities and Exchange Commission, 2014).

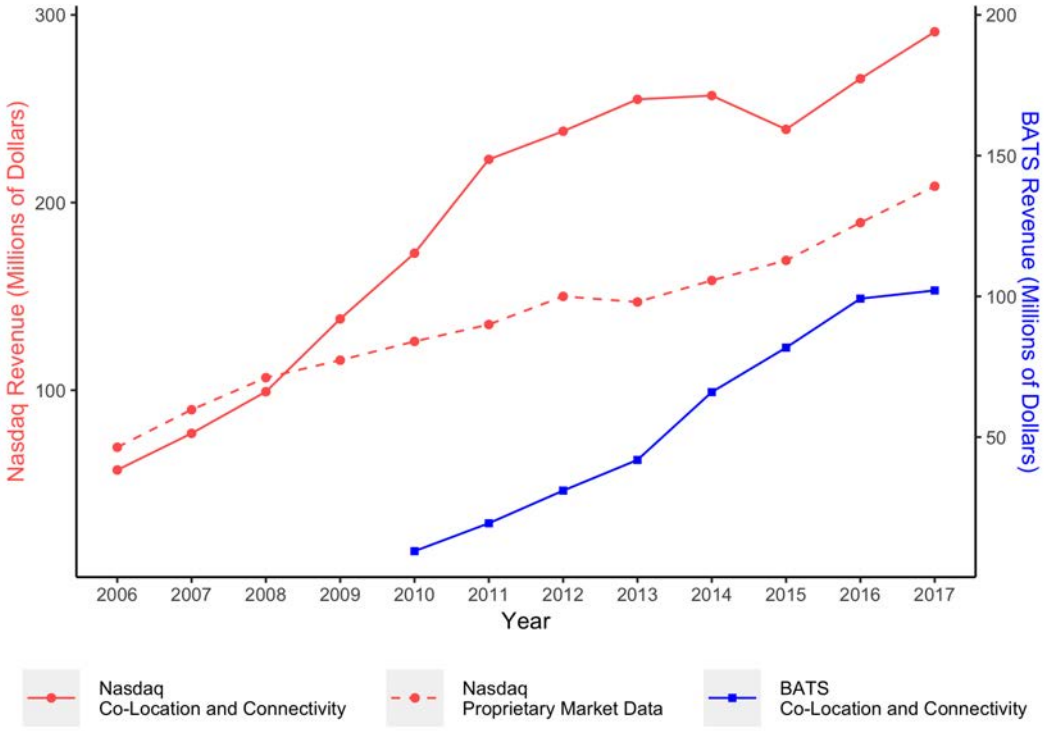
that data suggest that in 2015 total ESST revenue was between $675-790M. This is several times larger than regular-hours trading revenues.

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, and making some modest adjustments to attempt to provide apples-to-apples numbers over time. We are able to build meaningful time-series for Nasdaq co-location/connectivity and proprietary market data revenues and for BATS co-location/connectivity revenues. BATS only began charging for the proprietary market data that we think of as part of ESST relatively recently (Q3 2014). NYSE's financial reporting segments unfortunately changed too frequently in the Reg NMS era for the exercise to be instructive.[52]

For our analysis of Nasdaq we go back as far as 2006. 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). Nasdaq co-location and connectivity revenue was contained in the financial reporting category "Access Service Revenues" from 2006-2012, "Access and Broker Services Revenues" from 2013-2015, and "Trade Management Services Revenues" for 2016-2017. Nasdaq's segment reporting practices underwent some modest changes in fiscal years 2013 and 2015, with some revenue streams added to the category in 2013 and removed from the category in 2015, whereas the periods 2006-2012 and 2015-2017 appear to yield more reliable apples-to-apples growth rates for the reporting segment. Revenue in the category quadrupled in the 2006-2012 period (growth of 26.7% per year), was roughly flat in the 2012-2015 period, and then in 2015-2017 growth was 10.3% per year. Overall growth from 2006-2017 has been 15.9%

[52]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 CBOE/BATS Merger Proxy (Cboe Holdings, Inc. and BATS Global Markets, Inc., 2016) (2016) 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 D.2 for further discussion of all of these data.

per year.

For Nasdaq market data, from 2006-2012 Nasdaq separately reported revenue from U.S. tape plans and revenue from the U.S. proprietary market data products that we think of as part of ESST (see discussion in Stylized Fact #6). Notably, in 2006, tape revenue was nearly double Nasdaq's proprietary market data revenues ($129M "Net U.S. Tape Plan" vs. $70M "U.S. Market Data Products") whereas by 2012 proprietary revenue was about 30% higher than tape ($117M vs. $150M). Starting in 2013 Nasdaq reported only combined data revenue (i.e., tape plus proprietary), and made some other segment reporting changes. Starting in 2014, they reported only global data revenue. To get a roughly apples-to-apples time series for U.S. proprietary market data we make two adjustments. First, from 2014 onwards we use the 2013 ratio of U.S. to global market data revenue (72%) to get a U.S. market data revenue estimate. Second, from 2013 onwards we subtract out 2012 tape revenue of $117M to get to U.S. market data revenue excluding tape plan revenue.[53,54] We estimate that revenue growth for U.S. proprietary market data was 13.7% annually for the period 2006-2012 and 6.8% for the period 2012-2017, with the caveat that the 2006-2012 growth rate is based off of numbers directly in Nasdaq filings and the 2012-2017 growth rate is based on additional

[53]Nasdaq reports tape revenue of $117M in 2010, $115M in 2011, and $117M in 2012, and Nasdaq's market share is relatively flat from 2012 to 2017.

[54]These two assumptions together imply that any revenue growth in Nasdaq's global market data category since 2014 is attributed 72% to U.S. proprietary market data with the remaining 28% to Nasdaq's other sources of data revenues, international and index. Our sense from Nasdaq's financial reports is that this convention is conservative for our purposes of providing a sense of magnitudes for U.S. proprietary market data revenue growth. Please see further discussion in the Appendix.

assumptions. Overall growth from 2006-2017 has been 10.5% per year, and growth in the period since our SF#6 data, 2015-2017, has been 11.1% per year.

For BATS co-location and connectivity, the 2012 S-1 filing[55] reports that they began charging for co-location/connectivity in Q4 2009. Initially the revenue was reported in a segment called "Other Revenues,"[56] and starting in 2012 the segment "Port Fees and Other," and then in 2016, as part of CBOE, "Connectivity Fees and Other." From 2010, the first full year in which BATS earned these revenues, through 2013, the last full year before the Direct Edge acquisition, revenue more than quadrupled (growth of 64.0% per year). Revenue then doubled again from 2013 to 2015, but likely in large part due to the Direct Edge acquisition, and then grew 11.7% per year from 2015 to 2017. Overall growth from 2010-2017 has been 40.4% per year, but we caveat that this figure is inflated due to the Direct Edge acquisition.

For BATS market data, BATS reports in its 2016 S-1 filing that it only began charging for proprietary market data on the two BATS exchanges, BZX and BYX, in the 3rd quarter of 2014. Triangulating between 2015 data from the S-1 filing, 2016 data from the BATS/CBOE merger proxy, and 2017 data from CBOE's annual report, 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. We discuss the limited data that is available in more detail in the appendix.

Please see Appendix D.2 for further details of all of these data sources.

Overall the data, while imperfect and we hope appropriately caveated, 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 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

While our model is of course stylized and abstracts from many important issues such as agency frictions, tick-size frictions, fee complexity, strategic trading over time, etc., it does a reasonably good job empirically. In particular, Stylized Facts #1-#3 are broadly consistent with our equilibrium characterization of the trading game, Stylized Facts #4-#5 are consistent with the equilibrium prediction that trading fees are competitive and bound by the money-pump constraint, and Stylized Facts #6-#7 are consistent with the equilibrium prediction that exchanges earn significant economic rents from exchange-specific speed technology in the modern era of stock trading.

[55] BATS initially filed to go public in 2012 but then pulled the offering.

[56] Described as "Other revenues consist of port fees, which represent fees paid for connectivity to our markets, and, more recently, additional value-added products revenues [likely, co-location]."

We now discuss other potential models of exchange competition, many of which do not incorporate aspects of the U.S. equity regulatory environment and are at odds with certain 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 modern U.S. equity exchange competition specifically.

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.[57] 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. Last, these models are directly at odds with aspects of the regulatory environment for U.S. equities. Specifically, Reg NMS implies that market participants all multi-home, and UTP implies that all securities, in essence, multi-home.

The next class of models we discuss 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.[58] 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 as well as the regulatory environment for U.S. equities.

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

[57] 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.

[58] 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 (2018)). If Baldauf and Mollner (2018, 2019) are correct that the latency-arbitrage pie grows with the number of exchanges, that should only strengthen the arguments we make in Section 5.

"fills in the penny" for the market (i.e., makes tick-size constraints less binding) and gives exchanges market power. However, this model is inconsistent with the fact that the Top 5 exchanges, which control 83% of volume, all use essentially the same fee structure (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 (intuitively, to maximize distance on the Salop circle). This model is also inconsistent with trading fees being competitive and bound by a money-pump constraint, i.e., with our Stylized Facts #4-#5.[59] That said, while inconsistent with several aspects of the data, this model does seem central for understanding the co-existence of the maker-taker fee model and the taker-maker fee model.

Last, we emphasize that the Pagano (1989) framework seems consistent with many aspects of modern futures exchanges. The crucial difference between futures and equities, as emphasized earlier, is that futures contracts are proprietary to a particular exchange, i.e., there is no analog of UTP. As a result, futures exchanges are able to charge meaningful fees that exploit network effects.[60] Developing a better understanding of the IO of modern futures exchanges, and their incentives for market design innovation, seems a fruitful topic for future research.

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. In this section we now use the 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.[61] 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

[59] The Chao, Yao and Ye (2019) model may be inconsistent with Stylized Fact #1 as well, in the sense that it suggests that at any moment in time, only the exchange whose fee structure is "just right" given where fundamental value lies within the penny should have liquidity. Other exchanges' differentiated fees cause it to be impossible to provide liquidity at the right price within the penny.

[60] For example, whereas we showed in the discussion of Stylized Fact #4 that U.S. stock exchange trading fees are economically small and do not appear to cover exchange operating costs, the Chicago Mercantile Exchange's 2015 trading plus clearing revenue was $2,784M (84% of CME's total revenues), and the CME was significantly profitable on the basis of trading/clearing revenue less operating expenses alone ($1,446M of profit on this basis). (CME Group, Inc., 2016)

[61] The analysis can also be applied to the "asymmetric speed bump" or "asymmetric delay" market design, in which the exchange processes cancellations immediately upon receipt but processes new orders only after a fixed small delay (e.g., 350 microseconds). 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 technologically sophisticated enough to update within the delay window. 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 (2018) for a detailed theoretical analysis.

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). We then describe how we incorporate it formally into our model of exchange competition introduced in Section 3.

The FBA market design is similar to the continuous limit order book market design in many 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, and (iii) orders remain outstanding until either executed or canceled. As will become clear, two additional similarities are (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 (e.g., of 1 millisecond), and treat all orders received in the same interval as having been received at the same time. A way to think about this is that time is treated as a discrete variable rather than as a continuous variable. Second, orders are batch processed at the end of each discrete-time interval, using a uniform-price auction, rather than being serially processed upon receipt as in a limit order book. 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.[62] 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).

Formally, these differences relate to how an FBA exchange processes and prioritizes orders in the Stage 3 trading game of our model introduced in Section 3. We assume that an FBA exchange, which we call "Discrete" in our subsequent 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 during the batch interval), 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.

Everything else about the Stage 3 trading game is exactly 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

[62] 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.

as before by sending IOC orders and cancel messages to all exchanges, but because the Discrete exchange will process this activity in batch, there are no longer latency arbitrage rents from public information.

The only additional modification we make to our model of exchange competition from Section 3 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.[63]

In summary, in Stage 1, Continuous exchanges set both trading and ESST fees, while Discrete exchanges only set trading fees; in Stage 2, TFs make ESST adoption decisions for Continuous exchanges; and in Stage 3, the infinitely repeated trading game is played with the modifications described above. We also maintain the same solution concept as before: order book equilibrium for Stage 3, and subgame perfect Nash equilibrium for Stages 1 and 2 given anticipated Stage 3 play.

Discussion of the Batch Interval. We emphasized in the introduction and above that our analysis is of FBAs with very fast batch intervals. 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 (2019) is consistent with these magnitudes being enough to eliminate most latency arbitrage.

We caution that our model of competition between FBA exchanges and the continuous market would become more practically strained with a longer batch interval. One reason is that for an FBA exchange to have protected quotes under Reg NMS, the batch interval must satisfy the SEC's *de minimis* delay standard as discussed in Section 2.2 (see U.S. Securities and Exchange Commission, 2016*c*). With a longer 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.[64]

Thus, while we think there are good economic arguments for a longer batch interval if redesigning markets from scratch, in the context of competition with continuous markets under Reg NMS it is substantively important that we focus on very fast batch intervals.

[63] 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.

[64] As noted above in footnote 20, 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.

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.[65] 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, 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 Stage 3 trading game 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 trading firms have purchased exchange-specific speed technology from Continuous. In <u>any</u> order book equilibrium of the Stage 3 trading game, exactly one unit of liquidity is provided on Discrete at bid-ask spread $s^*_{discrete}$ around y following Period 1, and no liquidity is provided on the Continuous exchange. 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 $\lambda_{invest}\frac{\bar{s}_{continuous}}{2} - (\frac{N-1}{N}\lambda_{public} + \lambda_{private}) \cdot L(\bar{s}_{continuous}) = 0$.[66] 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 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

[65] 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.

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

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 $f > 0$.[67]

Given these results, we establish the following:

Proposition 5.2. *In <u>any</u> equilibrium of the overall multiple exchange game (i.e., Stages 1-3) with a single Continuous exchange and single Discrete exchange, (i) Discrete charges strictly positive trading fees; (ii) in every iteration of the trading game, exactly one unit of liquidity is provided on Discrete around y following Period 1, and no liquidity is provided elsewhere; (iii) Continuous earns zero profits; and (iv) Discrete 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.[68]

Please note that there are several reasons not to take the 100% aspect of Propositions 5.1-5.2 literally, especially tick-size constraints and agency frictions. We discuss such frictions in more detail below in Section 6.3.

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. *In <u>any</u> equilibrium of the full multi-exchange game with at least two Discrete exchanges, (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 y following Period 1; and (iii) all exchanges and trading firms earn zero economic 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 always provided

[67] Let $\bar{s}_{discrete}(f)$ (defined formally in the Appendix in (A.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.

[68] In the Glosten (1994) model, because investors sometimes consume large quantities at once, there is another difference between the limit order book and batch auctions which is that limit order books are pay-as-bid, or "discriminatory price", whereas in batch auctions all trade is cleared at the same market-clearing price. This difference is not present in our model because all trades are for "1" unit at a time. We view this modeling convention as appropriate given that modern investors commonly shred large orders into small orders placed in the market over time (see Kyle, Obizhaeva and Wang, 2018). That said, as we emphasize in the conclusion, incorporating investors who trade strategically over time as in Kyle-style models is an important direction for future research. In particular, it would be interesting to understand whether there are equilibrium differences between frequent batch auctions and asymmetric delay mechanisms — both of which eliminate latency arbitrage, but are uniform-price and discriminatory-price, respectively.

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 3.4 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 A.4 in the appendix) yields that $\Pi^D > NF^*_j$ for all exchanges j, for any equilibrium ESST revenues consistent with Proposition 3.4 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.[69] For the purpose of our analysis

[69] 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.

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. For example, if the fixed cost of adoption is \$100M, and we use a per-trading game discount factor of $\delta < 1$, then $(\sum_{t=0}^{\infty} \delta^t) c_{adopt} = \$100\text{M}$.

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.[70]

Adoption Incentives: A New Entrant Exchange. We first examine the adoption incentives for a de novo entrant exchange given the status quo where all incumbent exchanges employ Continuous.

Consider what would happen if a de novo entrant incurred entry and adoption costs to start a new Discrete exchange. In our model, all trading activity would then shift to the Discrete exchange, and Discrete earns revenues equal to Π^D per-trading game as long as all other exchanges remain Continuous. However, given that at least one incumbent exchange would imitate Discrete when able to do so, after T periods the entrant would earn zero profits. Hence, an entrant exchange would find entry profitable only if its expected revenues from entry (i.e., T periods of being the only Discrete exchange, followed by being one of multiple Discrete exchanges) exceeds its entry and adoption costs. 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 out of an infinitely repeated series of trading games.[71] The condition for a de novo to find it profitable to enter is:

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

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.

[70] 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 $\sigma_j^D R - c_{imitate} > 0$.

[71] 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 $\frac{EML}{ETL}$. In this context, if Discrete does not literally get 100% share when competing against Continuous but also does not go literally to zero profits when an incumbent imitates, ρ can be interpreted more broadly as the ratio of present-discounted private profits to present-discounted social value.

Adoption Incentives: An Incumbent Exchange. We next ask whether any incumbent exchange, when all exchanges are operating as Continuous, would wish to adopt Discrete. As emphasized, incumbents prefer the all-Continuous market to the all-Discrete market because of ESST revenues. Yet, if adoption costs of Discrete are sufficiently low and if rival exchanges cannot imitate without substantial delay, an incumbent exchange may find it profitable to adopt. The relevant condition for exchange j is:

$$\rho\Pi^D \geq c_{adopt} + NF_j^*. \tag{5.3}$$

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 it does not include entry costs, c_{entry}, but instead includes the foregone status-quo profits from remaining a Continuous exchange, NF_j^*. Since the net present value of large incumbents' ESST revenues is likely significantly larger than a de novo's cost of entry (i.e., $NF_j^* > c_{entry}$), condition (5.3) is likely more restrictive than condition (5.2). That is, incumbents with existing profits to protect may be less likely than de novo entrants to adopt.

That incumbent exchanges continue to operate Continuous is thus consistent with them maintaining the "cooperative" all-Continuous outcome of a repeated prisoner's dilemma. 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].*"[72]

(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 analysis in Section 5 implies that there are two key wedges between the private and social incentives to adopt market designs that address latency arbitrage and the associated arms race for speed.

First, a market design innovator, whether a de novo entrant or an incumbent, earns profits commensurate with the social value of eliminating latency arbitrage only for the period of time before the market design innovation is imitated. The market gets a benefit of $\Pi^*_{continuous}$ in perpetuity, but the innovator only gets a benefit commensurate with $\Pi^*_{continuous}$ for proportion ρ of time. Anecdotal evidence suggests that ρ might be quite small in practice.

[72] 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).

Second, incumbents face an additional wedge between private and social incentives, because they anticipate losing the net present value of economic rents from the speed race. This anticipated cost is reflected in the right-hand side of equation (5.3).

These gaps between private and social incentives are reasons why it is possible that private-market forces alone may not lead to adoption of market designs that address latency arbitrage and the arms race. On the other hand, the analysis in Section 5 indicates that if there is such an entrant, it will gain share. Taking the analysis of Section 5 literally, an initial entrant helps move the whole industry from the equilibrium with latency arbitrage to the equilibrium without latency arbitrage.

