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

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

May 6, 2019

Abstract

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 the continuous limit order book market design, a design that gives rise to latency arbitrage—arbitrage rents from symmetrically observed public information—and the associated high-frequency trading arms race (Budish, Cramton and Shim, 2015). Will the market adopt new market designs that address the negative aspects of high-frequency trading? This paper builds a theoretical model of stock exchange competition to answer this question. Our model, shaped by institutional and regulatory details of the U.S. equities market, shows that under the status quo market design: (i) trading behavior across the many distinct exchanges is as if there is just a single "synthesized" exchange, as opposed to traditional platform competition; (ii) as a result, trading fees are perfectly competitive; but (iii) exchanges capture and maintain significant economic rents from the sale of "speed technology" (i.e., proprietary data feeds and co-location)—arms for the arms race. Using a variety of data, we document seven stylized empirical facts that suggest that the model captures the essential economics of how U.S. stock exchanges compete and make money in the modern era. We then use the model to examine the private and social incentives for market design innovation. We show that the market design adoption game among incumbent exchanges is not a coordination game, but rather a repeated prisoner's dilemma. If an exchange adopts a new market design that eliminates latency arbitrage, it would win share and earn economic rents. However, imitation by other exchanges would result in an equilibrium that resembles the status quo with competitive trading fees, but now without the rents from the speed race. This means that although the social returns to market design innovation are large, the private returns are much smaller and may be *negative*, especially for incumbents that derive rents from the status quo. Despite this negative result, however, our analysis does not imply that a market-wide market design mandate is necessary to fix the problem. Rather, it suggests a modest regulatory "push" may be sufficient to tip the balance of incentives and encourage "the market to fix the market."

*Project start date: April 2015. We are especially grateful to Larry Glosten and Terry Hendershott for serving as discussants of an early version of this project. We also thank 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, Luigi Zingales, and numerous industry practitioners and seminar participants for helpful discussions and suggestions. Paul Kim, Cameron Taylor, Matthew O'Keefe, Natalia Drozdoff, and Ethan Che provided exceptional research assistance. Budish acknowledges financial support from the Fama-Miller Center, the Stigler Center, and the University of Chicago Booth School of Business. Disclosure: the authors declare that they have no relevant or material financial interests that relate to the research described in this paper. John Shim worked at Jump Trading, a high-frequency trading firm, from 2006-2011.

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

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

§University of Chicago Booth School of Business, john.shim@chicagobooth.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, and offends common economic intuitions about what constitutes an efficient market. Budish, Cramton and Shim showed that the fix is conceptually pretty simple, 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, *batch* process them using a standard uniform-price auction. However, despite the remarks of the SEC Chair above and encouragement from others,[1] the main U.S. stock exchanges have not shown much interest in changing their market design. In the two instances where a startup (IEX) and a small exchange (CHX) proposed market design ideas similarly motivated by concerns about aspects of high-frequency trading, the proposals were met with fierce resistance from the large incumbent exchanges and from many high-frequency trading firms.[2,3]

This begs the question: are private forces alone sufficient to foster the adoption of innovative and more

[1] See New York Attorney General Schneiderman (2014); Bloomberg Editorial Board (2014), which expressed views of both Bloomberg's editors and Goldman Sachs; 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] Our sense is that the resistance to the Investors' Exchange (IEX)—an unprecedented 477 SEC comment letters were filed regarding its exchange application, including one in which the New York Stock Exchange wrote "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*)—reflected a genuine mixture of incumbents' desire to preserve the status quo and legitimate concerns about the details of IEX's market design. The most important concern, from the perspective of the current paper, is that IEX's market design only protects against latency arbitrage for non-displayed pegged orders; it does *not* protect against latency arbitrage for conventional displayed limit orders. Its displayed ("lit") market is identical to a standard continuous limit order book, just 350 microseconds further away from market participants than would be the case from geography alone, due to the famous "speed bump". Please see Budish (2016*b*) for further details.

[3] The Chicago Stock Exchange (CHX) proposed to adopt an "asymmetric" speed bump for its exchange—in which liquidity taking orders are delayed but liquidity providing orders are not delayed—which would have protected against latency arbitrage in the displayed market, unlike IEX's symmetric speed bump. But it, too, met with fierce resistance from the larger exchanges and several high-frequency trading firms—for example, Citadel wrote that it "unfairly structurally and systematically discriminates against market participants that are primarily liquidity takers." (Citadel, 2016) CHX ultimately withdrew the proposal and was acquired by the New York Stock Exchange (Michaels and Osipovich, 2018). Please see Baldauf and Mollner (2018*a*) for a detailed theoretical analysis of asymmetric speed bumps, Section VIII.C-D of Budish, Cramton and Shim (2015) for additional discussion of speed bumps, and Budish (2016*a*) for further details of CHX's proposal.

efficient market designs? 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 (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). The question of whether or not sufficient innovation incentives exist for stock exchanges is timely and of significant importance. If they do, then it follows that "prescriptive regulation" should not mandate a specific market design; rather, regulators should ensure that regulation does not "stand in the way" of "competitive solutions." If not, intervention may be warranted—all the more so if the potential economic savings are substantial.[4]

In this paper, we argue that incumbent exchanges do not have sufficient incentives to adopt new market designs *precisely because they derive rents from the inefficiencies that these alternative designs seek to eliminate*. That is, even though there are multiple exchanges 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 at stake in the speed race. 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 (cf. 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.