This suggests that a potential option for policy is to provide a "push" rather than a market design mandate. By push we mean any policy that tips the balance of incentives sufficiently to cause either condition (5.2) or (5.3) to obtain.[73] 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;[74] that is, costs that are part of what we call c_{entry}. As evidence of the significant risk of a new stock exchange design not getting approved, again consider IEX and CHX. IEX went through a protracted fight over its exchange design, and ultimately had to make concessions such that, for the main part of its market (in industry parlance, the "lit" exchange part as opposed to the "dark" alternative trading system part), its market design was essentially a standard continuous limit order book (Budish, 2016*b*). CHX, too, went through a protracted fight over its proposed exchange design (which incorporated an asymmetric speed bump as modeled in Baldauf and Mollner, 2018), was essentially rejected by the SEC (CHX, 2017, 2018; U.S. Securities and Exchange Commission, 2017*c*), and then instead sold its exchange to NYSE (Michaels and Osipovich, 2018). Both of these considerations raise the risk-adjusted cost of entering as a new exchange with a new market design.

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).

[73] 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. 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.

[74] The Wall Street Journal reported, of the merger, "Analysts say CHX's most valuable asset is its license to run a national securities exchange. Applying for a new exchange license from the SEC can take years" (Michaels and Osipovich, 2018). At an industry conference attended by one of the authors around that time, numerous industry participants referred to CHX's value to NYSE as coming entirely from its "medallion," i.e., its license to run a stock exchange.

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, taking the model seriously, 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.[75]

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

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

Here, patents do not seem to be a viable way to create market exclusivity, for at least two reasons. First, the specific market design of frequent batch auctions is in the public domain. Second, even if frequent batch auctions were patented, to be effective the intellectual property protection would have to cover all possible market designs that eliminate latency arbitrage. As evidence of the difficulty of this, consider that the Chicago Mercantile Exchange filed for a patent (Hosman et al., 2017) in Jan 2016 for a market design idea that a close reader will recognize as, in essence, a form of batch auction, without using the word "auction" a single time.[76]

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

[75] 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 Budish (2016*b*) regarding the strengths and limitations of IEX's speed bump design.

[76] 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).

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 (2019) 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 3.4) also point to annual latency arbitrage profits of roughly this magnitude; see the discussion in Section 4.3. A discount rate of between 5% and 10% applied to the $2 billion per year figure implies that the net present value of latency arbitrage profits in U.S. equity markets is on the order of $20 to $40 billion. While admittedly rough, and based on extrapolation from UK equities which may not be comparable to U.S. equities, this gives a sense of magnitudes for the benefits of addressing latency arbitrage in the U.S. stock market.

This magnitude suggests that an entry subsidy almost surely passes a cost-benefit test. Getting a sense of magnitudes for the exclusivity period is more difficult. If we take Proposition 5.2 literally then even a very short exclusivity period would suffice. In that result, the adopter earns profits greater than $\frac{N-1}{N}\Pi^*_{continuous}$. If we use $\Pi^*_{continuous} = \$2$ billion annually and assume that entry plus adoption costs are about $100 million, this implies that an exclusivity period of as little as one month is sufficient to induce entry by a de novo. There are several reasons not to take the 100% tipping conclusion of Proposition 5.2 literally: most centrally, 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. In Appendix E, we analyze a scenario with tick-size constraints under the assumption that market participants always break ties in favor of Continuous over Discrete, i.e., they only trade on Discrete when they save a full tick and hence are mandated to do so under Reg NMS. This scenario is meant to capture potential agency frictions in a simple and worst-case way for Discrete. In this scenario, the Discrete exchange's market share with zero fee is about 20% instead of 100%, and with the revenue-maximizing fee Discrete's revenue is on the order of 5% of $\Pi^*_{continuous}$ instead of $\frac{N-1}{N}$. These numbers suggest that an exclusivity period on the order of 1-2 years might be sufficient to induce entry by a de novo. Clearly, both 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, but 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 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."

These ideas are already having some impact. The SEC recently (Oct 2019) 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.

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 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.[77] 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).

We also emphasize that while our model of U.S. stock exchanges is already reasonably complicated, there is much left out that would be valuable to incorporate in future research. We discussed in the main text the importance of tick-size constraints and asymmetric trading fees, a topic currently the subject of an SEC pilot test. It would also be valuable to incorporate investors with richer trading needs and information structures — e.g., institutional investors wishing to trade large quantities over a period of time, who need to trade off speed, price impact, and the risk of being detected (Kyle, 1985, 1989) — and the role of the broker-dealers who compete to serve them, and as such play a central role in directing trading volume. Such extensions, in addition to being of interest per se, may also shed light on the determination of equilibrium exchange market shares, which the current analysis is silent on (though it does yield an understanding of why such market shares may be relatively stable over time). Last, we hope that the model, given its novel focus on the source of exchange profits, will prove a useful starting point for future research related to the entry, merger, and investment incentives of stock exchanges, and can be a useful input into the recent policy debate about rising stock exchange market-data and co-location fees (Clayton, 2018; Jackson Jr., 2018; U.S. Securities and Exchange Commission, 2018*b*).

[77] 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

15 U.S.C. § 78a. 1934. Securities Exchange Act of 1934.

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.

Amihud, Yakov, and Haim Mendelson. 1996. "A New Approach to the Regulation of Trading Across Securities Markets." *New York University Law Review*, 71(6): 1411–1466.

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

Aquilina, Matteo, Eric Budish, and Peter O'Neill. 2019. "Quantifying the High-Frequency Trading "Arms Race": A Simple New Methodology and Estimates." Working Paper.

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.

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. 2018. "High-Frequency Trading and Market Performance." *Journal of Finance*. forthcoming.

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

BATS Global Markets, Inc. 2012. "Form S-1." Retrieved September 10, 2018 from https://www.sec.gov/Archives/edgar/data/1519917/000119312512125661/d179347ds1a.htm.

BATS Global Markets, Inc. 2015*a*. "BATS BYX Exchange Fee Schedule." Retrieved April 11, 2017 from http://www.bats.com/us/equities/membership/fee_schedule/byx/ through Wayback Machine (archived on February 28, 2015).

BATS Global Markets, Inc. 2015*b*. "BATS BZX Exchange Fee Schedule." Retrieved April 12, 2017 from http://www.bats.com/us/equities/membership/fee_schedule/bzx/ through Wayback Machine (archived on February 28, 2015).

BATS Global Markets, Inc. 2015*c*. "EDGA Exchange Fee Schedule." Retrieved April 11, 2017 from http://www.bats.com/us/equities/membership/fee_schedule/edga/ through Wayback Machine (archived on February 28, 2015).

BATS Global Markets, Inc. 2015*d*. "EDGX Exchange Fee Schedule." Retrieved April 11, 2017 from http://www.bats.com/us/equities/membership/fee_schedule/edgx/ through Wayback Machine (archived on April 27, 2015).

BATS Global Markets, Inc. 2016. "Form S-1." Retrieved September 12, 2018 from https://www.sec.gov/Archives/edgar/data/1659228/000104746916012191/a2228256zs-1a.htm.

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.

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*a*. "Re: Chicago Stock Exchange Liquidity Taking Access Delay (Release No. 34-78860; SR-CHX-2016-16)." Retrieved February 12, 2019 from https://www.sec.gov/comments/sr-chx-2016-16/chx201616-9.pdf.

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. 2019. "How to Fix the Market for Event Tickets." June 11. Keynote address to FTC workshop "That's the Ticket", Washington, D.C. Retrieved October 25, 2019 from https://faculty.chicagobooth.edu/eric.budish/research/Budish-FTC-Keynote.pptx.

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.

Canadian Securities Administrators. 2009. "Notice of Amendments to National Instrument 21-101 Marketplace Operation and National Instrument 23-101 Trading Rules." *Ontario Securities Commission*, Retrieved April 17, 2019 from `http://www.osc.gov.on.ca/documents/en/Securities-Category2/rule_20091113_21-101_new-noa-21-101and23-101.pdf`.

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.

Cboe Global Markets, Inc. 2018. "Fiscal Year 2017 10-K." Retrieved September 7, 2018 from `http://ir.cboe.com/~/media/Files/C/CBOE-IR-V2/documents/annual-proxy/2017-annual-report-and-form-10-k.pdf`.

Cboe Holdings, Inc. and BATS Global Markets, Inc. 2016. "Joint Proxy Statement on Merger Agreement." Retrieved September 10, 2018 from `http://ir.cboe.com/~/media/Files/C/CBOE-IR-V2/documents/special-proxy/joint-proxy-statement.pdf`.

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

Chao, Yong, Chen Yao, and Mao Ye. 2017. "Discrete Pricing and Market Fragmentation: a Tale of Two-Sided Markets." *American Economic Review: Papers and Proceedings*, 107(5): 196–199.

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.

Citadel. 2016. "Re: Proposed CHX Liquidity Taking Access Delay (Release No. 34-78860; File No. SRCHX-2016-16)." Retrieved February 15, 2019 from https://www.sec.gov/comments/sr-chx-2016-16/chx201616-7.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.

CME Group, Inc. 2016. "Fiscal Year 2015 10-K." Retrieved February 7, 2018 from https://www.sec.gov/Archives/edgar/data/1156375/000115637516000116/cme-2015123110k.htm.

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.

Diamond, Peter A. 1971. "A Model of Price Adjustment." *Journal of Economic Theory*, 3(2): 156–168.

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., 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.

Hasbrouck, Joel, George Sofianos, and Deborah Sosebee. 1993. "New York Stock Exchange Systems and Trading Procedures." NYSE Working Paper No. 93-01.

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.

Intercontinental Exchange, Inc. 2016. "Fiscal Year 2015 10-K." Retrieved September 18, 2018 from https://www.sec.gov/Archives/edgar/data/1571949/000157194916000020/ice2015123110k.htm.

Jackson Jr., Robert J. 2018. "Unfair Exchange: The State of America's Stock Markets." September 19. Speech at George Mason University, Arlington, Virginia. Retrieved January 11, 2019 from https://www.sec.gov/news/speech/jackson-unfair-exchange-state-americas-stock-markets.

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, Christopher A. Neilson, and Seth D. Zimmerman. 2019. "Heterogeneous Beliefs and School Choice Mechanisms." NBER Working Paper No. 25096.

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. 1989. "Informed Speculation with Imperfect Competition." *The Review of Economic Studies*, 56(3): 317–355.

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.

Nasdaq, Inc. 2015*a*. "Nasdaq BX Fee Schedule." Retrieved April 11, 2017 from `https://www.nasdaqtrader.com/Trader.aspx?id=bx_pricing` through Wayback Machine (archived on April 1, 2015).

Nasdaq, Inc. 2015*b*. "Price List - Trading Connectivity." Retrieved April 12, 2017 from `http://www.nasdaqtrader.com/Trader.aspx?id=PriceListTrading2` through Wayback Machine (archived on April 8, 2015).

Nasdaq, Inc. 2016. "Fiscal Year 2015 10-K." Retrieved October 16, 2018 from `https://www.sec.gov/Archives/edgar/data/1120193/000112019316000020/ndaq-20151231x10k.htm`.

Nasdaq, Inc. 2017. "Fiscal Year 2016 10-K." Retrieved September 10, 2018 from `https://www.sec.gov/Archives/edgar/data/1120193/000112019317000003/ndaq1231201610-k.htm`.

Nasdaq, Inc. 2018. "Fiscal Year 2017 10-K." Retrieved September 10, 2018 from `https://www.sec.gov/Archives/edgar/data/1120193/000112019318000003/ndaq1231201710-k.htm`.

Nasdaq OMX Group, Inc. 2010. "Fiscal Year 2009 10-K." Retrieved September 10, 2018 from `https://www.sec.gov/Archives/edgar/data/1120193/000119312510034340/d10k.htm`.

Nasdaq OMX Group, Inc. 2011. "Fiscal Year 2010 10-K." Retrieved September 10, 2018 from `https://www.sec.gov/Archives/edgar/data/1120193/000119312511045348/d10k.htm`.

Nasdaq OMX Group, Inc. 2012. "Fiscal Year 2011 10-K." Retrieved September 10, 2018 from `https://www.sec.gov/Archives/edgar/data/1120193/000119312512077518/d259668d10k.htm`.

Nasdaq OMX Group, Inc. 2013. "Fiscal Year 2012 10-K." Retrieved September 10, 2018 from `https://www.sec.gov/Archives/edgar/data/1120193/000119312513069357/d445717d10k.htm`.

Nasdaq OMX Group, Inc. 2014. "Fiscal Year 2013 10-K." Retrieved September 10, 2018 from `https://www.sec.gov/Archives/edgar/data/1120193/000112019314000002/ndaq-20131231x10k.htm`.

Nasdaq OMX Group, Inc. 2015. "Fiscal Year 2014 10-K." Retrieved September 10, 2018 from `https://www.sec.gov/Archives/edgar/data/1120193/000112019315000003/ndaq-20141231x10k.htm`.

Nasdaq Stock Market, Inc. 2007. "Fiscal Year 2006 10-K." Retrieved September 7, 2018 from `https://www.sec.gov/Archives/edgar/data/1120193/000119312507042803/d10k.htm`.

Nasdaq Stock Market, Inc. 2008. "Fiscal Year 2007 10-K." Retrieved September 10, 2018 from https://www.sec.gov/Archives/edgar/data/1120193/000119312508037364/d10k.htm.

Nasdaq Stock Market, Inc. 2009. "Fiscal Year 2008 10-K." Retrieved September 10, 2018 from https://www.sec.gov/Archives/edgar/data/1120193/000119312509039333/d10k.htm.

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

NYSE. 2015*a*. "Price List 2015." Retrieved April 12, 2017 from https://www.nyse.com/publicdocs/nyse/markets/nyse/NYSE_Price_List.pdf through Wayback Machine (archived on September 1, 2015).

NYSE. 2015*b*. "Re: Investors' Exchange LLC Form 1 Application (Release No. 34-75925; File No. 10-222)." Retrieved January 24, 2019 from https://www.sec.gov/comments/10-222/10222-19.pdf.

NYSE Arca Equities, Inc. 2015. "Schedule of Fees and Charges for Exchange Services." Retrieved April 19, 2017 from https://www.nyse.com/publicdocs/nyse/markets/nyse-arca/NYSE_Arca_Marketplace_Fees.pdf through Wayback Machine (archived on August 3, 2015).

NYSE Euronext. 2013. "Fiscal Year 2012 10-K." Retrieved September 13, 2018 from https://www.sec.gov/Archives/edgar/data/1368007/000136800713000005/nyx-20121231x10k.htm.

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.

O'Hara, Maureen, and Jonathan R. Macey. 1997. "The Law and Economics of Best Execution." *Journal of Financial Intermediation*, 6(3): 188–223.

O'Hara, Maureen, and Mao Ye. 2011. "Is Market Fragmentation Harming Market Quality?" *Journal of Financial Economics*, 100(3): 459–474.

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.

Petrella, Giovanni. 2010. "MiFID, Reg NMS and Competition across Trading Venues in Europe and the USA." *Journal of Financial Regulation and Compliance*, 18(3): 257–271.

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. 2018. "How Market Design Emerged from Game Theory." Stanford University Working Paper.

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.

Sannikov, Yuliy, and Andrzej Skrzypacz. 2016. "Dynamic Trading: Price Inertia and Front-Running." Working Paper.

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

Schneiderman, Eric. 2014. "Remarks on High-Frequency Trading and Insider Trading 2.0." Remarks to New York Law School Panel on "Insider Trading 2.0 – A New Initiative to Crack Down on Predatory Practices". Retrieved February 19, 2019 from `http://www.ag.ny.gov/pdfs/HFT_and_market_structure.pdf`.

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

U.S. Congress. 1994. "Unlisted Trading Privileges Act of 1994." H.R. 4535. 103rd Congress. Retrieved January 4, 2019 from `https://www.congress.gov/bill/103rd-congress/house-bill/4535/text`.

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. 2000. "Final Rule: Unlisted Trading Privileges, SEC Release No. 34-43217." Retrieved December 22, 2018 from `https://www.sec.gov/rules/final/34-43217.htm`.

U.S. Securities and Exchange Commission. 2005. "Regulation NMS, SEC Release No. 34-51808." Retrieved November 14, 2018 from https://www.sec.gov/rules/final/34-51808.pdf.

U.S. Securities and Exchange Commission. 2014. "Fee Amendments of the Consolidated Tape Association Plan and Consolidated Quotation Plan, SEC Release No. 34-73278." Retrieved November 9, 2018 from https://www.sec.gov/rules/sro/nms/2014/34-73278.pdf.

U.S. Securities and Exchange Commission. 2016*a*. "Comments on Investors' Exchange LLC; Notice of Filing of Application, as Amended, for Registration as a National Securities Exchange under Section 6 of the Securities Exchange Act of 1934." Retrieved December 22, 2018 from https://www.sec.gov/comments/10-222/10-222.shtml.

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*a*. "Comments on CHX Rulemaking: Notice of Filing of Proposed Rule Change to Adopt the CHX Liquidity Taking Access Delay." Retrieved December 22, 2018 from https://www.sec.gov/comments/sr-chx-2016-16/chx201616.shtml.

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. 2018*a*. "Comments on CHX Rulemaking: Notice of Filing of Proposed Rule Change to Adopt the CHX Liquidity Enhancing Access Delay." Retrieved December 22, 2018 from https://www.sec.gov/comments/sr-chx-2017-04/chx201704.htm.

U.S. Securities and Exchange Commission. 2018*b*. "Roundtable on Market Data Products, Market Access Services, and their Associated Fees." October 25. Retrieved on April 16, 2019 from https://www.sec.gov/spotlight/equity-market-structure-roundtables/roundtable-market-data-market-access-102518-transcript.pdf.

U.S. Securities and Exchange Commission. 2018*c*. "SEC Adopts Transaction Fee Pilot for NMS Stocks." Retrieved January 23, 2019 from https://www.sec.gov/news/press-release/2018-298.

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`.

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.

A Theory Appendix (For Online Publication)

A.1 Order Book Equilibrium (OBE)

Preliminaries. In the following definitions and proofs, we denote by $o_{ij} \in \mathcal{O}$ the *order* for TF i submitted to exchange j, where $\mathcal{O}$ is the set of all potential combinations of messages. We allow for three types of messages that TFs can send to a particular exchange j: (i) standard limit orders, which take the form (q_i, p_i) and indicate that the TF is willing to buy (if $q_i > 0$) or sell (if $q_i < 0$) up to $|q_i|$ units at price p_i; (ii) cancellations of existing limit orders in $\boldsymbol{\omega}_j$, i.e., exchange j's order book; and (iii) immediate-or-cancel orders (IOCs), which are standard limit orders that, if not fully executed in a given period, have any portion that is remaining cancelled by the exchange at the end of the period. An order submitted to a particular exchange may also contain no messages (i.e., $o_{ij} = \emptyset$); such an order simply maintains the TF's existing limit orders on that exchange, if any exist. A TF can adjust an existing limit order (e.g., change the price) by cancelling the old limit order and submitting a new one. Denote by $\boldsymbol{o}_i \equiv \{o_{ij}\}_{j \in \mathcal{M}}$ the set of orders submitted by TF i to all exchanges, where $\mathcal{M}$ represents the set of all exchanges.

We say that a *limit order provides liquidity* if it offers to buy (or sell) some positive quantity at a price less than (or greater than) y. Denote by $LO_{ij}(o_{ij}; \boldsymbol{\omega}_j)$ the set of TF i's liquidity-providing limit orders on exchange j, given the prior state of exchange j's order book $\boldsymbol{\omega}_j$ and the the processing of any messages contained in o_{ij}. We say that an *order o_{ij} provides liquidity* on exchange j if $LO_{ij}(o_{ij}; \boldsymbol{\omega}_j)$ is non-empty; this can occur if either (i) o_{ij} contains a limit order that provides liquidity; or (ii) TF i has outstanding liquidity-providing limit orders on exchange j and o_{ij} does not cancel all of these limit orders. As noted in the main text, because we have assumed that investors are equally likely to arrive needing to buy or sell one unit of the security and the distribution of jumps in y is symmetric about zero, it is convenient to focus on the provision of liquidity via combinations of two limit orders: i.e., for a given quantity l and fundamental value y, a limit order to buy the security at $y - s/2$, and a limit order to sell at $y + s/2$ for some (bid-ask) spread $s \geq 0$. We say *o_{ij} provides l units of liquidity at spread s* if $LO_{ij}(o_{ij}; \boldsymbol{\omega}_j)$ contains such a combination of limit orders.

There are two sets of relationships between orders that we refer to in this Appendix. We say that $\boldsymbol{o}_i'$ (weakly) *withdraws liquidity* relative to $\boldsymbol{o}_i$ if any limit order providing liquidity (at a given price and quantity on a particular exchange) contained in $\boldsymbol{o}_i'$ is also contained in $\boldsymbol{o}_i$, and any messages contained in $\boldsymbol{o}_i'$ but not in $\boldsymbol{o}_i$ are cancellations of existing limit orders. This implies that, for every exchange j, any limit order providing liquidity contained in $LO_{ij}(o_{ij}'; \boldsymbol{\omega}_j)$ is also contained in $LO_{ij}(o_{ij}; \boldsymbol{\omega}_j)$. We say that $\boldsymbol{o}_i'$ is a (strict) *price improvement* over $\boldsymbol{o}_i$ if, for any $q \in (0, 1]$ and any exchange $j \in \mathcal{M}$, buying and selling q units on exchange j is weakly cheaper trading against limit orders in $LO_{ij}(o_{ij}'; \boldsymbol{\omega}_j)$ than against limit orders in $LO_{ij}(o_{ij}; \boldsymbol{\omega}_j)$, and there exists some quantity $q \in (0, 1]$ and exchange j for which it is strictly cheaper to buy or sell q units trading against limit orders in $LO_{ij}(o_{ij}'; \boldsymbol{\omega}_j)$ than against limit orders in $LO_{ij}(o_{ij}; \boldsymbol{\omega}_j)$. (If there is no ability to buy (or sell) q units trading against limit orders in $LO_{ij}(\cdot)$ on exchange j, then the cost of buying (or selling) q units against limit orders in $LO_{ij}(\cdot)$ is considered infinite.) Note that if $\boldsymbol{o}_i$ does not provide liquidity on any exchange, then any $\boldsymbol{o}_i'$ providing liquidity at any (finite) price on any exchange represents a price improvement over $\boldsymbol{o}_i$.

In the following definition, we allow TFs to potentially withdraw liquidity "in response to" Period-1 deviations. In these circumstances, we assume that TFs are able to observe and condition withdrawals on the *interim state* of all exchanges' order books, defined to be the set of outstanding bids and asks and their respective queue priority on each exchange following the processing of Period-1 orders but prior to the start of Period 2.[78] When the distinction is important for the purposes of computing expected profits, we explicitly condition withdrawals on the interim state, denoted $\tilde{\boldsymbol{\omega}}$. Although queue priority does not play a role in "on-the-equilibrium-path" behavior for the equilibria constructed in Propositions 3.1 and 3.2, it is important for evaluating the potential profitability of "off-path" trading

[78] For example, consider a single continuous limit order book exchange and two TFs with the same general and exchange-specific speed technology, and say both TFs submit orders in Period 1 to provide a single unit of liquidity at the same price and bid-ask spread. The exchange's interim state reflects which order was processed first by the exchange and hence which TF's liquidity has higher priority to be filled first upon the arrival of an investor in Period 2.

game deviations and responses.

Let $E\pi_i(\boldsymbol{o}_i, \boldsymbol{o}_{-i})$ represent TF i's expected profits from a trading game given the Period-1 orders $\boldsymbol{o}_i$ that it submits and the Period-1 orders submitted by all other trading firms, denoted $\boldsymbol{o}_{-i} \equiv \{o_{kj}\}_{k \neq i, j \in \mathcal{M}}$, taking as given the state $(y, \boldsymbol{\omega})$ at the beginning of Period 1 of the given trading game, and assuming that all market participants employ their essentially unique optimal strategies in Period 2 of the trading game. Expectations are taken over the potentially random sequence in which the orders $(\boldsymbol{o}_i, \boldsymbol{o}_{-i})$ are processed by exchanges in Period 1, the random action of nature in Period 2, and, in the event of a sniping race in Period 2, the random sequence in which TFs' orders are processed by exchanges.

For thinking practically about the order book equilibrium concept, it is important to reiterate that our use of the language "submit orders" includes the possibility that a TF does not send any messages to one or more exchanges, and hence simply maintains its outstanding limit orders on those exchanges if it has any. As noted in the main text, trading games can be interpreted as lasting a sufficiently short amount of time (e.g., one millisecond or potentially even less) so that in most trading games, no exogenous Period-2 events occur. Hence, in the equilibria that we analyze, in most trading games TFs will not send any messages to any exchanges, and will simply maintain their existing limit orders in $\boldsymbol{\omega}$. It is in this sense that these equilibria capture the idea that each exchange's limit order book settles into a rest point between Period-2 events, i.e., between arrivals of an investor, informed trader, or public information.

Definition. We first provide an informal definition, and then a formal definition below (Definition A.1). Informally, an order book equilibrium is a set of orders $\boldsymbol{o}^* \equiv \{o_i^*\}$ submitted by trading firms in period 1 of each trading game that conditions only on the current state $(y, \boldsymbol{\omega})$, and satisfies two conditions. First, we require that there are no strictly profitable unilateral deviations for any TF that are *safe profitable price improvements*. We say that a price improvement is *profitable* if it is strictly profitable given the orders submitted by other TFs; and a profitable price improvement is *safe* if it remains strictly profitable even if some other TF withdraws liquidity in response to the deviation. Second, we require that there are no *robust deviations*, defined to be strictly profitable unilateral deviations that remain strictly profitable even if a rival TF withdraws liquidity or engages in a safe profitable price improvement in response. This requirement generates a constraint on the equilibrium level of spreads even without the presence of excess liquidity; however, it still allows a liquidity provider to widen its spread as long as in doing so, it doesn't induce another TF to provide liquidity at a strictly better price.

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

(i) *No safe profitable price improvements.* No TF i has a profitable price improvement that is *safe,* defined as remaining strictly profitable even if some other TF withdraws liquidity in response to TF i's deviation. Formally, for any TF i, if $\boldsymbol{o}_i'$ is a price improvement over $\boldsymbol{o}_i^*$ and is strictly profitable meaning that $E\pi_i(\boldsymbol{o}_i', \boldsymbol{o}_{-i}^*) > E\pi_i(\boldsymbol{o}_i^*, \boldsymbol{o}_{-i}^*)$, then there is some other TF $k \neq i$ and reaction $\boldsymbol{o}_k'(\cdot)$ that withdraws liquidity relative to $\boldsymbol{o}_k^*$ and renders TF i's deviation no longer strictly profitable: i.e., $E\pi_i(\boldsymbol{o}_i', (\boldsymbol{o}_k'(\tilde{\boldsymbol{\omega}}), \boldsymbol{o}_{-ik}^*)) \leq E\pi_i(\boldsymbol{o}_i^*, \boldsymbol{o}_{-i}^*)$, where $\tilde{\boldsymbol{\omega}}$ is the interim state arising from the processing of Period-1 orders $(\boldsymbol{o}_i', \boldsymbol{o}_{-i}^*)$, and $\boldsymbol{o}_{-ik}^*$ denotes orders in $\boldsymbol{o}^*$ for all TFs other than TF i and TF k. In our definition, we allow for $\boldsymbol{o}_k'(\cdot)$ to condition on the interim state $\tilde{\boldsymbol{\omega}}$.

(ii) *No robust deviations.* No TF i has any other strictly profitable deviation (i.e., not a price improvement) that is *robust*, defined as remaining strictly profitable if, in response to TF i's deviation, some other TF k either: (a) withdraws liquidity; or (b) engages in a safe profitable price improvement (as defined in (i)). Formally, for any TF i, if $E\pi_i(\boldsymbol{o}_i', \boldsymbol{o}_{-i}^*) > E\pi_i(\boldsymbol{o}_i^*, \boldsymbol{o}_{-i}^*)$ for some deviation $\boldsymbol{o}_i'$ that is not a price improvement over $\boldsymbol{o}_i^*$, then there is some TF k and reaction $\boldsymbol{o}_k'$ that renders TF i's deviation no longer strictly profitable, and either: (a) $\boldsymbol{o}_k'$ withdraws liquidity relative to $\boldsymbol{o}_k^*$ (allowing the withdrawal to condition on the interim state $\tilde{\omega}$, as in (i)); or (b) $\boldsymbol{o}_k'$ is a safe profitable price improvement, and hence is a profitable price improvement that remains strictly profitable for TF k even if any other TF, including TF i, withdraws liquidity in response.

Intuition for How OBE Restores Equilibrium Existence. Here, we provide intuition for why our order book equilibrium concept helps restore equilibrium existence in our Stage 3 trading game. Consider the single-exchange case analyzed in Section 3.2.1, where the exchange charges zero trading fees and all N TFs have purchased ESST. Consider a candidate equilibrium where a single unit of liquidity is provided by TF i at spread $s^*_{continuous}$ following period 1. Say, in this example, this implies TF i submits a bid at a price of 9 and an ask of 11 when $y = 10$ (thus, $s^*_{continuous} = 2$). Notice that TF i has a unilaterally profitable deviation of widening its spread to say $s' = 4$ (i.e., bid 8, ask 12). However, there is now a safe profitable price improvement by some other TF k that renders this deviation unprofitable: TF k could choose to provide a unit of liquidity at bid $8 + \varepsilon$, ask $12 - \varepsilon$ for sufficiently small $\varepsilon > 0$; this reaction remains profitable for k even if i were to withdraw any liquidity, and thus is safe. Alternatively, consider the unilaterally profitable deviation by TF k to undercut TF i's equilibrium order by adding a unit of liquidity at bid $9 + \varepsilon$, ask $11 - \varepsilon$: in doing so, TF k attempts to "have his cake and eat it too" (as discussed in Section 3.2.1), and earn revenues from liquidity provision at a strictly narrower spread while also sniping TF i's existing orders. However, TF i can withdraw liquidity in response to TF k's price improvement and render the deviation unprofitable (since k would prefer to snipe i's liquidity at $s^*_{continuous}$ than provide liquidity at a narrower spread); TF k's price improvement is thus not safe. Hence, these types of deviations that otherwise would have challenged the existence of an MPE no longer do so for an order book equilibrium.

A.2 Proofs for Section 3

A.2.1 Proof of Proposition 3.1 (Equilibrium of the Single-Exchange Trading Game)

Existence. We first prove that there exists an OBE of the single exchange trading game with $N \geq 2$ fast TFs, where a single unit of liquidity is provided at spread $s^*_{continuous}$. As discussed in the main text, Period 2 behavior for investors, informed traders, and (fast) trading firms is governed by essentially unique optimal strategies and described in the statement of the proposition.[79] Consider the following set of (fast) TF orders in Period 1 given state (y, ω). Some TF i submits an order such that he provides exactly one unit of liquidity at spread $s^*_{continuous}$ around y; if he has liquidity outstanding from the previous trading game in ω, he maintains, adjusts or withdraws it as necessary so that he provides exactly one unit at spread $s^*_{continuous}$ around y. All other TFs $k \neq i$ do not provide any liquidity (and withdraw any existing liquidity in ω, if present).

To show that these orders comprise an OBE, first consider deviations by TF i. Note that it is not profitable for TF i to adjust the amount of liquidity that it provides: withdrawing any amount of liquidity is unprofitable, since it earns strictly positive profits on its one unit provided at spread $s^*_{continuous}$; and offering additional liquidity beyond the initial one unit is unprofitable, as doing so only incurs additional adverse selection and sniping costs without any additional benefits. TF i also does not wish to reduce the spread on any amount of liquidity, as this strictly reduces its profits. Last, although there is a strictly profitable deviation by TF i to increase its spread to $s' > s^*_{continuous}$ for $l \leq 1$ units of liquidity that it provides, such a deviation is not robust. To see why, consider as a reaction the profitable price improvement by some TF $k \neq i$ to provide l units at spread $s^*_{continuous}$, and an additional $1 - l$ units as a *stub quote* (i.e., liquidity provided at a spread outside the support of J). This reaction renders TF i's deviation unprofitable; furthermore, the reaction is safe since k would prefer to offer such liquidity even if TF i were to withdraw any of its liquidity: providing l units of liquidity at $s^*_{continuous}$ is strictly preferable to sniping the same amount of liquidity at $s' > s^*_{continuous}$, and the stub quote ensures that k prefers to engage in its reaction even if TF i were to withdraw any of its liquidity.

Next, consider potential deviations for other TFs (who do not provide any liquidity in equilibrium):

1. No TF $k \neq i$ would wish to add any amount of liquidity at a strictly greater spread than $s^*_{continuous}$, as this incurs only adverse selection and sniping costs without any benefits of being traded against by an investor.

[79] Optimal strategies only differ in prescribed behavior for market participants when they are indifferent over consuming liquidity priced exactly at y.

2. Consider any strictly profitable deviation by TF $k \neq i$ that involves "undercutting" TF i by offering $l \leq 1$ additional units of liquidity at spread $s' = s^*_{continuous} - \varepsilon$; this deviation is strictly profitable for sufficiently small $\varepsilon > 0$ since TF k earns revenues from both liquidity provision (earning priority over i at a cost of just ε) and from sniping TF i's liquidity. But this deviation does not remain strictly profitable if TF i withdraws l units of its own liquidity offered at spread $s^*_{continuous}$: by (3.1), liquidity provision and stale quote sniping are equally profitable at $s^*_{continuous}$; hence TF k would have preferred to snipe at $s^*_{continuous}$ than provide liquidity at a strictly narrower spread, $s' < s^*_{continuous}$.

3. Consider the deviation by TF $k \neq i$ to provide l additional units of liquidity at $s^*_{continuous}$. Due to the random sequence in which messages are processed by the exchange, TF k's liquidity will be filled by an investor only if it is added to the order book before TF i's liquidity; since this occurs with positive probability and since liquidity provision at $s^*_{continuous}$ is strictly profitable when only one unit of liquidity is provided (by (3.1), it earns the same profits in expectation as sniping stale quotes at $s^*_{continuous}$), this deviation is strictly profitable for sufficiently small $l > 0$. Consider then the reaction by TF i to withdraw l units of its liquidity at $s^*_{continuous}$ only if it has lower queue priority than TF k (recall that withdrawals following a price improvement can condition on the interim state realized after the processing of Period-1 orders; cf. Section A.1). This reaction renders the deviation by TF k not strictly profitable: when k has higher queue priority, only 1 unit of depth is offered and k is indifferent between liquidity provision and sniping at $s^*_{continuous}$; and when k has worse queue priority, it only bears the sniping and adverse selection costs without the benefits of being filled by an investor upon arrival.

Hence, there are no robust deviations or safe profitable price improvements for any TF. Thus, these orders comprise an OBE of the single exchange trading game.

Uniqueness. As discussed in the main text, Period 2 behavior for investors, informed traders and (fast) trading firms is governed by essentially unique optimal strategies and described in the statement of the proposition. We next show that in *any* OBE, exactly a single unit of liquidity is provided at spread $s^*_{continuous}$ at the end of Period 1. (All references to liquidity provision are with respect to liquidity provided at a spread within the support of J; any liquidity offered outside this support has no economic role in equilibrium.)First, consider a candidate equilibrium where there are $l > 1$ units of liquidity offered at the end of Period 1. Consider any amount of liquidity offered at the worst price. If such liquidity would never be filled by an investor in Period 2—which can occur if there is at least one unit of liquidity offered at a strictly better price—then any TF offering such liquidity would have a strictly profitable deviation to withdraw this liquidity (which remains profitable even if others could respond with a price improvement), as such liquidity only bears adverse selection and sniping costs without liquidity provision benefits; thus, this cannot be an OBE. Hence, if there is greater than one unit of liquidity offered in total, all liquidity offered at the worst price must be in expectation filled by an investor in Period 2 with some probability that is strictly positive, but less than 1 (as $l > 1$). However, in this case, any TF offering liquidity at the worst price has a profitable price improvement to reduce the spread on its liquidity by some small amount $\varepsilon > 0$, thereby ensuring that its liquidity would be filled by an investor with certainty in Period 2; furthermore, this deviation remains profitable even if other TFs withdrew liquidity, and hence is safe. Thus there cannot be $l > 1$ units of liquidity offered at the end of Period 1 in any OBE. Next, consider a candidate equilibrium where there are $l < 1$ units of liquidity offered at the end of Period 1. Consider the strictly profitable unilateral deviation by any TF to offer $1 - l$ additional units of liquidity at spread $s^*_{continuous}$. This is a safe profitable price improvement, as reactions that withdraw offered liquidity do not render this deviation weakly unprofitable. This cannot be an OBE; contradiction. Thus, exactly a single unit of liquidity must be offered at the end of Period 1 in any OBE.

Now consider a candidate equilibrium where exactly one unit of liquidity is offered at the end of Period 1, but $l \leq 1$ units are offered at a spread $s < s^*_{continuous}$ by some TF i. Consider the strictly profitable unilateral deviation by TF i to increase its spread to $s^*_{continuous}$ on its offered liquidity. As above, there is no safe profitable price improvement that renders the deviation weakly unprofitable, as any TF considering a price improvement that undercuts TF i

would instead prefer to snipe TF i's liquidity at $s^*_{continuous}$ as opposed to providing liquidity at a narrower spread. This cannot be an OBE; contradiction. Next, consider a candidate equilibrium where exactly one unit of liquidity is offered at the end of Period 1, but $l \leq 1$ units are offered at a spread $s > s^*_{continuous}$ by some TF i . There is a safe profitable price improvement by TF $k \neq i$ to undercut and provide l units at spread $s^*_{continuous}$, as there are no withdrawals of liquidity that render the deviation weakly unprofitable (since k prefers to provide liquidity at $s^*_{continuous}$ to sniping liquidity provided at $s > s^*_{continuous}$). This cannot be an OBE; contradiction.