Central to our argument is a novel theoretical model of the stock exchange industry, tailored to the institutional and regulatory details that shape modern electronic trading, and built to understand the nature of exchange competition and the associated incentives for innovation. There are four types of players in our model, all strategic: exchanges, trading firms, investors, and informed traders. Initially, mirroring the *status quo* of the current market, we assume that all exchanges employ the continuous limit order book market design. Exchanges are undifferentiated, and 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 also 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.[5] Specifically: all

[4]While there is no exact consensus in the academic literature on the economic stakes in the high-frequency trading arms race, the academic estimates that are available suggest that it is in the single-digit billions of dollars per year for U.S. equities. Taking a net present value of this amount, and extrapolating across countries and other financial instruments, it is easy to get to a net present value figure in excess of $100 billion. See Budish (2017) for discussion.

[5]Glosten (1994) presciently foresaw that frictionless search and order-splitting across electronic markets (cf. his Assumption

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 linear 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).[6] 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 (cf. 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

4) could generate what we refer to as the single synthesized exchange (cf. his Proposition 8), well over a decade before the passage of Reg NMS. Please see Section 3.2.4 for a detailed discussion of the relationship between Glosten (1994) and this aspect of our analysis.

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

latency arbitrage.[7] 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.[8]

Although our model is highly stylized and abstracts from several real-world-complications (which include agency frictions, tick-sizes, asymmetric trading fees, and strategic trading over time as in Kyle-style models), we establish that our parsimonious model nonetheless does reasonably well in matching several empirical moments found in the data. Specifically, using a combination of trades-and-quotes (TAQ) data and exchange-company financial filings (e.g., 10-K's, S-1's, merger proxies, fee filings), we document seven stylized facts about modern-era stock exchange competition that align with the model. Our first series of facts relates to our result that the market behaves as if trading activity occurred on a single synthesized exchange. These facts, documented using a sample of reasonably highly traded stocks, include all major exchanges typically having displayed liquidity at the same best price; a close linear relationship between the quantity of liquidity on an exchange (i.e., its displayed depth) and its trading volume; and market shares that are interior (i.e., no tipping). Next, we document that trading fees across major exchanges are economically very small. Trading fees are quite complicated (cf. Chao, Yao and Ye (2019)), but using a variety of data sources to cut through this complexity, we compute that 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 implies that across approximately 1 trillion shares traded during regular hours each year, exchanges earn approximately \$200 million in trading fees. To put this in perspective, StubHub, the largest secondary-market venue for concert and sports tickets, has revenues exceeding \$1 billion; that is, StubHub's revenue is over five times that for all U.S. regular-hours equities trading, despite the secondary market for event tickets being a tiny fraction of the secondary market for U.S. equities. Last, we document that exchanges earn significant revenues from the sale of co-location services and proprietary data feeds. For the BATS exchange family, for which the data is the cleanest, revenue from co-location and data is about 69% of total revenue. In aggregate across the three major exchange families (BATS, Nasdaq, NYSE), we document significant growth in ESST fees during the Reg NMS era (post 2007), with 2018 speed technology revenues estimated to be on the order of \$1 billion.

Overall, the empirical facts suggest that our simple model—though explicitly abstracting away from many aspects of modern U.S. stock trading—captures the essential economics of how U.S. stock exchanges compete and make money in the modern era. Also importantly, the empirical facts we document for the U.S. stock exchange industry, taken in total, are *not* consistent with many other models of financial exchange competition in prior literature. These include models that feature single-homing, network effects, market tipping, supra-competitive trading fees, and so forth (e.g., Pagano (1989); Ellison and Fudenberg (2003); Cantillon and Yin (2008)); models in which exchanges are meaningfully horizontally or vertically differentiated (e.g., Baldauf and Mollner (2018*b*); Pagnotta and Philippon (2018)); and models in which tick-size frictions are central (Chao, Yao and Ye (2017, 2019)). Nor are these facts consistent with standard models of platform competition from other economic contexts (cf. Rochet and Tirole (2003, 2006); Farrell and Klemperer (2007)).

[7] 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 about 20% of the total latency arbitrage prize. Please see Section 3.2.4 for discussion.

[8] A particularly extreme version of this move was announced very recently in January 2019 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*). 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).