Thus, any OBE has a single unit of liquidity provided at bid-ask spread $s^*_{continuous}$ following Period 1.

A.2.2 Supporting Lemmas for Proposition 3.2

The proof of Proposition 3.2 relies on the following two supporting lemmas.

Lemma A.1. *Consider any Stage 3 trading game where all N TFs have purchased ESST from the same set of exchanges. Further assume that trading fees are zero for all exchanges contained in the non-empty set $\mathcal{J} \subseteq \mathcal{M}$, and strictly positive for all exchanges $m \notin \mathcal{J}$. Then:*

1. Existence: for any vector of market shares $\boldsymbol{\sigma^*} = (\sigma^*_1, \ldots, \sigma^*_M)$ such that $\sum_{j\in\mathcal{J}} \sigma^*_j = 1$ and $\sigma^*_m = 0$ if $m \notin \mathcal{J}$, there exists an OBE in which TFs in aggregate provide σ^*_j units of liquidity on each exchange j at spread $s^*_{continuous}$ in Period 1.
2. Uniqueness: any OBE of the trading game has exactly one unit of liquidity provided in aggregate at spread $s^*_{continuous}$ in period 1, where liquidity is provided across exchanges according to some vector of market shares $\boldsymbol{\sigma^*} = (\sigma^*_1, \ldots, \sigma^*_M)$ such that $\sum_{j\in\mathcal{J}} \sigma^*_j = 1$ and $\sigma^*_m = 0$ if $m \notin \mathcal{J}$.

(We do not require the uniqueness portion of Lemma A.1 for our main results, but state and prove it here for completeness.)

Proof. Condition on state (y, ω) at the beginning of this trading game.

Existence. Consider any vector of exchange market shares $\boldsymbol{\sigma^*} = (\sigma^*_1, \ldots, \sigma^*_M)$ such that $\sum_{j\in\mathcal{J}} \sigma^*_j = 1$ and $\sigma^*_m = 0$ if $m \notin \mathcal{J}$. Consider the following candidate strategies. In Period 1, a single TF i submits an order to each exchange $j \in \mathcal{J}$ to provide exactly σ^*_j units of liquidity at spread $s^*_{continuous}$ around y (maintaining, adjusting or withdrawing any outstanding liquidity from the previous trading game as necessary). All other TFs $k \neq i$ do not provide any liquidity (and withdraw any existing liquidity, if present). In Period 2: investors trade at least one unit, prioritizing across exchanges based on the lowest value of $s_j/2 + f_j$ (where for each exchange j, s_j is the lowest spread at which liquidity is offered and f_j is the trading fee), breaking ties according to routing table strategies given by $\boldsymbol{\gamma}^* = \boldsymbol{\sigma}^*$, and then trading against any remaining profitable orders; informed traders trade against any profitable orders; and if there is a publicly observable jump in y, TF i sends messages to cancel all liquidity providing orders, and all other TFs attempt to trade against any profitable orders. As discussed in the main text, Period-2 strategies are essentially unique, and there are no strictly profitable Period-2 deviations. Consider now strictly profitable Period-1 deviations. The arguments here are analogous to those used above in the proof of Proposition 3.1. If TF i increases its spread on any exchange, there is a safe profitable price improvement by some TF $k \neq i$ to provide liquidity at spread $s^*_{continuous}$, rendering the deviation not strictly profitable. If some TF $k \neq i$ adds additional liquidity on any exchange, TF i can withdraw any amount of liquidity whenever profitable to do so (i.e., any liquidity with low enough queue priority to not be filled by an investor in Period 2), rendering the deviation not strictly profitable. Thus, these strategies comprise an OBE. Note that in this equilibrium, in each trading game each TF earns (gross ESST fees) expected profits of $\sigma^*_j \times \Pi^*_{continuous}/N$ on exchange j from either liquidity provision or sniping activity; this implies that each TF earns in aggregate $\Pi^*_{continuous}/N$ per-trading game across all exchanges—the same amount that each TF would earn in equilibrium if there was only a single exchange.

Uniqueness. Consider any OBE where $\boldsymbol{l} = (l^*_1, \ldots, l^*_M)$ units of liquidity are provided across exchanges at the end of Period 1, investors use routing table strategies given by $\boldsymbol{\gamma}^* = (\gamma^*_1, \ldots, \gamma^*_M)$, and market shares are given

by $\boldsymbol{\sigma}^* = (\sigma_1^*, \ldots, \sigma_M^*)$, where σ_j volume is transacted on each exchange j in the event that investors arrive. In any equilibrium, we show that exactly a single unit of liquidity is provided in aggregate among all exchanges with zero trading fees (i.e., $\sum_{j\in\mathcal{J}} l_j^* = 1$ and $l_m^* = 0$ if $m \notin \mathcal{J}$) at spread $s^*_{continuous}$ around y following Period 1; and transaction volume upon the arrival of an investor coincides with liquidity provision for all exchanges (i.e., $\sigma_j^* = l_j^*$ for all $j \in \mathcal{M}$). We prove this by ruling out the following cases. First, any amount of liquidity $l > 0$ cannot be provided at spread $s' \neq s^*_{continuous}$ on some exchange $j \in \mathcal{J}$ by any TF i, as the same arguments used in the proof of Proposition 3.1 establish that there would then exist either a safe profitable price improvement or a robust deviation for some other TF. Second, in Period 1, exactly one unit of liquidity must be provided: if less than one unit is provided, then any TF would have a safe profitable price improvement to add some small amount of liquidity at spread $s^*_{continuous}$ to some exchange $j \in \mathcal{J}$; if more than one unit is provided, then the same arguments used in the proof of Proposition 3.1 establish that some TF offering liquidity at the worst price has a robust deviation to either withdraw such liquidity, or reduce the spread by some positive amount $\varepsilon > 0$ to guarantee that it would be transacted against by an investor in Period 2. Third, positive liquidity cannot be provided on any exchange m where $f_m > 0$. Assume not, and some amount of liquidity $l > 0$ is provided on exchange m at spread s' by some TF i in equilibrium. For this to be an equilibrium, it cannot be that $s'/2 + f_m > s^*_{continuous}/2$, as otherwise there would be a safe profitable price improvement by another TF k to provide the same amount of liquidity l on any exchange $j \in \mathcal{J}$ at spread $s^*_{continuous}$ (since both an investor would strictly prefer to transact on exchange j at spread $s^*_{continuous}$ than on exchange m at s', and TF k would strictly prefer to provide liquidity on exchange j at $s^*_{continuous}$ than snipe liquidity on exchange m at s'). However, if $s'/2 + f_m \leq s^*_{continuous}/2$, then TF i has a robust deviation to withdraw all liquidity on m and offer the same amount of liquidity on any exchange $j \in \mathcal{J}$ at spread $s^*_{continuous}$ (and avoid trading fees) since there is no safe profitable price improvement by any other TF (e.g., any TF $k \neq i$ prefers sniping liquidity on j at $s^*_{continuous}$ than offering it on any other exchange at a lower spread). Thus, any equilibrium involves exactly a single unit of liquidity provided in aggregate at spread $s^*_{continuous}$, and liquidity is only provided on exchanges with zero trading fees.

Lemma A.2. *("Lone-Wolf Lemma") Consider any Stage 3 trading game where: (i) trading fees on all exchanges are zero; (ii) TF i, referred to as the "lone-wolf," has purchased exchange-specific speed technology (ESST) only on exchanges contained in the set $\mathcal{J} \subset \mathcal{M}$; and (iii) all other TFs $k \neq i$ have purchased ESST on all exchanges. There exists an OBE of this trading game where exactly one unit of liquidity is provided only on exchanges contained in $\mathcal{J}$ by TF i at spread $\tilde{s}_N$ in Period 1, where $\tilde{s}_N$ solves:*

$$\lambda_{invest}\frac{\tilde{s}_N}{2} - (\frac{N-2}{N-1}\lambda_{public} + \lambda_{private})L(\tilde{s}_N) = \frac{\lambda_{public}L(\tilde{s}_N)}{N}\,, \tag{A.1}$$

and TF i earns in expectation at least $\pi_N^{lone-wolf} = \frac{N-1}{N^2}\lambda_{public}L(\tilde{s}_N)$ per-trading game gross of ESST fees, where $\pi_N^{lone-wolf} \in (\frac{N-2}{N-1} \times \frac{\Pi^*_{continuous}}{N}, \frac{\Pi^*_{continuous}}{N})$. Furthermore, in any OBE in which only TF i provides liquidity in Period 1 of each trading game, a single unit is provided only on exchanges contained in $\mathcal{J}$ by TF i at spread $\tilde{s}_N$.

Proof. In this proof, all references to TF profits are in expectation for each trading game, gross ESST fees.

Preliminaries. Define the spread $\bar{s}_N$ to be the minimum spread TF i must charge on exchange j for one-unit of liquidity so that i breaks even in expectation when $N-1$ other trading firms have also purchased ESST from j and no liquidity is provided on any other exchange; i.e., $\bar{s}_N$ is the solution to:

$$\lambda_{invest}\frac{\bar{s}_N}{2} - (\frac{N-1}{N}\lambda_{public} + \lambda_{private})L(\bar{s}_N) = 0\,, \tag{A.2}$$

and we refer to $\bar{s}_N$ as the zero-variable profit spread. The difference between the definition of $\bar{s}_N$ and the definition of $s^*_{continuous}$ in (3.1) is that $\bar{s}_N$ does not incorporate the opportunity cost of sniping, worth $\frac{1}{N}\lambda_{public}L(\cdot)$.

Next, we provide intuition for the spread $\tilde{s}_N$ defined in (A.1). Assume that conditions (i)-(iii) in the statement of the Lemma hold, and the lone-wolf TF i provides one unit of liquidity at spread $\tilde{s}_N$ on some exchange $j \in \mathcal{J}$. Then any other TFs $k \neq i$ would be indifferent between (i) sniping TF i on exchange j (earning the right-hand side

of (A.1)), and (ii) TF i not providing any liquidity, and TF k instead providing one unit of liquidity at $\tilde{s}_N$ on some exchange $j' \notin \mathcal{J}$ (earning the left-hand side of (A.1), where TF k only risks being sniped by $N-2$ other TFs who have ESST on exchange j').

We now prove that $\bar{s}_N < \tilde{s}_N < s^*_{continuous}$. The first inequality, $\bar{s}_N < \tilde{s}_N$, follows from comparing (A.2) to (A.1), which can be re-written as $\lambda_{invest}\frac{\tilde{s}_N}{2} - ((\frac{1}{N} + \frac{N-2}{N-1})\lambda_{public} + \lambda_{private})L(\tilde{s}_N) = 0$. It is straightforward to show that the coefficient on λ_{public} is greater in (A.1) than in (A.2): $\frac{1}{N} + \frac{N-2}{N-1} = 1 - \frac{1}{N(N-1)} > 1 - \frac{1}{N} = \frac{N-1}{N}$. Hence, it follows that $\tilde{s}_N > \bar{s}_N$. The second inequality, $\tilde{s}_N < s^*_{continuous}$, follows using similar logic: in (3.1), which defines $s^*_{continuous}$, λ_{public} enters the equation with a coefficient of 1; however, in (A.1), which defines $\tilde{s}_N$, λ_{public} enters with a coefficient strictly less than 1.

The rest of the proof proceeds in three parts. First, we establish that an equilibrium with the properties outlined in the statement of the Lemma exists (Existence). Second, we establish that any equilibrium in which only TF i provides liquidity in Period 1 of each trading game must have these properties (Uniqueness). Last, we prove that $\pi_N^{lone-wolf} \in (\Pi^*_{continuous} \times (N-2)/((N-1) \times N), \Pi^*_{continuous}/N)$ (Profit Bound).

Existence. We prove that there is an OBE of the Stage 3 trading game in which TF i provides one unit of liquidity at spread $\tilde{s}_N$ across exchanges according to any arbitrary vector of shares $\boldsymbol{\sigma^*} = (\sigma_1^*, \ldots, \sigma_M^*)$ s.t. $\sum_{j \in \mathcal{J}} \sigma_j^* = 1$ and $\sigma_j^* = 0$ if $j \notin \mathcal{J}$, and no additional liquidity is provided by any other TF. Consider equilibrium strategies where in Period 1, TF i submits orders to provide one unit of liquidity at spread $\tilde{s}_N$ across exchanges in $\mathcal{J}$ according to $\boldsymbol{\sigma^*}$ (maintaining, adjusting or withdrawing any outstanding liquidity from the previous trading game as necessary), and other TFs do not provide any liquidity (and withdraw any existing liquidity, if ever present); and in Period 2, strategies follow those described in the proof of Lemma A.1 (where investors break ties across exchanges using routing table strategies $\boldsymbol{\gamma}^* = \boldsymbol{\sigma}^*$). The right-hand-side of (A.1) represents the gross expected payoffs that any TF $k \neq i$ expects to obtain by sniping TF i across all exchanges; the left-hand-side represents the gross expected payoffs that any TF k would anticipate if TF k were instead the sole liquidity provider on some other exchange $m \notin \mathcal{J}$ at spread $\tilde{s}_N$. Hence, no TF $k \neq i$ has a strictly profitable deviation—for example, by undercutting i or providing additional liquidity at a spread weakly smaller than $\tilde{s}_N$ on any exchange—that remains profitable if TF i reacts by withdrawing any liquidity that is no longer profitable to offer. By similar arguments used to establish our single exchange results, there is also no robust deviation for TF i: if TF i widened its spread on any amount of liquidity, the deviation would be rendered unprofitable by another TF k's safe reaction to provide that amount of liquidity on some exchange $m \notin \mathcal{J}$ at spread $\tilde{s}_N$ (which, by (A.1), is more profitable for TF k than sniping TF i at any spread strictly greater than $\tilde{s}_N$); and TF i reducing its spread or adjusting the amount of liquidity that it provides would strictly reduce profits. Thus, these strategies comprise an OBE.

Uniqueness. We prove that in any OBE in which only TF i provides liquidity in Period 1 of each trading game, a single unit of liquidity is provided by TF i across only exchanges contained in $\mathcal{J}$ at spread $\tilde{s}_N$. First, note that TF i must offer exactly one unit of liquidity in aggregate: otherwise, TF i would find it profitable to withdraw liquidity (if it offered strictly greater than one unit of liquidity) or have a safe profitable price improvement to add liquidity at spread $\tilde{s}_N$ on some exchange in $\mathcal{J}$ (if it offered strictly less than one unit of liquidity). Second, note that such liquidity must be offered only on exchanges in $\mathcal{J}$. Assume not, and some positive amount of liquidity $l_m > 0$ was offered by TF i on some exchange $m \notin \mathcal{J}$. Any such liquidity on exchange m must be offered by TF i at a spread $s^*_{continuous}$: if it were offered at a greater spread, there would be a safe profitable price improvement by some other TF $k \neq i$ to undercut TF i and offer this amount of liquidity on exchange m at spread $s^*_{continuous}$; if it were offered at a lower spread, then TF i would find it profitable to withdraw such liquidity, since without ESST on exchange m, a TF who provides liquidity at $s^*_{continuous}$ earns zero expected profits—i.e., the revenue from investor arrivals is exactly offset by the costs of adverse selection and sniping. However, if TF i offered positive liquidity on exchange m at spread $s^*_{continuous}$, it would then have a robust deviation to withdraw that liquidity and offer instead the same amount of liquidity on some exchange in $\mathcal{J}$ at spread $\tilde{s}_N$ (as discussed above when establishing Existence, no TF $k \neq i$ would find offering liquidity at any spread less than $\tilde{s}_N$ on any exchange strictly preferable to sniping TF i's liquidity on an exchange in $\mathcal{J}$ at spread $\tilde{s}_N$). Hence, TF i must offer a single unit of liquidity only across exchanges

in $\mathcal{J}$. Last, TF i must offer a single unit of liquidity at spread $\tilde{s}_N$. If any amount of liquidity were offered at a lower spread, TF i would have a robust deviation to increase its spread to $\tilde{s}_N$; and if any amount of liquidity were offered at a strictly greater spread, there would be a safe profitable price improvement by some TF $k \neq i$ to provide the same amount of liquidity on some exchange $m \notin \mathcal{J}$ at spread $\tilde{s}_N$.

Profit Bound. Define

$$\pi_N^{lone-wolf} \equiv \lambda_{invest}\frac{\tilde{s}_N}{2} - (\frac{N-1}{N}\lambda_{public} + \lambda_{private})L(\tilde{s}_N) \tag{A.3}$$

to be the expected profits per trading game (gross ESST fees) that a lone-wolf liquidity provider (TF i) makes providing a single unit of liquidity at spread $\tilde{s}_N$ across exchanges contained in $\mathcal{J}$ when there are N total trading firms (including him) that also have purchased ESST on exchanges contained in $\mathcal{J}$, and TF i is the sole liquidity provider. We now prove bounds on $\pi_N^{lone-wolf}$. First, the upper-bound $\pi_N^{lone-wolf} < \Pi^*_{continuous}/N = \lambda_{invest}\frac{s^*_{continuous}}{2} - (\frac{N-1}{N}\lambda_{public} + \lambda_{private})L(s^*_{continuous})$ follows since $\tilde{s}_N < s^*_{continuous}$ and $L(\tilde{s}_N) > L(s^*_{continuous})$. Next, consider the lower-bound $\pi_N^{lone-wolf} > \frac{N-2}{(N-1)N}\Pi^*_{continuous}$. Solving for $\lambda_{invest}\frac{\tilde{s}_N}{2}$ in (A.1) and substituting this expression into the right-hand side of (A.3) yields:

$$\begin{aligned}\pi_N^{lone-wolf} &= (\frac{1}{N} + \frac{N-2}{N-1} - \frac{N-1}{N})\lambda_{public}L(\tilde{s}_N) \\ &= \frac{N-2}{(N-1)N}\lambda_{public}L(\tilde{s}_N)\ .\end{aligned}$$

The bound then follows from $\lambda_{public}L(\tilde{s}_N) > \lambda_{public}L(s^*_{continuous}) = \Pi^*_{continuous}$.

A.2.3 Proof of Proposition 3.2 (Equilibrium Existence for the Multiple-Exchange Game)

For any vector of ESST fees $\boldsymbol{F}^*$ that satisfies (3.2) and market shares $\boldsymbol{\sigma}^*$ such that $\sum_{j\in\mathcal{M}}\sigma_j^* = 1$, consider the following candidate equilibrium strategies:

- In Stage 1, each exchange j charges F_j^* for ESST and sets trading fees $f_j = 0$;
- In Stage 2, all TFs buy ESST from exchange j only if (i) its ESST fee $F_j \leq F_j^*$, (ii) $f_j = \min_{k\in M} f_k$, and (iii) $\sigma_j^* > 0$;
- In Stage 3,
    1. If all TFs purchase ESST from the same set of exchanges $\mathcal{J} \subseteq \mathcal{M}$ where $f_j = 0$ for all $j \in \mathcal{J}$, then in Period 1 of each trading game, some TF i submits orders to provide $\sigma_j^*/(\sum_{k\in\mathcal{J}}\sigma_k^*)$ amount of liquidity on each exchange $j \in \mathcal{J}$ at spread $s^*_{continuous}$ around y (maintaining, adjusting or withdrawing any outstanding liquidity from the previous trading game as necessary), all other TFs submit orders such that they provide no liquidity on any exchange, and no liquidity is provided elsewhere (this occurs on the candidate equilibrium path).
    2. If one TF i purchases ESST from a non-empty strict subset of exchanges $\mathcal{J}' \subset \mathcal{M}$, and all other TFs $k \neq i$ purchase ESST from a strictly greater set of exchanges $\mathcal{J}$ (so that $\mathcal{J}' \subset \mathcal{J} \subseteq \mathcal{M}$) where $f_j = 0$ for all $j \in \mathcal{J}$, then in Period 1 of each trading game, TF i is the "lone-wolf" liquidity provider and submits orders to provide one unit of liquidity on some exchange $j \in \mathcal{J}'$ at spread $\tilde{s}_N$ (defined in (A.1)) around y (maintaining, adjusting or withdrawing any outstanding liquidity from the previous trading game as necessary), all other TFs submit orders such that they provide no liquidity on any exchange, and no liquidity is provided elsewhere.
    3. If one TF i purchases ESST from a set of exchanges $\mathcal{K} \subseteq \mathcal{M}$ and all other TFs $k \neq i$ purchase ESST from a strict subset of exchanges $\mathcal{J} \subset \mathcal{M}$, where $f_j = 0$ for all $j \in \mathcal{J}$ and $\mathcal{K} \nsubseteq \mathcal{J}$, then in Period 1 of each trading game:
        (a) If $\mathcal{J} \subset \mathcal{K}$ (so that TF i purchases from a strictly greater set of exchanges than all other TFs), strategies are as in Case 1. above and liquidity is provided only on exchanges in $\mathcal{J}$;

(b) If $\mathcal{J} \cap \mathcal{K} = \emptyset$ (so that TF i purchases from no exchanges contained in $\mathcal{J}$), strategies are analogous to Case 1. above: some TF $k \neq i$ which has purchased ESST on exchanges in $\mathcal{J}$ submits orders to provide $\sigma_j^*/(\sum_{k\in\mathcal{J}} \sigma_k^*)$ amount of liquidity on each exchange $j \in \mathcal{J}$ at spread $s^*_{continuous}$ around y (maintaining, adjusting or withdrawing any outstanding liquidity from the previous trading game as necessary) and all other TFs submit orders such that they provide no liquidity on any exchange;

(c) Otherwise (which occurs if $\mathcal{K}$ contains a non-empty strict subset of exchanges in $\mathcal{J}$ and at least one exchange outside of $\mathcal{J}$), strategies are as in Case 2. above where TF i is the lone-wolf liquidity provider, and provides one unit of liquidity at spread $\tilde{s}_N$ on some exchange contained in $\mathcal{J}' = \mathcal{J} \cap \mathcal{K}$ (maintaining, adjusting or withdrawing any outstanding liquidity from the previous trading game as necessary), all other TFs submit orders such that they provide no liquidity on any exchange, and no liquidity is provided elsewhere.

- In Stage 3, in Period 2 of each trading game, investors upon arrival trade one unit, prioritizing based on spreads and trading fees and breaking ties according to routing table strategies $\boldsymbol{\gamma}^* = \boldsymbol{\sigma}^*$ (with uniform tie-breaking across all exchanges where $\gamma_k = 0$), and trade against any remaining profitable orders given y; informed traders upon arrival trade against any profitable orders given y; and upon the arrival of a publicly observed jump in y, the sole liquidity providing TF attempts to cancel its liquidity providing orders while all other TFs engage in stale-quote sniping and attempt to trade against any profitable orders.

Note that these strategies dictate play in all subgames that are reachable via any sequence of unilateral deviations in Stages 1 and 2.

To show that these strategies comprise an equilibrium, we now consider potential sequences of unilateral deviations by all market participants. First, consider Stage 1 deviations involving exchanges and their choice of ESST fees and trading fees. Given equilibrium strategies, any exchange j would strictly reduce profits by lowering its ESST fee (as outcomes would otherwise remain the same); exchanges also would earn negative profits by reducing trading fees (as it would create a "money pump," as discussed in the main text). Furthermore, if any exchange increased its ESST fee, the exchange would earn zero profits as no TF would purchase ESST from it, and liquidity would only be provided on other exchanges (which is an OBE outcome of the multi-exchange trading game; see Lemma A.1); and if any exchange increased is trading fee, it also would earn zero profits for the same reasons. Hence, exchanges have no strictly profitable unilateral deviations.

Next, we turn to Stage 2 strategies for TFs. By following candidate strategies in Stage 2 given exchanges did not deviate in Stage 1, all TFs earn $\frac{1}{N}\Pi^*_{continuous} - \sum_j F_j^*$ which, by condition (3.2), is positive. Potentially profitable unilateral deviations for any TF involve the purchase of ESST from a strict subset of exchanges (as purchasing ESST from no exchanges yields no profits, and being the only TF to purchase ESST from an exchange yields no benefit due to our fair-access assumption). In subgames following such deviations, prescribed strategies comprise the unique OBE of the subsequent "lone-wolf" Stage 3 trading game given no liquidity is provided by other TFs (Lemma A.2), and the deviating TF earns in expectation $\pi_N^{lone-wolf}$ per trading game (gross trading fees). Condition (3.2) ensures that this deviation is not profitable for any TF. Similar arguments establish that there are no strictly profitable deviations for TFs in Stage 2 given at most one exchange engaged in any deviation in Stage 1.

Finally, given equilibrium play in Stages 1 and 2, Lemma A.1 establishes that Stage 3 strategies comprise an OBE.

A.2.4 Proof of Proposition 3.3

We first prove that condition (3.2) must hold in any equilibrium where all TFs purchase ESST from all exchanges and trading fees are zero for all exchanges. Consider any equilibrium in which all TFs purchase ESST from all exchanges and trading fees are zero for all exchanges. Assume by contradiction that condition (3.2) does not hold. Then one of the two following cases must be true.

(i) $\frac{\Pi^*_{continuous}}{N} < \sum_{k:\sigma^*_k>0} F_k$. This implies that TFs earn negative expected profits from equilibrium strategies; as a result, any TF has a profitable deviation of not purchasing speed technology from any exchange. Contradiction.

(ii) $\pi_N^{lone-wolf} - \min_j F_j > \frac{\Pi^*_{continuous}}{N} - \sum_{k:\sigma^*_k>0} F_k$. Consider a deviation by TF i to purchase ESST only from an exchange with the lowest ESST fee. By Lemma A.2, such a deviation earns TF i at least expected profits $\pi_N^{lone-wolf} - \min_j F_j$, which is higher than what it would earn via equilibrium strategies. Contradiction.

We next establish that if condition (3.2) is satisfied but not binding in some equilibrium where all TFs purchase ESST from all exchanges, then some exchange j can increase its ESST fee to $F'_j = F^*_j + \varepsilon$ for sufficiently small $\varepsilon > 0$, and there would still exist a subgame equilibrium beginning in Stage 2 where all TFs still purchase ESST from all exchanges (hence proving the second part of the Proposition). This follows because, for sufficiently small $\varepsilon > 0$, condition (3.2) would still hold for the vector of ESST fees $\boldsymbol{F}' = (F'_j, \boldsymbol{F}^*_{-j})$.

A.2.5 Proof of Proposition 3.4

Consider any vector of ESST fees $\boldsymbol{F}' = (F'_{1,\ldots,}F'_M)$ that maximizes $\sum_{j\in\mathcal{M}} F_j$ among all vectors of ESST fees that satisfy condition (3.2), which can be rewritten as: $\sum_{j:\sigma^*_j>0} F^*_j \leq \frac{\Pi^*_{continuous}}{N} - \max(0, \pi_N^{lone-wolf} - \min_j F^*_j)$. Since the upper bound on the total sum of ESST fees across exchanges (the left-hand side) is increasing in the minimum ESST fee (on the right-hand side), ESST fees must be equal across exchanges, and there must be a constant $\tilde{F}$ such that $F'_j = \tilde{F}$ for all $j \in \mathcal{M}$. Hence, any vector of ESST fees that maximizes the sum over all ESST fees and satisfies condition (3.2) is unique and involves each exchange charging the same amount $\tilde{F}$. This implies that each trading firm pays at most $M \times \tilde{F} \leq \frac{1}{N}\Pi^*_{continuous} - (\pi_N^{lone-wolf} - \tilde{F})$ in ESST fees across all exchanges if condition (3.2) holds. Since $\pi_N^{lone-wolf} > \frac{N-2}{(N-1)N}\Pi^*_{continuous}$ by Lemma A.2, it follows that $\tilde{F} < \frac{1}{(M-1)(N-1)N}\Pi^*_{continuous}$ and $M \times N \times \tilde{F} < \frac{M}{(M-1)(N-1)}\Pi^*_{continuous}$, where $M \times N \times \tilde{F}$ is an upper bound on the total amount of ESST earned by all exchanges if condition (3.2) holds.

A.3 Proofs For Section 5

Preliminaries: Equilibrium Spreads on Discrete. Denote by $\bar{s}_{discrete}(f)$ the zero-variable profit spread for a liquidity provider on Discrete given Discrete charges a trading fee $f \geq 0$; such a spread solves:

$$\lambda_{invest}(\frac{\bar{s}_{discrete}(f)}{2} - f) - \lambda_{private}L(\bar{s}_{discrete}(f), f) = 0\ , \tag{A.4}$$

where $L(s,f) \equiv E(J - \frac{s}{2} + f | J > \frac{s}{2} + f)Pr(J > \frac{s}{2} + f)$ represents the expected loss to a liquidity provider providing liquidity at spread s on an exchange with trading fee f in the event of being adversely traded against. The first term on the left-hand-side of (A.4) represents the revenues a liquidity provider earns when an investor arrives (i.e., half the spread less the trading fee), and the second term is the expected loss from informed trading. A unique solution $\bar{s}_{discrete}(f)$ exists for any $f \geq 0$ (and is strictly positive) by the same arguments used to establish the existence and uniqueness of $s^*_{continuous}$ in the main text (see the discussion following equation (3.1)). Let $\bar{s}_{continuous}$ denote the zero-variable profit spread $\bar{s}_N$ on Continuous, defined in equation (A.2).

Define $f^*_{discrete}$ to be the trading fee so that an investor is indifferent between trading on Discrete at spread $\bar{s}_{discrete}(f^*_{discrete})$ with trading fee $f^*_{discrete}$, and trading on Continuous at the zero-variable profit spread $\bar{s}_{continuous}$ with no trading fee. As the following lemma establishes, under a technical assumption, $f^*_{discrete}$ exists and is unique.

Lemma A.3. *Assume that the jump size distribution is continuously differentiable. Then there exists a unique solution $f^*_{discrete}$ to:*

$$\frac{\bar{s}_{discrete}(f^*_{discrete})}{2} + f^*_{discrete} = \frac{\bar{s}_{continuous}}{2}\ . \tag{A.5}$$

Furthermore, if $f <(>) f^*_{discrete}$, then $\frac{\bar{s}_{discrete}(f)}{2} + f <(>) \frac{\bar{s}_{continuous}}{2}$.

Proof. Let $H(s,f) = \lambda_{invest}(\frac{s}{2} - f) - \lambda_{private}L(s,f)$. Define $s(f)$ to be the solution to $H(s(f),f) = 0$ (hence, $\bar{s}_{discrete}(f) = s(f)$). Since the jump size distribution is continuously differentiable, $H(\cdot)$ is as well, and by the im-

plicit function theorem, the function $s(f)$ exists and is continuously differentiable with $s'(f) = -(\partial H/\partial f)/(\partial H/\partial s) = \frac{\lambda_{invest}+\lambda_{private}L_f(s,f)}{\lambda_{invest}/2-\lambda_{private}L_s(s,f)}$, where $L_f(\cdot)$ and $L_s(\cdot)$ represent partial derivatives of $L(\cdot)$. We next establish that $\frac{\bar{s}_{discrete}(f)}{2}+f$ is strictly increasing in f: differentiating this expression with respect to f implies that a sufficient condition for it to be strictly increasing in f is $s'(f) > -2$. Substituting in for $s'(f)$ and re-arranging terms yields $s'(f) > -2 \Leftrightarrow \lambda_{invest}/\lambda_{private} > (L_s(s,f) - L_f(s,f)/2)$. This inequality always holds since the left-hand-side is strictly positive, and the right-hand-side is weakly negative.[80] Since $\frac{\bar{s}_{discrete}(f)}{2} + f$ is thus strictly increasing and continuous in f, and since it is less than $\bar{s}_{continuous}/2$ for $f = 0$ but greater than $\bar{s}_{continuous}/2$ when $f = \bar{s}_{continuous}/2$, there exists a unique solution to (A.5). The rest of the statement directly follows. □

We maintain the assumption that the jump size distribution is continuously differentiable for the rest of this section.

A.3.1 Proof of Proposition 5.1

The statement of Proposition 5.1 assumes that both Continuous and Discrete charge zero trading fees. Here, we prove a more general version of Proposition 5.1, and allow Discrete to charge a weakly positive trading fee. Formally, we prove that in any equilibrium of any Stage 3 trading game given state $(y, \boldsymbol{\omega})$ with a single Continuous and single Discrete exchange where all TFs have purchased ESST from Continuous and trading fees are zero on Continuous and equal to $\tilde{f} \in [0, f^*_{discrete})$ on Discrete, exactly one unit of liquidity is provided on Discrete at bid-ask spread $\bar{s}_{discrete}(\tilde{f})$ around y following Period 1, and no liquidity is provided elsewhere. Such an equilibrium exists. (It is straightforward to use the same arguments below to establish that the Proposition statement holds even if there exist other Discrete exchanges with trading fees that are strictly greater than $\tilde{f}$).

Existence. Consider the following Stage 3 strategies given state $(y, \boldsymbol{\omega})$. In period 1, a single TF i submits an order to Discrete to provide one unit of liquidity at spread $\bar{s}_{discrete}(\tilde{f})$ around y; if he has liquidity outstanding from the previous trading game, he maintains, adjusts or withdraws it as necessary so that he provides exactly one unit at spread $\bar{s}_{discrete}(\tilde{f})$ around y. All other TFs $k \neq i$ do not provide any liquidity (and withdraw any existing liquidity if present). In period 2: an investor trades at least one unit of liquidity, prioritizing orders across exchanges indexed by j based on the lowest value of $s_j/2 + f_j$ and breaking ties in any arbitrary fashion, and then also trades against any remaining profitable orders; informed traders trade against any profitable orders; and if there is a publicly observable jump in y, TF i sends messages to cancel all liquidity providing orders, and all other TFs attempt to trade against any profitable orders (but are unable to do so in equilibrium). Using similar arguments used in the proof of Lemma A.1, it is straightforward to show that there are no safe profitable price improvements or robust deviations in Period 1, or profitable deviations in Period 2, and hence these strategies comprise an OBE for the Stage 3 trading game. In particular, in Period 1, any increase by TF i in its spread on Discrete to $\bar{s}_{discrete}(\tilde{f}) + \varepsilon$ for any $\varepsilon > 0$ and any amount of liquidity is not a robust deviation, as it is rendered unprofitable by a safe profitable price improvement from another TF to provide the same amount of liquidity at spread $\bar{s}_{discrete}(\tilde{f}) + \varepsilon/2$ on Discrete.

Uniqueness. We now establish that in any Stage 3 OBE, exactly one unit of liquidity is provided on Discrete following Period 1 at spread $\bar{s}_{discrete}(\tilde{f})$, and no liquidity is provided elsewhere. By the same arguments in Lemma

[80] Let G_{jump} denote the jump size distribution and g_{jump} its associated density. To establish that $(L_s(s,f) - L_f(s,f)/2) \leq 0$, note that

$$L(s,f) = E(J - \frac{s}{2} + f | J > \frac{s}{2} + f) Pr(J > \frac{s}{2} + f) = \int_{\frac{s}{2}+f}^{\infty} [t - \frac{s}{2} + f] g_{jump}(t) dt \, .$$

Hence,

$$\begin{aligned} L_s(s,f) &= -\int_{\frac{s}{2}+f}^{\infty} \frac{g_{jump}(t)}{2} dt - [(\frac{s}{2}+f) - \frac{s}{2} + f] \times \frac{g_{jump}(s/2+f)}{2} = -\frac{(1 - G_{jump}(s/2+f))}{2} - f \times g_{jump}(\frac{s}{2}+f), \\ L_f(s,f) &= \int_{\frac{s}{2}+f}^{\infty} g_{jump}(t) dt - [(\frac{s}{2}+f) - \frac{s}{2} + f] \times g_{jump}(\frac{s}{2}+f) = (1 - G_{jump}(\frac{s}{2}+f)) - 2f \times g_{jump}(\frac{s}{2}+f), \end{aligned}$$

and $(L_s(s,f) - L_f(s,f)/2) = -(1 - G_{jump}(s/2+f))$, which is weakly negative since $G_{jump}(x) \leq 1$ for all x.

A.1, we establish that exactly one unit of liquidity must be provided in any trading game equilibrium. Now consider a candidate equilibrium where some positive amount of liquidity is provided on Continuous: such liquidity cannot be provided at spread strictly less than $\bar{s}_{continuous}$ (the zero-variable profit spread on Continuous), else the liquidity provider would be better off withdrawing its order; at any spread weakly greater than $\bar{s}_{continuous}$, there is a safe profitable price improvement by any slow TF to provide the same amount of liquidity on Discrete at some spread $s' \in (\bar{s}_{discrete}(\tilde{f}), \bar{s}_{continuous} - 2\tilde{f})$, implying these strategies cannot be an equilibrium.[81] Last, consider a candidate equilibrium where any amount of liquidity on Discrete is provided at some spread $\tilde{s} \neq \bar{s}_{discrete}(\tilde{f})$: if provided at a smaller spread, such liquidity is better off being withdrawn (as it is less than the zero-variable profit spread on Discrete given informed trading); and if provided at a greater spread, there is a safe profitable price improvement by any slow TF to provide the same amount of liquidity at any spread $s' \in (\bar{s}_{discrete}(\tilde{f}), \tilde{s})$.