The last part of our analysis uses the model to address our overarching question: will the market adopt new market designs, such as frequent batch auctions (Budish, Cramton and Shim, 2015), that address the negative aspects of high-frequency trading? How do exchanges' private innovation incentives relate to social incentives? Our model suggests that exchanges are unlikely to embrace such innovation with open arms. However, it is not because a new market design that eliminates latency arbitrage would fail to be utilized by market participants—indeed, if introduced, it would gain significant market share—but rather because such innovation would destroy the rents that incumbent exchanges currently earn from speed technology. To conduct this analysis, we extend our theoretical model to allow for exchanges to operate one of two market designs: either the continuous-time limit order book (Continuous), or discrete-time frequent batch auctions (Discrete). Importantly, in the context of competition with the Continuous market, we consider frequent batch auctions with a very short batch interval: long enough to effectively batch process if multiple trading firms react to the same public signal at the same time, but otherwise essentially as short as possible.[9]

We show that if only one exchange employs Discrete while all others employ Continuous, the Discrete exchange will capture a large share of trading volume and large economic rents. 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. That is, the standard coordination problems associated with getting a new market off the ground do not apply here, and the Discrete exchange is able to earn trading fees commensurate with the tax that it eliminates.[10,11] However, this frictionless world is also a double-edged sword: although it rewards a first-mover, it also implies that any subsequent adoption of Discrete by other exchanges—which our model suggests is likely—leads to the same Bertrand competition on trading fees as before, but now without the industry rents from the speed race. Thus, we establish that the market design adoption game among incumbent exchanges is essentially a repeated 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. Similar arguments imply that a de novo entrant exchange using Discrete would have difficulty recouping any substantial fixed costs of entry if imitation by other exchanges is likely and rapid. Thus, we conclude that private incentives to adopt a new market design that eliminates latency arbitrage are dramatically lower than social incentives, and potentially even negative.

Finally, we discuss policy implications. The basic question is whether (i) there will be a private-market solution to latency arbitrage and the arms race (i.e., "will the market fix the market"), or (ii) would some sort of regulatory intervention—ranging from a market-wide market design mandate to something more circumscribed—be required. Our analysis suggests that although private incentives alone may be insufficient,

[9] In practice, given advances in speed technology over the last several years, 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 batch interval of 1 millisecond or less would also allow the frequent batch auction exchange to operate within the framework of 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.

[10] 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 5.1.2 for further discussion.

[11] In our model, with continuous prices, the unique equilibrium is for Discrete to win 100% market share as long as the sniping tax is strictly positive. In a richer model, with tick-size constraints (i.e., a restriction that prices must be in increments of $0.01), if the sniping cost per share is smaller than the tick size—which seems empirically to be the case—then there also may exist equilibria without complete tipping in which Discrete's market share depends on the ratio of sniping costs per share to the tick size. Please see Section 5.1.2 for further discussion.

a modest regulatory "push"—one that tips the balance of incentives enough to get a de novo exchange to enter or an incumbent to adopt—might suffice. The rough intuition for why a push may suffice, as opposed to needing an all-out mandate, is that *investors* strictly prefer markets without the latency arbitrage tax, and investors are ultimately who exchanges and trading firms make their money from—so, once such a market enters, private-market forces can take over. Such pushes might include: (i) reducing the entry and adoption costs of launching a new stock exchange, for example, by lowering the risk of failed entry by clarifying which types of market designs would be admissible under Reg NMS, or finding some way to subsidize the fixed costs of entry; and (ii) a modest regulatory exclusivity period for the innovator, during which competing exchanges would not be able to imitate the design. Analogous to FDA exclusivity periods for non-patentable drugs, an SEC exclusivity period could induce a first-mover exchange to invest the fixed costs associated with developing, implementing and gaining regulatory approval for a new market design.

Our paper makes several contributions to the literature. First is our theoretical industrial organization model of the stock exchange industry, described by some as the single most iconic market in global capitalism (e.g., 60 Minutes (2014)). 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 (2018*b*)), or by financial instruments that are specific to a single exchange (as in Cantillon and Yin (2008)). This single-homing is often (though not always) accompanied by some meaningful differentiation across exchanges, either horizontally or vertically. By contrast, in our model, motivated by the regulatory environment for modern electronic U.S. stock trading, stocks are fungible across exchanges, market participants can frictionlessly multi-home across exchanges, and exchanges are undifferentiated. This modeling approach also leads to "economics of the status quo" that are 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 approximately zero 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 (cf. 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, in our analysis we develop and motivate an equilibrium solution concept which we refer to as an *order-book equilibrium* to address equilibrium existence issues. This solution concept is closely related to alternative solution concepts employed in the insurance market literature (e.g., Wilson (1977); Riley (1979)), and may prove useful analyzing other markets with 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). Though we acknowledge that these particular contributions (and our modeling exercise overall) may be highly tailored for a specific market, we believe that this specificity is justified by the importance of the industry.

Our paper's second contribution is the seven stylized facts that, to our knowledge, have not been documented in this form elsewhere. 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 (cf. 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 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 constitutes a prisoner's dilemma (as opposed to, e.g., a coordination game), the rest of the analysis and discussion is straightforward. We use these insights to identity a modest policy response, well short of the "prescriptive regulation" that SEC Chair White expressed wariness of. 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.[12]

Roadmap. The remainder of this paper is organized as follows. In Section 2, we describe the key institutional features of the stock market that shape our theoretical model of exchange competition. We introduce and analyze the model of the status quo in Section 3. Section 4 provides our seven stylized empirical facts. In Section 5, we use our theoretical model to examine competition among alternative market designs. Section 6 proposes potential policy responses, and Section 7 concludes.