A.3.2 Proof of Proposition 5.2

Existence. First, we establish that if Discrete charged any trading fee $f' > f^*_{discrete}$ (where $f^*_{discrete}$ is the solution to equation (A.5)) and Continuous had zero trading fees, then in any Stage 3 equilibrium, no liquidity can be provided on Discrete. To see why, assume instead that there is positive liquidity provided on Discrete in some equilibrium. The lowest spread at which liquidity could be profitably offered on Discrete is the zero-variable profit spread $\bar{s}_{discrete}(f')$. At this spread, the total price considered by an investor contemplating trading on Discrete is $\bar{s}_{discrete}(f')/2 + f' > \bar{s}_{continuous}/2$ (by Lemma A.3 and the definition of $f^*_{discrete}$). This implies that there exists a safe profitable price improvement for some fast TF on Continuous to provide liquidity on Continuous at spread $s' \in (\bar{s}_{continuous}, \bar{s}_{discrete}(f') + 2f')$, as such liquidity at spread s' on Continuous would be preferred by investors than that on Discrete; contradiction.

Now consider the following equilibrium strategies. In Stage 3, market participants use strategies described in the Proof of Proposition 5.1, with the modification that investors break ties in favor of Discrete.[82] In Stage 2, all fast TFs do not purchase ESST from Continuous. In Stage 1, Discrete charges positive trading fees $f^*_{discrete}$; Continuous charges zero trading fees and zero ESST fees. In Stage 1, Continuous has no strictly profitable deviations: any attempt to charge positive trading or ESST fees does not affect profits; negative fees result in strictly negative profits. Discrete also has no strictly profitable deviations: by Proposition 5.1, reducing trading fees yields lower profits for Discrete as it does not affect Stage 3 trading game behavior; and any higher trading fees results in all trading activity in Stage 3 occurring on Continuous and zero profits (as established above). There are no strictly profitable Stage 2 deviations by any TF (as purchasing ESST does not affect profits), and similar arguments used in the Existence portion of the proof for Proposition 5.1 establish that these strategies comprise an OBE of the Stage 3 trading game.

Uniqueness. We establish that in any equilibrium, (i) Discrete charges trading fees equal to $f^*_{discrete}$; (ii) in every iteration of the trading game, exactly one unit of liquidity is offered on Discrete at spread $\bar{s}_{discrete}(f^*_{discrete})$ and no liquidity is provided elsewhere; and (iii) Continuous exchanges earn zero profits. For claim (i), first consider a candidate equilibrium where Discrete charges trading fee $f < f^*_{discrete}$. Since $\frac{\bar{s}_{discrete}(f)}{2} + f < \frac{\bar{s}_{continuous}}{2}$, then by continuity of $\bar{s}_{discrete}(\cdot)$, there exists $f' = f + \varepsilon$ for sufficiently small $\varepsilon > 0$ such that $\frac{\bar{s}_{discrete}(f')}{2} + f' < \frac{\bar{s}_{continuous}}{2}$ and would yield Discrete strictly higher profits as it would still capture all trading volume but obtain higher trading revenues; contradiction. Thus, Discrete cannot charge any trading fee $f < f^*_{discrete}$. Next, consider a candidate equilibrium where Discrete charges trading fee $f > f^*_{discrete}$. In this equilibrium, either Discrete has zero trading volume in Stage 3, or positive trading volume. In the case that Discrete has zero trading volume, since there exists some strictly positive $f' < f^*_{discrete}$ such that $\frac{\bar{s}_{discrete}(f')}{2} + f' < \frac{\bar{s}_{continuous}}{2}$ and yields positive trading volume on Discrete in any Stage 3 equilibrium (by Proposition 5.1), there is a profitable deviation for Discrete to charge f'

[81] As long as the spread on Discrete $s' < \bar{s}_{continuous} - 2\tilde{f}$, an investor would prefer to transact on Discrete (paying $\frac{s'}{2} + \tilde{f}$) than transact on Continuous (paying $\frac{\bar{s}_{continuous}}{2}$), since $\frac{s'}{2} + \tilde{f} < \frac{\bar{s}_{continuous} - 2f}{2} + \tilde{f} = \frac{\bar{s}_{continuous}}{2}$.

[82] If trading fees are restricted to be in discrete units (e.g., in units of \$0.0001), then there also exist equilibria in which investors always break ties in favor of Continuous: in such equilibria, Discrete charges the greatest trading fee f such that $\frac{\bar{s}_{discrete}(f)}{2} + f < \frac{\bar{s}_{continuous}}{2}$, and liquidity is only offered on Discrete in each trading game.

instead; contradiction. In the case that Discrete has positive trading volume (which results if Continuous charges a sufficiently high trading fee), then using similar arguments as in Lemma A.3 and above, there is some positive trading fee $f' > 0$ that Continuous could charge such that $\frac{\bar{s}_{continuous}(f')}{2} + f' < \frac{\bar{s}_{discrete}(f)}{2} + f$, where $\bar{s}_{continuous}(f')$ is the analogous zero-variable profit spread on Continuous given Continuous charges trading fees f', that results in all trading volume occurring on Continuous in any Stage 3 OBE and higher profits for Continuous; contradiction. Thus, Discrete cannot charge any trading fee $f > f^*_{discrete}$. Hence, Discrete must charge trading fees equal to $f^*_{discrete}$. For claim (ii), any equilibria in which strictly less than or strictly greater than one unit of liquidity is provided in aggregate across all exchanges can be ruled out using similar arguments as in Lemma A.1. Now consider a candidate equilibrium in which exactly one unit of liquidity is offered in aggregate, but strictly positive liquidity is offered on Continuous. Then Discrete has a profitable deviation in Stage 1 by instead charging a trading fee $f' = f^*_{discrete} - \varepsilon$ for sufficiently small $\varepsilon > 0$, guaranteeing that all volume transacts on Discrete in Stage 3 and increasing profits for Discrete; contradiction. Claim (iii) directly follows from (i) and (ii).

Profit Bound for Discrete. We now establish that when Discrete charges trading fees equal to $f^*_{discrete}$ and one unit of liquidity is provided in each trading game on Discrete at spread $\bar{s}_{discrete}(f^*_{discrete})$ and no liquidity is provided elsewhere, Discrete earns (in expectation) at least $\frac{N-1}{N}\Pi^*$ per trading game. Substituting in the expression for $\frac{\bar{s}_{continuous}}{2}$ given by (A.5) into (A.2) yields:

$$\lambda_{invest}(\frac{\bar{s}_{discrete}(f^*_{discrete})}{2} + f^*_{discrete}) - \frac{N-1}{N}\lambda_{public}L(\bar{s}_{continuous}) - \lambda_{private}L(\bar{s}_{discrete}(f^*_{discrete}), f^*_{discrete}) = 0.$$

Subtracting equation (A.4) (which characterizes the zero-variable profit spread on Discrete at $f^*_{discrete}$) from this expression and re-arranging yields:

$$\lambda_{invest}(2f^*_{discrete}) = \frac{N-1}{N}\lambda_{public}L(\bar{s}_{continuous}) \ ,$$

where the left-hand side of the equation represents the Discrete exchange's expected revenues from trading fees $f^*_{discrete}$ obtained from only investors (and not from informed traders), and thus is strictly *less than* Discrete's *total* expected revenues per-trading game (which includes trading revenues from both investors and informed traders). The right-hand side of the equation represents $(N-1)/N$ share of the total "sniping prize" at a spread of $\bar{s}_{continuous}$; since $\bar{s}_{continuous} < s^*_{continuous}$ and $L(\cdot)$ is decreasing in the spread, the right-hand side is strictly greater than $(N-1)/N$ share of $\Pi^*_{continuous}$, and the result follows.

A.3.3 Proof of Proposition 5.3

Existence. Consider the following strategies. In Stage 3, market participants use strategies described in the Proof of Proposition 5.1, where investors break ties in favor of one particular Discrete exchange labeled j for convenience, and one unit of liquidity is provided by a single TF i on the Discrete exchange with the lowest trading fees, and solely on exchange j in the case of equal trading fees. In Stage 2, no fast TFs purchase ESST from Continuous. In Stage 1, all Discrete exchanges charge zero trading fees; Continuous charges zero trading fees and zero ESST fees. In Stage 1, Continuous has no profitable deviations: any attempt to charge positive trading or ESST fees does not affect profits; negative fees result in strictly negative profits. Any Discrete exchange also has no strictly profitable deviations: lowering trading fees to be negative incurs losses, and increasing trading fees results in no trading volume and revenues given equilibrium strategies (i.e., all other exchanges charge zero trading fees). Finally, there are no strictly profitable Stage 2 deviations by any TF (as purchasing ESST does not affect profits), and arguments similar to those used in Proposition 5.1 establishes that Stage 3 strategies comprise an OBE.

Uniqueness. We establish that in any equilibrium, (i) at least one Discrete exchange charges zero trading fees; (ii) in every iteration of the trading game, exactly one unit of liquidity is offered only on Discrete exchanges with zero trading fees at spread $\bar{s}_{discrete}(0)$; and (iii) all exchanges and trading firms earn zero profits. For claim (i), consider an equilibrium where all Discrete exchanges charge strictly positive trading fees, and the minimum trading fee is $f > 0$. The same logic underlying why undifferentiated Bertrand competition results in marginal cost pricing implies

that this cannot be an equilibrium: for some Discrete exchange, there exists a profitable Stage 1 deviation to charge a slightly lower trading fee $f' = f - \varepsilon$ for some $\varepsilon > 0$ as this would guarantee that all subsequent trading volume would occur on that exchange (as the same arguments used in Proposition 5.1 straightforwardly establish that all Stage 3 trading game equilibria involve all trading volume occurring on the Discrete exchange with the lowest trading fees). Contradiction. Claim (ii) follows directly from the arguments used in Proposition 5.1. Claim (iii) directly follows from claims (i) and (ii).

A.3.4 Prisoner's Dilemma Payoffs

Lemma A.4. *Assume that ESST fees $\boldsymbol{F}^*$ satisfy condition (3.2). Then $NF_j^* < \Pi^D$ for any exchange j.*

Proof. Let $\bar{F}$ be the most that any exchange j can charge for ESST fees given (3.2). This maximum is realized for exchange j when all other exchanges charge 0; condition (3.2) then becomes $\bar{F} \leq \frac{1}{N}\Pi^*_{continuous} - \pi_N^{lone-wolf} < \frac{1}{(N-1)N}\Pi^*_{continuous}$, where the last inequality follows from the lower bound on $\pi_N^{lone-wolf}$ given by Lemma A.2. Hence, $N\bar{F} < \frac{1}{N-1}\Pi^*_{continuous}$. Since $\Pi^D > \frac{N-1}{N}\Pi^*_{continuous}$ by Proposition 5.2 (and since $N \geq 3$), the result follows. □

B Additional Empirical Evidence on the Stage 3 Trading Game

In this appendix, we report additional evidence on the Stage 3 Trading Game presented in Section 4.1. Specifically, we present Stylized Facts #1 and #2 for the "Top 8" exchanges in 2015, which include the 3 taker-maker exchanges in addition to the 5 maker-taker exchanges that we study in the main text. We also present a version of Figure 4.3 from Stylized Fact #3 from the beginning of the Reg NMS era. Before going into the the additional results, we describe the institutional details of taker-maker exchanges, which are necessary to interpret Stylized Facts #1 and #2 with the expanded Top 8 sample.

Taker-Maker Exchanges. In our sample of the Top 8 exchanges in 2015, there were 3 exchanges which utilized the "taker-maker" pricing model, in which the taker of liquidity (i.e., the submitter of an order that trades against a resting bid or offer) gets a rebate and the maker pays a fee (i.e., the resting bid or offer). This is in contrast to the "maker-taker" pricing model, which is used by the other 5 exchanges in the Top 8, where the taker pays a fee and the maker receives a rebate. Together, these 8 exchanges account for 98% of share volume, with the taker-maker exchanges holding 15% volume share.

The difference between taker-maker and maker-taker fee structures results in slightly different prices for takers of liquidity. When a liquidity provider offers liquidity at the same price on both a maker-taker and a taker-maker exchange, it is in effect offering a better price on the taker-maker exchange despite being quoted at the same price. For example, say the rebate on the two types of exchanges is \$0.0029 (with the rebate going to the taker on the taker-maker exchange and to the maker on the maker-taker exchange), the fee is \$0.0030 (paid by the liquidity provider on a taker-maker and by the taker on a maker-taker), and the net fee collected by the exchange, after receiving the fee and paying the rebate, is \$0.0001. Then the taker of liquidity would save \$0.0059 on the taker-maker exchange relative to the maker-taker exchange when both have quotes at the same price — the taker would receive \$0.0029 as a rebate from trading on the taker-maker as opposed to paying a fee of \$0.0030 on the maker-taker.

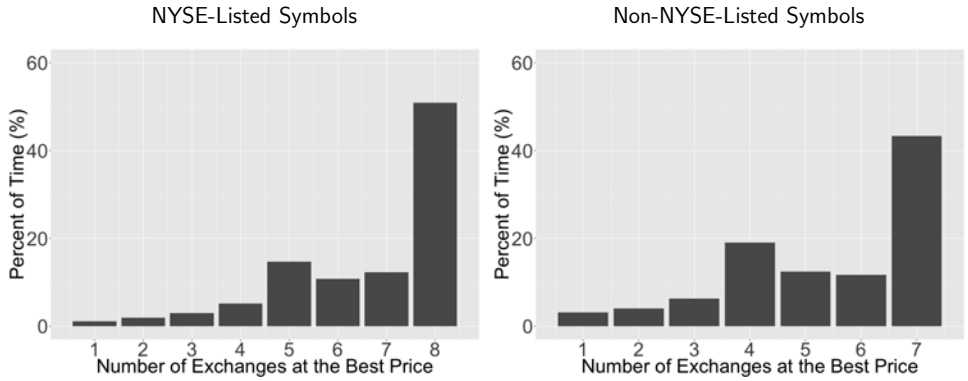
In case this is not clear, imagine there is depth on both a taker-maker exchange and a maker-taker exchange at the national best offer price, say \$12.34. Also as above, let the rebate be \$0.0029 and the fee be \$0.0030. Then the net-of-fee purchase price on the taker-maker exchange for the taker of liquidity when the rebate is \$0.0029 is \$12.34 - \$0.0029 = \$12.3371, while the net-of-fee price on the maker-taker exchange is \$12.34 + \$0.0030 = \$12.343. The difference between the two exchanges is \$12.343 - \$12.3371 = \$0.0059. For the liquidity provider, the net-of-fee sale price on the taker-maker exchange is \$12.34 - \$0.0030 = \$12.3370, while the net-of-fee price on the maker-taker exchange is \$12.34 + \$0.0029 = \$12.3429. This example shows that depth on a taker-maker exchange is economically more attractive and would be consumed first by attentive traders. This effect is observable in the the additional evidence for Stylized Facts #1 and #2 below.

We next present Stylized Facts #1 and #2 for the Top 8 exchanges, as well as Stylized Fact #3 over an extended sample.

Stylized Fact #1: Many Exchanges Simultaneously at the Best Bid and Best Offer (Additional Evidence) Figure B.1 presents results for the Top 8 exchanges, i.e., all exchanges with meaningful market share. We present the results separately for NYSE-listed symbols and non-NYSE listed symbols because, as mentioned in the main text, non-NYSE listed symbols do not trade on NYSE (but do trade everywhere else) and NYSE listed symbols trade everywhere. Hence, for NYSE listed symbols the maximum number of exchanges out of the Top 8 that could be at the best bid or offer is 8, whereas for non-NYSE listed symbols the maximum is 7.

If we look at the Top 8, all exchanges are at the best bid (similarly, best offer) in about 50% of milliseconds. There is a small peak at 5 exchanges for NYSE-listed and 4 exchanges for non-NYSE listed stocks. As mentioned above, if a liquidity provider quotes the same price on a taker-maker exchange as on a maker-taker exchange, it is in effect offering a price that is roughly half a penny (the sum of the fee and rebate) better for the taker of liquidity and worse for itself as the provider of liquidity (see Table 4.1 in Section 4.2 for exact numbers). Therefore it makes economic sense that it will often be the case that the best price is found on all of the maker-taker exchanges, 5 exchanges for

Figure B.1: Multiple Exchanges at the Same Best Price
Top 8 Exchanges (Main "Maker-Takers" and "Taker-Makers")



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 price was equal to N. The figure considers the Top 8 exchanges; for discussion of Top 8 see the text. 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 eight exchanges, all measured at the end of the millisecond. Sample is 100 highest volume symbols that satisfy data-cleaning filters (see text for description) on all dates in 2015.

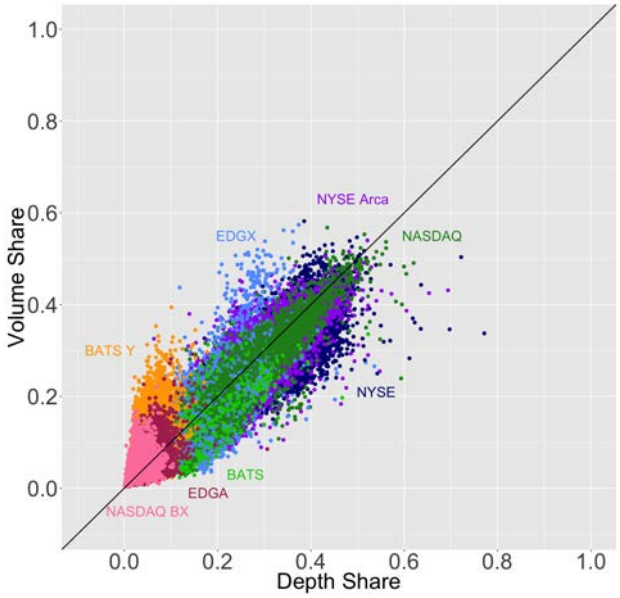
NYSE-listed and 4 exchanges for non-NYSE listed, but not on the taker-maker exchanges. It is rare that only one or a few exchanges are at the best price.

Stylized Fact #2: Depth Equals Volume. (Additional Evidence) Figure B.2 presents a scatterplot of $VolumeShare_{ijt}$ against $DepthShare_{ijt}$ for the Top 8 exchanges (see the main text in Section 4.1 Stylized Fact #2 for details on how we calculate shares). Each dot in the figure represents a symbol-exchange-date tuple. Dots are color coded by exchange and are labeled in the figure.

The figure shows that most of the depth-volume data falls along the 45 degree line. Formally, 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. Just as the figure in the main text, the Top 5 exchanges (NYSE, Nasdaq, NYSE Arca, BATS, and EDGX) are tightly scattered along the 45 degree line. The taker-maker exchanges (EDGA, BATS Y, Nasdaq BX) are clustered at the bottom left of the figure and have a steeper slope than the maker-taker exchanges. That is, the taker-maker exchanges have volume shares that are typically greater than depth shares. The reason for this, as mentioned in the example above, is that the taker-maker exchanges pay a rebate to the taker of liquidity. Thus, while depth on taker-maker exchanges is comparatively rare, when there is depth on taker-maker exchanges it is more economically attractive, after accounting for fees, than depth at the same pre-fee price on the larger maker-taker exchanges. For this reason, it makes sense that taker-maker exchanges have volume shares that are larger than their depth shares.

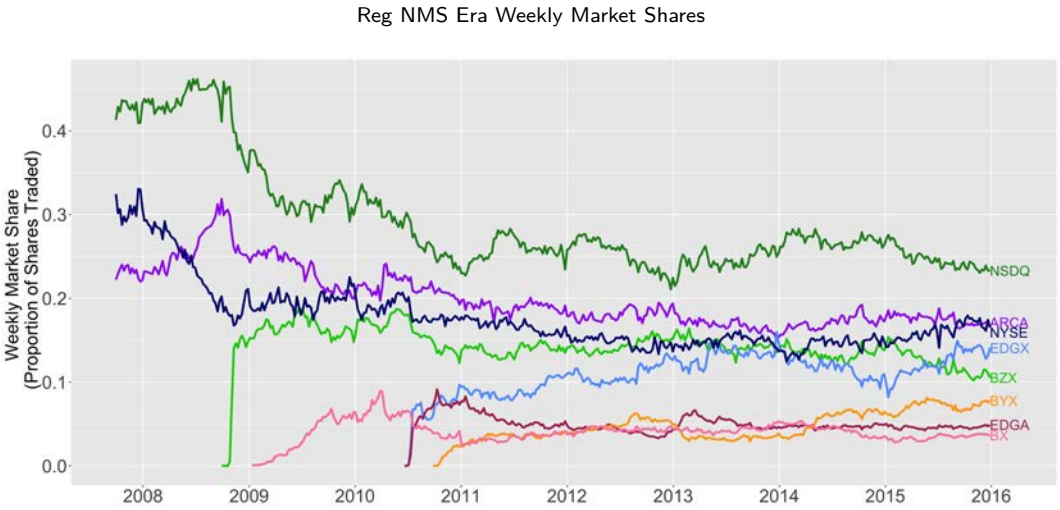
Stylized Fact #3: Exchange Market Shares are Interior and Relatively Stable. (Additional Evidence) In the main text for Stylized Fact #3, we presented aggregate weekly exchange market shares from January 2011 to December 2015 for the Top 5 maker-taker exchanges. Figure B.3 presents exchange market shares from October 2007, the start of the Reg NMS era, through the end of 2015. There are several "jumps" in the data, in particular for BZX in late 2008 and EDGA and EDGX in 2010. Although these exchanges were officially approved at the time of the jump, they operated as off-exchange venues, or ATS's, and had significant market shares before they were officially approved. Thus, although the data show jumps in market share when these exchanges were approved, the market share change in going from an ATS to an exchange was likely much more smooth.

Figure B.2: 2015 Daily Volume Share vs. Depth Share: Top 8 Exchanges



Notes: The data is from NYSE TAQ for the Top 8 exchanges. The dark line depicts the 45-degree line which is the depth share to volume share relationship predicted by the theory. Observations are symbol-date-exchange shares, with shares calculated among the Top 8. For details of share calculations see the text. Sample is 100 highest volume symbols that satisfy data-cleaning filters (see text for description) on all dates in 2015.

Figure B.3: Exchange Market Shares: 2007 – 2015

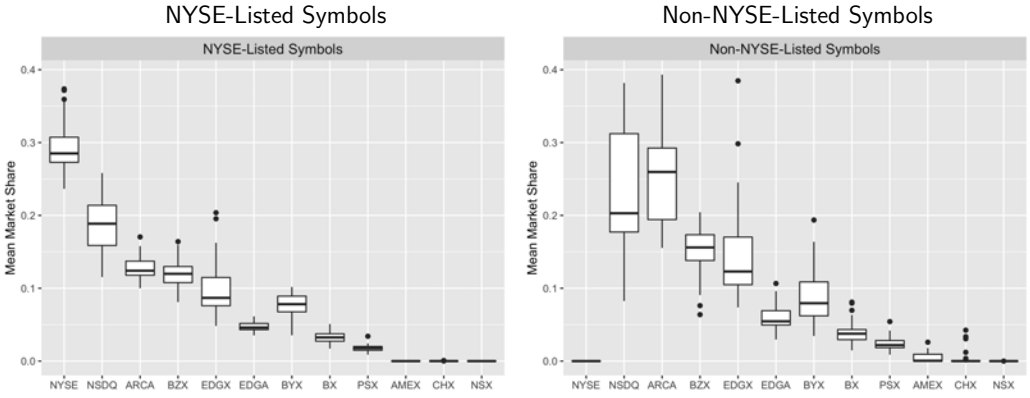


Notes: The data is from NYSE TAQ and covers October 2007 to December 2015 for the Top 8 exchanges. The market shares are based on all on-exchange trading volume in shares.

Figure B.4 explores how market shares vary across stocks. For each stock in the top 100, we compute its average market share per exchange over all dates in 2015. We then present this data as a box plot. Each box represents the 25th-75th percentile range for symbol market shares on that exchange, with the solid horizontal line in the middle of the box representing the median. The lines above and below the box represent the full range, with dots for outliers.

As can be seen, while there is of course variation across symbols, most of the variation in the data is driven by the exchange.

Figure B.4: 2015 Exchange Market Shares: Per Stock



Notes: The data is from NYSE TAQ. Observations are symbol-exchange averages of symbol-exchange-date market shares in 2015. In a given box, the middle line is the median and the edges of the box are the 25th and 75th percentiles. The lines on top of and below the box (whiskers) go out to the interquartile range multiplied by +/-1.5. The dots are symbol-exchange outliers that fall outside of that range.

C Supporting Details for Evidence on Trading Fees (f)

In this appendix, we provide supporting details for the evidence in Stylized Fact #4 presented in Section 4.2. Appendix C.1 provides supporting material for the range of trading fees from historical exchange fee schedules, and Appendix C.2 provides supporting material for the average f per exchange family.

C.1 Expanded Table of Range of U.S. Equity Exchange Trading Fees

Table C.1 shows the range of regular-hours trading fees for the top 8 exchanges as reported in historical fee schedules. This table extends Table 4.1 in Section 4.2 by adding the fee ranges for special fee programs, as well as the fee ranges by listing exchange (referred to as the "Tape" a stock is included on: Tape "A" represents stocks listed on NYSE, "B" on NYSE Arca, and "C" on Nasdaq). As discussed in the main text, 7 of the 8 exchanges have total fees that are negative for at least one fee program. The only exception is EDGA, which has a minimum total fee per share per side of \$0.00005 of 0.5 mills.

C.2 Details for Average Trading Fee Calculations in Table 4.1

In this appendix, we provide supporting details for the calculation of average per-share per-side trading fees as reported in Table 4.1. The calculations themselves are available in a supporting spreadsheet available in the online appendix.

BATS. For BATS, the April 2016 S-1 provides a net trading revenue figure of \$81.0M, as we reported in the text of Section 4.3 and a matched share volume figure of 1.5 billion shares per day which corresponds to 378 billion shares per year (252 trading days). We cross-checked the volume figure with the NYSE TAQ data and found 367.9 billion in that data set, which is within rounding error. Using the S-1 figures for consistency with what follows, we obtain net revenue per share of \$81M/378B=\$0.000214 which corresponds to \$0.000107 per-share per side. This figure includes revenue from regular-hours trading, which is what we want, but it also includes revenue from opening and closing auctions and routing, which we want to strip out.

For BATS, the auction volume is minimal (0.13B per NYSE TAQ), so even under the assumption that all auction volume pays the maximum auction fee (which, depending on the order type utilized, ranges from zero to 5 mills for the opening auction and 10 mills for the closing auction), auction revenue does not move the needle. Routing volume on the other hand is significant, at approximately 25.2 billion shares per the S-1 (0.1B per day times 252 trading days). BATS reports routing and clearing costs of \$43.7M in their S-1, which is 17.3 mills per share. We use a variety of data regarding routing fees based on the ultimate destination of the trade (e.g., a directed ISO versus a take on another exchange versus a take on a dark venue) to obtain a back-of-envelope estimate for BATS's routing revenue per share of 22.8 mills per share and hence net routing revenue of 5.5 mills per share. This in turn implies net routing revenue for BATS overall of 5.5 mills * 25.2 B shares = \$13.8 million. Subtracting this revenue from BATS's total revenue as reported above yields \$67.2 million of regular-hours trading revenue, or \$0.000178 per share and \$0.000089 per-share per-side, as reported in the Table. We caveat that the routing estimate is particularly back-of-envelope, so the reader may prefer to utilize the \$0.000107 figure reported above or to adjust for routing in some other way.

Nasdaq. Nasdaq last reported U.S. cash equity net trading revenues in 2013; in 2015, they only report equity net trading revenues globally, not for the U.S. Our first step therefore is to take the 2015 number for global cash equity trading revenues less transaction-based expenses, of \$253 million, and multiply it by the 2013 ratio of U.S.:Global equity trading revenues, which was \$107M/\$193M = 55%. This yields \$140.3M for 2015 U.S. cash equity net trading revenues. Nasdaq reports matched U.S. equity share volume of 327.7 billion; this is close to the figure we obtain in the TAQ as a cross-check (329.4 B). We thus obtain net revenue per share of \$140.3M/327.7B = \$0.000428, or \$0.000214 per-share per-side. We caveat that this figure will be incorrect if the 2015 U.S.:Global ratio is meaningfully different from the 2013 ratio.

Table C.1: U.S. Equity Exchange Trading Fees Per Share ("f")

				Maker Fee		Taker Fee		Total fee per share		Total fee per share per side	
Exchange	Fee Type	Program	Tape	Min	Max	Min	Max	Min	Max	Min	Max
NASDAQ	Maker-Taker	Regular	C	-0.00325	-0.00150	0.00300	0.00300	-0.00025	0.00150	-0.00013	0.00075
NASDAQ	Maker-Taker	DLP	C	-0.00400		0.00300	0.00300	-0.00100		-0.00050	
NASDAQ	Maker-Taker	Regular	A/B	-0.00325	-0.00200	0.00295	0.00300	-0.00030	0.00100	-0.00015	0.00050
BATS BZX	Maker-Taker	Regular	NA	-0.00320	-0.00200	0.00300	0.00300	-0.00020	0.00100	-0.00010	0.00050
BATS BZX	Maker-Taker	NBBO Setter	NA	-0.00360		0.00300	0.00300	-0.00060		-0.00030	
EDGX	Maker-Taker	Regular	NA	-0.00320	-0.00200	0.00300	0.00300	-0.00020	0.00100	-0.00010	0.00050
NYSE	Maker-Taker	Regular	A	-0.00220	-0.00140	0.00270	0.00270	0.00050	0.00130	0.00025	0.00065
NYSE	Maker-Taker	SLP	A	-0.00290		0.00270	0.00270	-0.00020		-0.00010	
NYSE	Maker-Taker	DMM	A	-0.00350		0.00270	0.00270	-0.00080		-0.00040	
NYSE Arca	Maker-Taker	Regular	B	-0.00270	-0.00200	0.00280	0.00300	0.00010	0.00100	0.00005	0.00050
NYSE Arca	Maker-Taker	LMM	B	-0.00450		0.00250	0.00250	-0.00200		-0.00100	
NYSE Arca	Maker-Taker	Regular	A/C	-0.00300	-0.00200	0.00300	0.00300	0.00000	0.00100	0.00000	0.00050
BATS BYX	Taker-Maker	Regular	NA	0.00140	0.00180	-0.00160	-0.00160	-0.00020	0.00020	-0.00010	0.00010
BATS BYX	Taker-Maker	NBBO Setter	NA	0.00130		-0.00160	-0.00160	-0.00030		-0.00015	
EDGA	Taker-Maker	Regular	NA	0.00030	0.00050	-0.00020	-0.00020	0.00010	0.00030	0.00005	0.00015
NASDAQ BX	Taker-Maker	Regular	NA	0.00165	0.00200	-0.00150	-0.00040	0.00015	0.00160	0.00008	0.00080
NASDAQ BX	Taker-Maker	QMM	NA	0.00140		-0.00150	-0.00040	-0.00010		-0.00005	

Notes: This table summarizes the fee schedules for the top 8 exchanges retrieved from Internet Archive (Wayback Machine) dated from February 28, 2015 to September 1, 2015 (BATS Global Markets, Inc., 2015a,b,c,d; Nasdaq, Inc., 2015a,b; NYSE, 2015a; NYSE Arca Equities, Inc., 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 consider all volume-based incentives for regular-hours liquidity provision, but we do not include additional incentives for trading off hours, trading at the open or close, creating non-displayed midpoint liquidity, sending retail orders, routing, or for trading securities with a share price below $1. The "Regular" program corresponds to the fees and rebates a firm would receive if it does not qualify for additional incentive programs detailed below, which often either involve an additional volume threshold, a National Best Bid and Offer quoting requirement, or an off-hours trading requirement. The Designated Liquidity Provider (Nasdaq DLP) program rewards market participants who maintain a one or two-sided quote on specified Nasdaq-listed ETFs for at least 15% of the trading day. The National Best Bid or Offer Setter (BZX/BYX NBBO Setter) program rewards participants who send orders that set the new national best bid or offer, as well as fulfill an additional volume requirement. The Supplemental Liquidity Provider (NYSE SLP) program rewards participants who quote at the NBBO at least 10% of the trading day, as well as fulfill an additional volume requirement. The Designated Market Maker (NYSE DMM) program rewards participants who make commitments to satisfy a wide variety of requirements involving market depth, volume, NBBO quoting, capital, and others every month. The Qualified Market Maker (Nasdaq BX QMM) program grants a discount on making liquidity for participants who actively quote at the NBBO. We also separately report fees by "Tape" or listing exchange. Tape "A" represents stocks listed on NYSE, "B" on NYSE Arca, and "C" on Nasdaq. In the table, NA indicates that an exchange does not charge different fees by listing exchange.

The next step is to deduct auction volume and revenue, which are both significant. We obtain auction volume from TAQ, of 5.3B annually for the opening auction and 20.2B for the closing auction. For the opening auction, we use a fee of 15 mills per-share per-side, which is the fee for regular market-on-open and limit-on-open orders that participate in the auction. This ignores fees for some other less-common order types as well as a fee cap for high-volume users of $20,000 per month. For the closing auction, Nasdaq has a fee schedule with 6 tiers based on volume levels. The fee ranges from 8 mills for the highest-volume tier to 15 mills for the lowest. We assume an equal six-way split across the six tiers to obtain 12.1 mills. Together, the opening and closing auction account for 25.5B shares traded and $64.6 million of revenue.

Last, we deduct routing revenue. Routing is prominently discussed in Nasdaq financial statements but they do not report any specific numbers. Since the Nasdaq routing business appears to be at least somewhat similar to the BATS routing business, we utilize the BATS net routing revenue per share number computed above (5.5 mills) and the BATS routed volume as a % of total volume (6.7%), to obtain net routing revenue of $12.0 million on 21.8 billion shares.

When we subtract auction revenue and volume, and subtract routing revenue, we obtain 302.2 billion regular-hours shares traded and $63.6 million of regular-hours net trading revenue, for $0.000211 per share and $0.000105 per-share per-side, as reported in the table. As a sensitivity analysis, we assume that we have overestimated auction revenues by 25%, for example, due to the monthly fee caps. This would change the figure to $0.000132 per-share per-side.

NYSE NYSE's parent company, Intercontinental Exchange (ICE), reports in its 2015 10-K that NYSE's U.S. cash equities revenues, net of transaction based expenses, were $220 million in 2015. The ICE 10-K reports average daily matched volume of 1,187M shares for Tape A, 296M shares for Tape B, and 206M for Tape C. Multiplied by 252 trading days this yields annual volume of 425.6 billion shares, which is close to the TAQ number. This yields revenue per share of $220M/425.6B=$0.000517, or $0.000258 per-share per-side.

Next, we deduct auction revenue and volume. We get opening and closing auction volume for NYSE, NYSE Arca, and NYSE Mkt from the TAQ data. These volumes are significant for both NYSE and NYSE Arca, with 11.1B and 1.9B of volume for the open, and 48.4B and 9.7B of volume for the close, respectively. For the opening auction, we use a fee of 10 mills for NYSE and NYSE Mkt and 15 mills for NYSE Arca, based on their fee schedules. As with Nasdaq, there are some discounts (in particular for NYSE designated market makers) and monthly caps, which we do not attempt to account for here, but rather do so below in a sensitivity analysis. For the closing auction, NYSE has a range of fees from 6 mills to 10 mills depending on volume tier; we use an equal-weighted average of the tiers to obtain 7.7 mills. NYSE Arca's closing auction fee is 10 mills and NYSE Mkt's is 8.5 mills. Combined across these three venues and combining both the open and close, we obtain $123.3M of total auction revenue.

For routed volume, we utilize that the ICE 10-K reports both matched volume and handled volume; the difference is what is routed. This comes to 10.8 billion shares annualized across the 3 tapes. We utilize the same 5.5 mills net routing fee number from BATS, lacking any better source. This comes to $5.9M of total routing revenue.

When we subtract auction revenue and volume, and subtract routing revenue, we obtain 353.5 billion regular-hours shares traded and $90.7 million of regular-hours net trading revenue, for $0.000257 per share and $0.000128 per-share per-side, as reported in the table. As a sensitivity analysis, we assume that we have overestimated auction revenues by 25%, for example, due to the monthly fee caps. This would change the figure to $0.000172 per-share per-side.

D Supporting Details for Evidence on Exchange-Specific Speed Technology Revenue (F)

In this appendix, we provide supporting details for the evidence presented in Section 4.3. Appendix D.1 provides supporting material for the magnitude of ESST revenue documented in Stylized Fact #6. Appendix D.2 provides supporting material for the magnitude of ESST revenue documented in Stylized Fact #7.

D.1 Details for Data and Co-Location Revenue Estimates for Nasdaq and NYSE

In this appendix, we provide supporting details for our calculations of market data and co-location/connectivity revenues for Nasdaq and NYSE in 2015, which we reported in Section 4.3.

Nasdaq's fiscal year 2015 10-K reports market data and co-location/connectivity revenue only at the global level – $399M and $239M, respectively.[83] To get from global to the U.S., for market data, we utilize information in its 2013 10-K filing that breaks out its market data business geographically: U.S. is 72% of the total in 2013, and we assume this ratio holds in 2015. For co-location/connectivity, we use Nasdaq's overall 2015 U.S.:global revenue ratio, of 71%. Last, we need to separate out Nasdaq's U.S. Equities business from its U.S. Options business. We take two approaches. First, we assume that Nasdaq's market data and co-location revenue from U.S. Equities vs. U.S. Options is proportional to its trading volume in U.S. Equities vs. U.S. Options. Second, we assume that Nasdaq's U.S. Options business generates the same market data and co-location revenue revenue as BATS's U.S. Options business, scaled up for Nasdaq's larger U.S. Options volume than BATS. The first approach assumes that every 1 option traded on Nasdaq generates the same market data and co-location revenue as 100 shares of stock; the second approach assumes that 1 option traded on Nasdaq generates the same market data and co-location/connectivity revenue as 1 option traded on BATS. These two approaches yield a range for Nasdaq's U.S. Equities revenue of $222.4M-$267.3M for Market Data, $121.0M-$139.0M for Co-Location/Connectivity, and $343.3M-$406.4M combined.