2 Institutional Background

Readers of this paper—especially researchers who are less familiar with financial market microstructure—may have in mind, when thinking of stock exchanges and how they compete, the old New York Stock

[12]In this spirit, our paper belongs to an active literature at the intersection of finance and market design, with some recent works including 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), Glode and Opp (2016), Hendershott and Madhavan (2015), Hortaçsu, Kastl and Zhang (2018), Kastl (2017), Kyle and Lee (2017), Kyle, Obizhaeva and Wang (2018), and Menkveld, Yueshen and Zhu (2017).

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,[13] a large majority of its volume transacted on the Nasdaq exchange (86% in 1993).[14] 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 (see also Ellison and Fudenberg, 2003).

This model, however, is less applicable for the modern era of stock trading.[15] 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 (cf. 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. 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.[16]

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.[17] Regulations for futures exchanges, on the other hand, are quite different from those for stock exchanges, both in the U.S. and abroad. In particular, there is no analogue of UTP in futures markets because each contract is proprietary to a particular exchange. This has a significant effect on the industrial organization of futures markets as distinct from stock markets, as we will discuss briefly under Stylized Fact #6 in Section 4.3 (see especially Figure 4.5). 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,

[13] 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.

[14] 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.

[15] 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).

[16] 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.

[17] 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.

etc.; in particular, the information dissemination provisions of Reg NMS are often economically different in these asset classes. Again, our focus will be on U.S. stock exchanges, though we think there is much interesting future research to do on the industrial organization of financial exchanges for other kinds of assets, under other regulatory regimes, and so forth.

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. 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).

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 routes tracing to the Securities Exchange Act Amendments of 1975 and the SEC's "Order Handling Rules" promulgated in 1996.[18] For the purpose of the present paper, however, there are two core features to highlight.[19]

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

[18]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)

[19]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.[20]

A June 2016 rules clarification issued by the SEC indicated that exchanges can use market designs that impose delays on the processing of orders and still qualify as "immediate and automatic" so long as (i) the delay is of a de minimis level of 1 millisecond or less, and (ii) the purpose of the delay is consistent with the efficiency and fairness goals of the 1934 Exchange Act (U.S. Securities and Exchange Commission, 2016*b*). This rules clarification suggests that quotes in a frequent batch auction exchange would be protected under Rule 611 so long as the batch interval was no longer than the de minimis threshold of 1 millisecond; however, this specific market design has not yet been put before the SEC for explicit approval.[21]

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.[22]

The second key provision to highlight is the Access Rule, or Rule 610. Intuitively, for an exchange to be able to comply with the Order Protection Rule it must be able to efficiently obtain the necessary information about quotes on other exchanges, and, if necessary, be able to efficiently route orders to trade against quotes on other exchanges. Similarly, a broker-dealer seeking to comply with the Order Protection Rule using ISOs must be able to efficiently obtain information about quotes from all exchanges, and efficiently trade against quotes on all exchanges. 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)[23]

[20] A central issue in the debate over IEX's exchange application was whether IEX's quotes, given that its market design included a "speed bump", would count as immediately and automatically accessible under Reg NMS. See U.S. Securities and Exchange Commission (2016*a*) for the (unprecedented number of) public comments on IEX's exchange application. See also footnote 2 for additional details regarding IEX's market design.

[21] To date, the one market design that has been approved by the SEC that imposes a de minimis delay is that of IEX; the SEC issued its rules interpretation of "immediate and automatic" simultaneously with its approval of IEX's exchange application, with both issued 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, 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. 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.

[22] For further details on intermarket sweep orders see the text of Reg NMS (U.S. Securities and Exchange Commission, 2005). Formally, the relevant aspects of the regulation are Rule 600(b)(30) for the definition of ISOs, Rule 611(b)(5) for the exchange's exemption from ensuring compliance with the Order Protection Rule for ISOs, and Rule 611(c) for this compliance obligation instead residing in the sender of the ISO.

[23] 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,

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. As the SEC writes (pg. 27), "The adopted rule thereby assures order routers that displayed prices are, within a limited range, true prices." 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, cf. 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.[24] 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, which are not allowed to be discriminatory. 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 develop a simple model of stock exchange competition to better understand the *status quo* of the market. The model is a necessary building block for our motivating question about market design innovation,

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.

[24] The prohibition against discriminatory trading fees enables what Reg NMS describes as "private linkages" among exchanges, which, roughly, are services that provide data about and access to quotes from all exchanges. The text of Reg NMS describes (pg. 166) that "many different private firms have entered the business of linking with a wide range of trading centers and then offering their customers access to those trading centers through the private firms' linkages. Competitive forces determine the types and costs of these private linkages." Given our focus on the economics of stock exchanges our model will abstract from the competition among linkage providers (e.g., broker-dealers, retail brokers) to offer access to end investors; as we discuss in the conclusion, this seems a fruitful avenue for future research.

analyzed later in Section 5. When presenting our model, we first consider trading on a single non-strategic exchange that operates a continuous limit order book. We then generalize the analysis to consider multiple exchanges that compete with one another in an environment shaped by the key institutional details reviewed in Section 2. We restrict all exchanges to employ the continuous limit order book market design in this section; later when examining the question of market design innovation in Section 5, we allow exchanges to be strategic with respect to their market design choice.