NYSE was acquired by ICE, a large futures exchange conglomerate, in Nov 2013. ICE's 2014 10-K filing therefore gives significant detail on the contribution of the NYSE business to the overall ICE business, for 2014, the first full year of integration (and also for the Nov-Dec 2013 period). The filing reports that NYSE's U.S. businesses (not including Euronext, which ICE divested) contributed $430M to its data services business in 2014; this includes both market data and co-location/connectivity, for both U.S. equities and U.S. options. The filing also reports that $202M of this was for co-location/connectivity, implying $228M for market data. ICE's 2015 10-K filing reports that it reclassified an additional $60M of revenue, for 2014, from its "other" category to its data services business, and that this revenue corresponds to "NYSE connectivity fees and colocation service revenues".[84] Therefore the adjusted 2014 totals are $262M for co-location/connectivity and $228M for market data. Comparison of ICE's 2014 and 2015 10-K filings suggest a growth rate of its overall data services business from 2014 to 2015, of which the NYSE business was by far the largest component, of 12.3%.[85] For comparison, BATS's U.S. equities growth rate for the 2014 to 2015 period was 19.2% for co-location/connectivity and 12.4% for market data,[86] which suggests that the 12.3% growth rate computed from ICE data is reasonable for NYSE. We use this growth rate to obtain estimates for 2015 for NYSE's

[83] The $239M for global co-location/connectivity also contains revenues from a small Nordic region Broker Services business, which when last reported separately was $19M; we subtract out this $19M from Nasdaq's global "Access and Broker Services" business in the analysis that follows.

[84] It is hard to know but we guess that this adjustment reflects post-merger alignment of accounting practices between NYSE and ICE, that in principle should have been reflected in the 2014 10-K but that was not completed until the 2015 10-K.

[85] This growth figure accounts for several other ICE acquisitions in this time period. 2014 ICE data services revenue was $691M but includes just 12 weeks of the SuperDerivatives business, which contributed $12M in those 12 weeks; therefore 2014 revenue pro forma for the SuperDerivatives business was $731M. 2015 data services revenue was $871M but includes $50M of revenues from 2015 acquisitions of Interactive Data and Trayport; therefore a like-for-like 2015 revenue number is $821M, or 12.3% more than the adjusted 2014 figure.

[86] BATS's 2014 numbers include just 11 months of Direct Edge revenue versus 12 months in 2015. If we conservatively assume that the Direct Edge business is 50% of BATS's overall business, then we can take the unadjusted 2014-to-2015 growth rates, of 23.7% for co-location/connectivity and 16.9% for market data, and reduce them by $50\% \cdot \frac{1}{11} \approx 4.5$ percentage points, to obtain 2014 to 2015 growth rates that are apples-to-apples.

overall U.S. business, and then utilize the same two methods described above for Nasdaq to obtain estimates for NYSE's U.S. Equities business. This yields a range of $218.9-$241.5M for U.S. equities market data, $251.6-$281.5M for U.S. equities co-lo/connectivity, and $470.5-$523.0M combined.

D.2 Details for Data and Co-Location Revenue Growth Estimates for Nasdaq and BATS

In this appendix we provide supporting details for the data on Nasdaq and BATS exchange-specific speed technology (ESST) revenue growth discussed in Stylized Fact #7. As discussed in the main text, our goal is to get a sense of magnitudes for ESST growth over time by looking at revenue growth in the financial reporting categories that contain U.S. equities ESST revenues.

From 2006 to 2012, Nasdaq reported co-location/connectivity revenues in the category "Access services revenues." In 2013, Nasdaq changed the reporting category to "Access and Broker Services Revenues," which incorporated Nasdaq's small Nordic broker services business ($19M in 2012) into the category. In 2015, Nasdaq appears to have adjusted its accounting practices to reclassify some revenue in "Access and Broker Services Revenues" to "Technology Solutions," which led to a downward revision of $18M for the 2014 revenue figures based on the 2015 reporting method as reported in Nasdaq's fiscal year 2015 10-K, so it is of a similar magnitude as the upward revision in 2013. In 2016, Nasdaq changed the reporting category again to "Trade Management Services Revenues," but this appears to be a change in the category name only with no revision to revenue figures reported in previous 10-Ks. As noted in the main text, we view the periods 2006-2012 and 2015-2017 as yielding reliable apples-to-apples growth rates, and the period 2012-2015 seems relatively flat with the caveat that there were multiple reporting changes.

From 2006 to 2012, Nasdaq reported its U.S. equities proprietary market data revenue in the category "U.S. market data products." Over this period, Nasdaq also separately reports tape revenues as "Net U.S. tape plans." Starting in 2013, Nasdaq reports only combined U.S. data revenue (i.e., including both proprietary and tape) in the segment "U.S. market data products," and also made some segment reporting changes, moving $27M of revenue from "Index data products" out of U.S. market data products into its own reporting category. Starting in 2014, Nasdaq reports only total market data revenue instead of separating out U.S., international, and index market data separately. To get a roughly apples-to-apples time series, we make the following two adjustments. First, from 2014 onwards, we use the 2013 ratio of U.S. to total market data revenue (72%) to get a U.S. market data revenue estimate. Second, from 2013 onwards, we subtract out 2012 tape revenue of $117M to get to U.S. market data revenue excluding tape plan revenue.[87] These two assumptions together imply that any revenue growth in Nasdaq's total market data category since 2014 is attributed 72% to Nasdaq's U.S. proprietary market data segment with the remaining 28% to international and index market data revenues. Our sense is that this convention is conservative since Nasdaq reports that both international market data revenue and index data revenue were relatively flat in the years leading up to 2013 (International: $83M in 2011 to $77M in 2013; Index: $24M in 2011 to $27M in 2013).

We use four data sources for BATS co-location/connectivity revenue: BATS's 2012 S-1 statement (revenue from 2009-2011), BATS's 2016 S-1 statement (revenue from 2010-2015), the CBOE/BATS 2016 proxy statement (revenue for 9 months of 2016, which we annualize), and CBOE's 2017 annual report (which reports BATS's contribution to CBOE revenues for 10 months, which we annualize; CBOE's acquisition of BATS was finalized on Feb 28 2017). BATS states in its 2012 S-1 that it began charging for co-location/connectivity revenue, described as "port fees," in Q4 2009 (pg. 64) so we report numbers starting in 2010. Before 2012, the reporting segment was called "Other Revenues" in BATS's 2012 S-1; BATS describes the category by stating "Other revenues consist of port fees, which represent fees paid for connectivity to our markets, and, more recently, additional value-added products revenues." The reporting

[87] We feel comfortable treating tape revenues as flat since 2012 since tape revenues are based on depth and market shares, and we show in Stylized Fact #3 that market shares are roughly flat and show in Stylized Fact #2 that depth shares line up one-for one with market shares. Nasdaq's market shares by share volume from 2011 to 2017 were: 29.1%, 29.5%, 28.9%, 31.3%, 28.0%, 26.0%, 28.0%. Moreover, in the last three years that Nasdaq did report tape revenues separately, they were essentially constant: $117M in 2010, $115M in 2011, and $117M in 2012.

segment changed to "Port Fees and Other" in BATS's 2016 S-1; the revenue reported for 2011 in BATS's 2016 S-1 is within 1% of the 2011 revenue reported in BATS's 2012 S-1 so we conclude that the change was almost entirely a renaming of the reporting category rather than a substantive change. The reporting segment changed again to "Connectivity Fees and Other" for 2016 in the CBOE/BATS proxy statement, which was a change in name only (revenue reported from previous years are consistent with the 2016 S-1). In 2017, as a part of CBOE, BATS's revenue from co-location/connectivity is split across two CBOE segments, "Access fees" and "Exchange services and other fees." CBOE separately reports BATS's contributions to these categories and we report their sum, annualized to twelve months.

BATS reports in its 2016 S-1 that its two BATS exchanges, BZY and BYX, only began charging for proprietary market data in Q3 2014 (pg. 94). We thus use numbers starting in 2015 (this is also the first full year that revenue associated with Direct Edge is included in BATS's filings). BATS market data revenue for U.S. equities is included in the category "Market Data Fees" in BATS's 2016 S-1 (2015 revenue) and in the 2016 CBOE/BATS proxy statement (9 months of revenue from 2016, which we annualize). BATS's market data revenue for 2017 comes from CBOE's 2017 annual report, which provides BATS's contribution to the category "Market Data Fees" for 10 months in 2017 (which we annualize). We know that tape revenue is a significant fraction of market data revenue, and utilize percentages provided in BATS's filings to estimate tape revenue, which we can subtract from overall market data revenue as reported in BATS's filings. BATS reports in its 2016 S-1 that 84% of market data revenue in 2015 comes from tape revenue (pg. 21), and reports in the CBOE/BATS merger proxy that 79% comes from tape revenue in 2016 (pg. 56). Using these percentages, we estimate that 2015 tape revenue is $110M and 2016 tape revenue is $116M. We also assume that 2017 tape revenue is flat from 2016 levels.[88] If we subtract these tape revenue estimates from the overall market data revenue reported in BATS's filings, we get $21.0M in 2015, $30.8M in 2016 and $38.3M in 2017 (growth of 35.3% per year). These numbers include data revenues related to BATS's European equities and U.S. options business, so they overstate the level but likely understate the growth rate of BATS's U.S. proprietary market data since 2015.[89]

[88]BATS's combined market share for its four exchanges by share volume declined slightly from 36.3% in 2016 to 34.5% in 2017, so if anything 2017 tape revenues would be slightly smaller than 2016.

[89]The CBOE/BATS proxy statement reports on pg. 311-312 that, of the growth in market data revenue of $10.7M for the first 9 months of 2016 versus the same period in 2015, $7.4M came from U.S. proprietary market data ("pricing changes in proprietary market data that were implemented in the third quarter of 2015 and the first quarter of 2016") versus $0.7M from U.S. options (pg. 312) and $0.7M from European equities (pg. 318). Unfortunately, we are not able to ascertain any specific information about growth in BATS's proprietary market data revenues from 2016 onwards; once BATS became incorporated into CBOE reporting became even less granular.

E Discussion of Discrete vs. Continuous with Tick-Size Constraints and Agency Frictions

In this Appendix we discuss competition between a Discrete exchange and one or more Continuous exchanges when there are tick-size constraints and agency frictions. As emphasized in the main text, these frictions are reasons why we suggest that the reader not take the 100% tipping aspect of Propositions 5.1-5.2 literally.

For the purpose of this discussion we make the following assumptions:

- Tick-size constraints. We assume that stocks trade in increments of $ticksize$ = \$0.01. This reflects tick-size regulations in U.S. stock markets for stocks with a nominal share price greater than \$1 (Reg NMS Rule 612).
- Agency frictions. We assume that whenever there is liquidity at the same quoted price on Continuous and Discrete, investors always break ties in favor of Continuous. This tie-breaking in favor of Continuous is independent of any fee differences (discussed below). Investors route orders to Discrete only if Discrete has a quoted price that is at least a full tick better for the investor, i.e., when they are mandated to do so under the Reg NMS order protection rule. This assumption is meant to capture, in a simple and worst-case way, any agency frictions that might favor trading on Continuous markets over Discrete markets (in the spirit of Battalio, Corwin and Jennings (2016)).
- Maker-Taker fees. We assume that the Continuous market uses a maker-taker fee schedule with a take fee equal to the regulatory maximum under the Reg NMS Access Rule of +30 mills (\$0.0030) per share, and a make fee equal to -30 mills per share (i.e., make rebate of 30 mills), for a net fee of 0. This approximates current practice as discussed in Section 4.2. Additionally, in conjunction with the tie-breaking assumption in the previous bullet point, this assumption about fees is a worst-case for Discrete. Since we have assumed that investors only route to Discrete when liquidity is a full tick more attractively priced, Discrete would like to charge investors a higher fee than they pay on Continuous — investors would be willing to pay a higher fee conditional on trading, since whenever they trade on Discrete they are saving a full tick. However, the Continuous market is already charging investors the regulatory maximum. Therefore, if Discrete is to charge a positive fee, it does so by charging a take fee of 30 mills, just as on Continuous, and a make fee of $-(30 - f_D)$ mills (i.e., rebate of $30 - f_D$ mills), where f_D denotes the net fee to Discrete per share.
- Distribution of fundamental values. For the purpose of discussion, we assume that the fundamental value y of the security is uniformly distributed such that all values between any two relevant ticks are equally likely.
- Magnitude of latency arbitrage. For the purpose of discussion, we utilize the estimate from Aquilina, Budish and O'Neill (2019) that the latency arbitrage tax on liquidity is 0.42 basis points (0.0042%) of traded volume.[90]

Given our assumptions about tick-sizes and fees, if there were a single Discrete exchange operating in isolation, the equilibrium best offer would be $\left\lceil y + \frac{s^*_{discrete}}{2} - (\$0.0030 - f_D) \right\rceil$, where the notation $\lceil x \rceil$ denotes rounding up to the nearest whole-penny increment. (For simplicity, we focus discussion on just the offer, the bid being symmetric.) If there were a single Continuous exchange operating in isolation, the equilibrium best offer would be $\left\lceil y + \frac{s^*_{continuous}}{2} - (\$0.0030) \right\rceil$. TFs would only be able to offer liquidity on Discrete at a strictly better whole-penny increment than on continuous if

$$\left\lceil y + \frac{s^*_{discrete}}{2} - (\$0.0030 - f_D) \right\rceil < \left\lceil y + \frac{s^*_{continuous}}{2} - (\$0.0030) \right\rceil , \tag{E.1}$$

that is, if the magnitude of latency arbitrage, represented by $\frac{s^*_{continuous}}{2} - \frac{s^*_{discrete}}{2}$, is large enough to "cross a tick" given the current fundamental value y, and accounting for any difference in fees and rebates.

[90] As shown in Section 5.5 of Aquilina, Budish and O'Neill (2019), for the purpose of computing the amount by which eliminating latency arbitrage would reduce the cost of liquidity, it is technically more accurate to use latency-arbitrage profits as a proportion of non-race traded volume as opposed to as a proportion of all traded volume. This non-race volume latency arbitrage tax is 0.53 basis points as opposed to 0.42 basis points. To err on the side of conservatism we use the 0.42 basis points figure.

Table E.1: FBA Market Share and Revenue with Tick-Size Constraints and Agency Frictions

FBA Net Fee Per Share (Mills) (*)	FBA Share (% of Share Volume)	FBA Share (% of Dollar Volume)	FBA Annual Revenue ($ Millions)
0	18.7%	37.2%	0.0
1	17.7%	36.2%	15.7
2	16.7%	35.2%	29.6
3	15.8%	34.3%	41.8
4	14.8%	33.3%	52.5
5	14.0%	32.4%	61.8
6	13.2%	31.4%	69.9
7	12.4%	30.5%	76.8
8	11.7%	29.6%	82.8
9	11.0%	28.8%	87.8
10	10.4%	27.9%	92.1
15	7.8%	23.9%	103.8
20	6.0%	20.5%	105.9
25	4.7%	17.6%	103.0
30	3.7%	15.2%	97.2

(*) FBA Net Fee = Take Fee + Make Fee. As discussed in the text, we hold fixed the Take Fee at 30 mills ($0.0030) per share, and vary the Make Fee from -30 mills (i.e., rebate of $0.0030 per share) to 0.

Notes: Data from NYSE TAQ in 2015. The symbol universe is all stocks with $shareprice_i$ >$5 that traded continuously throughout the year under the same ticker. We first calculate $FBAshare_i$ for each symbol i according to equation (E.3), with $shareprice_i$ calculated as the volume-weighted average trade price of the symbol over all trading days in 2015, and f_D set to the values in the "FBA Net Fee Per Share" column. The 0.0042% figure in (E.3) for the magnitude of latency arbitrage is from Aquilina, Budish and O'Neill (2019) as discussed in the text. We then use the FBA market shares for each symbol to compute overall FBA market shares by share volume and dollar volume, with overall market shares expressed relative to the symbol universe. FBA annual revenue is computed as overall FBA share volume times the net fee f_D.

Given the assumption that fundamental values are uniformly distributed between any two relevant ticks, the probability that the fundamental value y satisfies condition (E.1) is:

$$\frac{\frac{s^*_{continuous}}{2} - \frac{s^*_{discrete}}{2} - f_D}{ticksize}, \tag{E.2}$$

truncated below at 0 (in case f_D is too large) and above at 1 (in case $\frac{s^*_{continuous}}{2} - \frac{s^*_{discrete}}{2} - f_D$ exceeds the tick size). For example, if $\frac{s^*_{continuous}}{2} - \frac{s^*_{discrete}}{2} - f_D = \$0.0020$, then the probability that y is such that condition (E.1) holds is 20%. Our assumption about the magnitude of latency arbitrage enables us to compute Discrete's market share for symbol i from (E.2) as:

$$FBAshare_i = \frac{(0.0042\% \cdot shareprice_i - f_D)}{\$0.01} \tag{E.3}$$

again truncating below at 0 and above at 1.

We use TAQ data in 2015 to compute (E.3) for all symbols with $shareprice_i$ >$5 that traded continuously throughout the year under the same ticker, with $shareprice_i$ calculated as the volume-weighted average trade price for symbol i and net fees per share f_D ranging from 0 mills to 30 mills. A net fee f_D of 0 mills corresponds to a take fee of +30 mills and a make fee of -30 mills (i.e., a rebate) as on the Continuous market, whereas a net fee f_D of 30 mills corresponds to a take fee of +30 mills and a make fee of 0. We then use $FBAshare_i$ to compute overall FBA market shares by share volume and dollar volume (relative to the symbol universe). We also use $FBAshare_i$ to compute annual FBA revenue, which is $FBAshare_i$ multiplied by f_D and by overall share volume for symbol i, then summed over all symbols. The results are summarized in Table E.1.

At a net fee of $f_D = 0$, i.e., the same maker-taker fee structure as the Continuous market, the Discrete exchange's share is computed as 18.7% in share volume and 37.2% in dollar volume. The reason why dollar volume is meaningfully higher than share volume is that tick-size constraints are less binding for high nominal share price stocks.

At a net fee of $f_D = 10$ mills, the Discrete exchange's share is computed as 10.4% in share volume and 27.9% in dollar volume. Fee revenues are $92.1 million per year. Fee revenues are maximized at about $106 million per year, using a net fee of about 20 mills.

Clearly, this exercise is very back-of-the-envelope. In particular, it utilizes a magnitude for latency arbitrage taken from UK equity markets, whereas if better data were available in U.S. equity markets it would be possible to directly calculate both the average level and the cross-sectional heterogeneity across stocks and ETFs of latency arbitrage in the U.S.. It also incorporates complicated market frictions in a very stylized manner. Nevertheless, we hope that this exercise provides the reader with a useful sense of magnitudes both for interpreting the Discrete vs. Continuous theoretical results in Section 5.2 and for thinking about the market design exclusivity period idea in Section 6.2.