3.1 A Single Non-Strategic Continuous Limit Order Book Exchange

Our baseline model of trading on a single exchange adapts the framework introduced in Budish, Cramton and Shim (2015) (hereafter, BCS), and departs from it in two important ways.

Our first departure is to introduce a stylized version of informed trading in the spirit of Copeland and Galai (1983) and Glosten and Milgrom (1985). The purpose of this modification is to parsimoniously provide a rationale for positive equilibrium bid-ask spreads that is independent of latency arbitrage considerations.

Second, rather than working with a continuous-time model in which certain events—representing the arrival of investors, or jumps in the fundamental value of a security—occur according to exogenous Poisson processes, we work with an infinitely repeated two-period *trading game* in which, 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.[25] This modeling approach will retain the economic interpretability of the continuous-time Poisson model while making some of our assumptions regarding speed and latency more transparent, especially in the multi-exchange case.

One technical contribution of this paper is the development of an alternative equilibrium solution concept that we refer to as an *order-book equilibrium*. This solution concept, in a manner analogous to the alternative equilibrium notions of Wilson (1977) and Riley (1979) used to restore existence in models of insurance (see also Rothschild and Stiglitz (1976)), restricts the set of deviations that an equilibrium in our environment must be robust to, and circumvents existence issues that would otherwise arise. As we discuss further below, our approach can be viewed as complementary to others—which include adopting alternative assumptions, such as perfectly competitive liquidity provision (Glosten and Milgrom, 1985) or "immediate" responses to deviant actions (BCS)—used to generate equilibrium predictions in trading markets.

3.1.1 Setup

We first consider a single exchange that uses the continuous limit order book market design. There is a security, x, that trades on this exchange and a signal, y, which is perfectly correlated to the fundamental value of x. The signal y evolves as a compound jump process, occurring with probability λ_{jump} per trading game and with jumps drawn from a symmetric distribution with bounded support and mean zero. What will matter economically is the absolute value of jump-sizes, denoted by random variable J, which is drawn from what we refer to as the jump-size distribution. We assume that x can always be costlessly liquidated at its fundamental value, can be traded in continuous units (which will facilitate extending our analysis to multiple exchanges), and that prices are continuous so that x can be traded at any price.[26]

[25] Even for the highest activity symbol in all of US equity markets, SPY, on its highest-volume day of 2018 (Feb 6th), 95.2% of milliseconds have neither any trade nor change in the national best bid or offer (price or quantity). For the median volume symbol in the S&P 500 index on its median volume trading day (OMC on November 1st), 99.9% of milliseconds have neither any trade nor change in the national best bid or offer.

[26] We discuss the role of tick-sizes in Section 3.2.5.

Initially we assume that the single exchange is non-strategic and does not charge any trading fees (so that x can be costlessly traded). There are then three types of players (who we refer to as *market participants*) whose actions we focus on: Investors, Informed Traders, and Trading Firms. 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 in the period he arrives 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.[27] Formally, if an investor arrives to market needing to buy one unit of x, buys a unit at price p, and the fundamental value is y, then her payoff is $v + (y - p)$, where v is a large positive constant that represents her inelastic need to trade. If she needs to sell a unit and does so at p when the fundamental value is y, her payoff is $v + (p - y)$.[28] Note that what we call investors could also be termed "noise traders" since they are essentially mechanical.

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. In our current model, 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 $p - y$.[29]

Finally, *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 $N \geq 2$ "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. We model speed as a tie-breaker (cf. Baldauf and Mollner (2018*a*)), meaning that if two firms submit messages to the exchange in the same time period of our trading game, and one is fast while the other is not, the message of the one that is fast will get serially processed first; if both firms are fast or both are slow, the processing order is uniformly random. In this Section, there will be no role in equilibrium for trading firms who do not possess speed technology; hence, we will use TFs to refer to the N fast trading firms unless explicitly noted otherwise.

The following objects are primitives of our single-exchange 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 (for random variable J); and (iii) the number of fast TFs (N).

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

[28] 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.

[29] 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).

3.1.2 Timing

At the beginning of each trading game, there is a publicly observed *state*, which consists of a pair (y, ω) that represents the fundamental value of the security (y), and the currently outstanding bids and asks in the exchange's limit order book (ω); ω is also referred to as the *state of the order book*. If it is the first play of the trading game, the initial fundamental value is y_0, and the order book is initially empty. Otherwise, the state is determined at the conclusion of the previous trading game and ω contains all limit orders that remain outstanding.

Each trading game has two "periods" that happen sequentially:

1. **Period 1:** Trading firms simultaneously send orders to the exchange after observing the state (y, ω) at the beginning of the trading game. Each TF i's *order* is a set of messages denoted by $o_i \in \mathcal{O}$, where $\mathcal{O}$ represents the set of all potential combinations of messages and may contain standard limit orders, cancellations of existing limit orders, and immediate-or-cancel orders. Immediate-or-cancel orders (abbreviated IOCs) behave similarly to standard limit orders, but with proxy instructions to cancel the order immediately if it is not executed (or to cancel whatever portion is not immediately executed). Limit orders and IOCs take the form (q_i, p_i), where such an order states 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. All orders are then processed by the exchange. If multiple market participants send orders in the same period, the messages are serially processed by the exchange in a random sequence, with speed serving as a tie-breaker: if multiple fast TFs send orders in the same period, the sequence in which they are serially processed is uniformly random; if other market participants (including slow trading firms and investors) send orders in the same period, they are processed after those of fast TFs, also in a random sequence. Orders submitted by TFs may affect the state of the order book, ω.

2. **Period 2:** After period-1 orders have been processed by the exchange and incorporated into the state of the order book ω, nature moves and selects one of four possibilities:

   (a) 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 the exchange. The investor's activity may affect ω; y is unchanged.

   (b) With probability $\lambda_{private}$: an informed trader privately observes a jump in y. The informed trader has a single opportunity to send IOCs to the exchange. The informed trader's activity may affect ω; the jump in y then is publicly observed.

   (c) With probability λ_{public}: there is a publicly observable jump in y. All TFs have a single opportunity to send IOCs and cancellation messages to the exchange, potentially affecting ω.[30]

   (d) With probability $1-\lambda_{invest}-\lambda_{private}-\lambda_{public} \geq 0$: there is no event; y and ω are both unchanged.

The state (y, ω) at the end of the trading game remains the state for the beginning of the next trading game.

[30] In period 2 of each trading game, agents are only permitted to send marketable orders (IOCs) and cancellation messages to the exchange; all non-marketable orders that provide liquidity are restricted to be sent in period 1. This restriction does not affect the equilibrium outcome of our game: as investors and informed traders exist only for at most one trading game, they will never wish to employ non-marketable orders; and TFs are allowed to send messages again immediately in period 1 of the following trading game before any additional exogenous events occur.

3.1.3 Equilibrium of the Single Exchange Trading Game

In our infinitely repeated trading game with a single exchange, we restrict attention to pure Markov strategies: market participants are only able to condition their pure strategies on the publicly observable state (y, ω), and not on the history of play in previous trading games. This implies that in period 1, when sending orders to the exchange, all TFs condition their actions only on the state at the beginning of the trading game (including possibly their own outstanding orders in ω); 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.

We proceed by first analyzing each trading game in isolation of others, and ignore the possibility that actions in one trading game may affect continuation payoffs in subsequent games. We then check and show that repeated play of the equilibrium that we construct for a single trading game remains an equilibrium for the infinitely repeated trading game when such interactions are accounted for.

Regardless of which outcome nature chooses in period 2, market participants' optimal strategies in period 2 are straightforward to characterize:

- Upon arrival, an investor and informed trader have 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 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 "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 "speed race" will be one of the fast TFs.

Thus, as market participants have 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.

As noted above, we initially assume that period-1 behavior in each trading game can be analyzed independently of other trading games. In the equilibria that we construct, at the beginning of period 1, there will be a set of TFs choosing to provide liquidity via non-marketable limit orders. As investors are equally likely to arrive needing to buy or sell one unit of x and the distribution of jumps in y is symmetric about zero, it is convenient to focus on equilibria in which any TF wishing to provide liquidity does so via two

limit orders: for a given quantity q and fundamental value y, a TF will submit an order to buy x at $y - s/2$, and an order to sell x at $y + s/2$ for some *bid-ask spread* $s \geq 0$.[31]

To understand how the bid-ask spread is determined in equilibrium, consider a trading firm offering to either buy or sell 1 unit of x at a spread of s (i.e., a bid of $y - s/2$ and an ask of $y + s/2$) when there is no additional liquidity offered in the order book. In traditional models of adverse selection (Copeland and Galai, 1983; Glosten and Milgrom, 1985), the benefit of such liquidity provision 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 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. 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. Here, 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.[32] 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.

Solution Concept. Given our restriction to Markov strategies, a natural solution concept for our trading game is pure-strategy Markov-perfect equilibrium (MPE). However, a pure-strategy MPE does not exist. To see why, consider a potential equilibrium in which, 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

[31] Because there are no profitable trading opportunities at the beginning of any trading game in equilibrium (given market participants' optimal period-2 strategies), optimal period-1 messages sent by TFs will (without loss) consist only of new liquidity providing (non-marketable) limit orders, or cancellations of existing liquidity. With regards to liquidity providing orders, such a pair of orders could also be supplied by two different TFs; this distinction will not matter for our analysis, as equilibrium payoffs for all TFs will be equivalent.

[32] If public and private information had different jump distributions, denoted J_{public} and $J_{private}$, the RHS 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.

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.

This non-existence result arises in our environment because of adverse selection. In a standard model of undifferentiated Bertrand competition among multiple firms without adverse selection, an equilibrium exists with marginal-cost pricing as there are no concerns that offering a product for sale at such a price generates economic losses. In a sense, excess "liquidity" provision does not incur risk, and serves to constrain the price that any given firm can charge. In contrast, in our environment the expected cost of providing liquidity depends on the actions of rivals. Hence, TFs are not willing to provide excess liquidity in the order book to constrain others' spreads—doing so exposes them to adverse selection and sniping risk without the full benefit of being filled by an uninformed investor if one arrived. Thus, because there is no excess liquidity in the order book, the TFs that are providing the non-excess liquidity have a profitable deviation to increase their spreads.

To address this issue while still maintaining the tractability of our discrete time model, we introduce an alternative equilibrium solution concept that modifies MPE. Our new solution concept, which we refer to as an *order book equilibrium,* attempts to capture the spirit of "competitive" liquidity provision, as assumed in Glosten and Milgrom (1985), by allowing for TFs to react to deviations in certain ways. In particular, if a liquidity-providing TF deviates—for example by widening its spread—our concept allows other TFs to provide additional liquidity at a better price if they would wish to do so; thus, even if excess liquidity is not provided in equilibrium, the presence of other *potential* liquidity providers will discipline equilibrium price levels. More subtly, our concept also handles a profitable deviation 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. It does so by allowing any TF whose quotes were undercut by the deviation to withdraw if it would like to do so (e.g., if its liquidity would no longer be filled by an investor), for the purpose of evaluating the profitability of the deviation.[33]

We provide a formal definition of our order book equilibrium concept in Appendix A.1 and an informal one here. An order book equilibrium is a set of orders $\boldsymbol{o}^* \equiv \{o_i^*\}$ submitted by TFs in period 1 of each trading game that conditions only on the current state (y, ω), and satisfies two conditions. First, we require that there are no *safe profitable price improvements.* We say that a strictly profitable unilateral deviation by some TF is a *profitable price improvement* if it provides any quantity of liquidity at a weakly better price than before, and some quantity of liquidity at a strictly better price. (All profit considerations here hold fixed the strategies of all other firms and take expectations over outcomes and optimal play in period 2.) For example, any TF adjusting its orders to provide the same amount of liquidity at a narrower spread represents a price improvement; and any deviation that only adds liquidity (at some finite price) also represents a price improvement. We say that a profitable price improvement is *safe* if it remains strictly profitable even if some other TF were able to withdraw liquidity in response to the deviation. Second, we require that there are no other strictly profitable unilateral deviations—including those that involve the widening of existing spreads—that remain strictly profitable even if a rival TF is able to withdraw liquidity or engage in a safe

[33]The non-existence issues that arise in our setting are abstracted away from in other models of securities trading that we build upon. For example, Glosten and Milgrom (1985) assumes that trading firms (specialists) provide liquidity at the competitive spread, and "do not specify why the specialist should be competitive"; allowing TFs to provide additional liquidity and "undercut" if equilibrium spreads widen captures this competitive notion. Additionally, BCS obtain their results in continuous time, assuming that any deviations—e.g., widening spreads or undercutting—would be met by "immediate" responses; allowing TFs to react to deviations formalizes this idea, in a manner similar to Wilson (1977) and Riley (1979) (see footnote 34 for further discussion).

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.

In addition to being inspired by competitive liquidity provision, our solution concept embodies the idea that each exchange's limit order book settles into a "rest point" in any given trading game, whereby no TF wishes to adjust liquidity provision between any changes in a security's fundamental value or the arrivals of investors and informed traders. The first condition of our equilibrium concept implies that no TF would wish to replace liquidity offered by a rival at a strictly narrower spread; and the second ensures that no TF can profitably widen the spread on any offered liquidity without inducing a rival to wish to replace it.

Similar restrictions on the set of allowable deviations have been employed by alternative solution concepts developed to address equilibrium existence issues 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)).[34] Our relation to this literature is not accidental: both financial and insurance markets are characterized by adverse selection, and in both settings firms that are "undercut" by a rival (who offers a better price, or who offers a product that attracts less adversely selected consumers) may wish to withdraw from the market rather than face an adversely selected set of trading partners.

Equilibrium. We now are in a position to summarize equilibrium behavior in our single-exchange trading game:

Proposition 3.1. *Any order book equilibrium given state (y, ω) with a single continuous limit order book exchange has a single unit of liquidity provided at bid-ask spread $s^*_{continuous}$ (defined in (3.1)) around y following period 1. 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 trading firms endogenously sort themselves into either liquidity provision and latency arbitrage roles in equilibrium; at the equilibrium spread $s^*_{continuous}$, TFs are indifferent between these roles. 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 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 $\lambda_{public} \cdot L(s^*_{continuous})$.

Before continuing, it is helpful to briefly discuss why our order book solution concept helps restore equilibrium existence. Consider a candidate equilibrium where a single unit of liquidity is provided by TF

[34] Both Wilson (1977) and Riley (1979) examine equilibria among firms providing insurance policies, and introduce solution concepts that admit dynamic responses to deviations. A set of policies comprises an *E2 equilibrium* (Wilson, 1977) if there are no strictly profitable unilateral deviations that remain so even if policies, rendered unprofitable by the deviation, are withdrawn. A set of policies comprises a *reactive equilibrium* (Riley, 1979) if there are no unilaterally profitable 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 liquidity could then be withdrawn (similar to Riley).

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 cancel its orders, 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 above), and earn revenues from liquidity provision at a strictly narrower spread while also sniping TF i's existing orders. However, TF i can cancel its own orders in response to TF k's price improvement and render it 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.

Last, we note that repeated play of any order book equilibrium strategies that condition only on the state (y, ω) comprises an equilibrium of the infinitely repeated trading game. This need not have been true, since orders resting in the order book in one period have priority that carries over to subsequent periods, 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.[35]

3.2 Modeling the Status Quo: Multiple Competing Exchanges

The previous single-exchange analysis accomplished three main goals. The first was to adapt the BCS model of trading to a discrete time model with a stylized version of informed trading. The second was to introduce our order book equilibrium concept. The third was to characterize equilibrium behavior in our trading game.

We now extend our analysis to multiple exchanges where the same security x can be traded on multiple exchanges. The key extension here is allowing exchanges to be strategic and compete for rents by charging fees to market participants. We show that there exist equilibria in which the trading game outcomes of the single-exchange case—i.e., a single unit of depth is provided (and transacted by investors) at spread $s^*_{continuous}$—is replicated *across* multiple exchanges. Critically, we also show that exchanges are able to capture a share of the sniping prize without dissipating rents through competition with one another; in turn, this implies that they have incentives to maintain the existing status quo.

3.2.1 Setup and Market Participants

Assume now that there is a set $\mathcal{M}$ of exchanges indexed by j, where the number of exchanges $M = |\mathcal{M}| \geq 2$. Exchanges are exogenously present in the market and undifferentiated from one another.[36] All exchanges use the continuous limit order book market design. As in the single-exchange analysis there is a single security, x, only now that security can be bought or sold on any of the M exchanges, i.e., the asset is completely *fungible* across exchanges. This fungibility assumption captures the economics of Unlisted Trading Privileges regulation, as discussed in detail in Section 2. We continue to assume that shares are perfectly divisible; this allows for a market participant to split his desired order, regardless of size, across multiple exchanges. It is

[35] 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 is no longer appropriate.

[36] We discuss entry incentives for exchanges later in Section 5.

substantively important for the analysis, and also realistic, that participants can split orders across multiple exchanges.

Exchanges operate as strategic players alongside Investors, Informed Traders, and Trading Firms.

Exchanges. Prior to play of the multi-exchange variant of our infinitely repeated trading game (described in detail below), we assume that each exchange j simultaneously sets two prices: a per-share trading fee denoted by f_j, and 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.[37] 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 the exchange's servers), access to fast exchange-specific proprietary data feeds, and connectivity/bandwidth fees.[38] It is modeled as a rental cost per trading game charged to TFs, capturing that in practice exchanges typically assess these fees on a rental basis.

Investors, Informed Traders, and Trading Firms. As before, there are N TFs exogenously in the market endowed with a general-purpose speed technology (enabling their messages to be processed by the exchange before those sent by other market participants in the same period). After exchanges determine their fees, we assume that each TF chooses which exchanges to purchase ESST from, and then the infinitely repeated trading game occurs. In each trading game, we assume that the TFs with highest order priority on a given exchange are those with both general-purpose speed technology and ESST on that exchange; next are those TFs with just the general-purpose speed technology; and last are market participants with neither. As above, we will use speed as a tie-breaker among market participants who act at the same time in any trading game.

In each iteration of our trading game, as before, we assume that an investor arrives to the market in each iteration of our trading game with probability λ_{invest}, and has an inelastic need to buy or sell a unit of x; an investor trades in the period he arrives using IOCs, and then exits the game. Similarly, an informed trader with private information about y arrives each trading game with probability $\lambda_{private}$, and trades immediately using IOCs if profitable. Finally, with probability λ_{public}, a publicly observed jump in y occurs (which, as we discuss, leads to a multi-exchange variant of a sniping race among TFs), where the jump-size distribution remains the same as before.

The primitives of our game are the same as in the single exchange case, with the addition of the number of exchanges (M).

Reg NMS. We capture the key provisions of Reg NMS (Rules 610 and 611, see Section 2 for discussion) by assuming that all market participants face, on an order-by-order basis, what we call *frictionless search and access*. More specifically, frictionless search means that trading firms, investors, and informed traders (upon arrival to the market) observe the current state of the order book on all exchanges at zero cost prior to taking any action. Frictionless access means 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, even if an order is split across multiple exchanges.

[37] We discuss asymmetric fee schedules, potentially including fees and rebates, below in Section 3.2.5.

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

Additionally, we assume that investors and informed traders, upon arrival to the market, can costlessly 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'.[39] 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.[40]

We believe that these assumptions are consistent with the technology and sophistication of modern trading, and capture the essential economics of Regulation National Market System. We emphasize that, in practice, investors and informed traders need not do the multi-exchange search and access themselves, but rather may rely on their broker-dealer's routing algorithms or an exchange's routing logic to perform this function on their behalf. We also note that our way of modeling speed and latency, with speed essentially playing the role of a tiebreaker, circumvents much of the complexity (and associated controversy) of Reg NMS in practice; this is an interesting and important debate but not central to the present paper.[41]

ESST Fair-Access Assumption. Last, we require that each exchange sell ESST to at least 2 trading firms, or not sell ESST at all. If only a single TF purchases ESST from a given exchange j (which we will show occurs only off-path in equilibrium), 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 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" (Clayton, 2018). For this reason, we also assume that the number of TFs endowed with general-purpose speed technology is at least $N \geq 3$.[42]

3.2.2 Timing

We now present the timing of our multiple-exchange game. There are three *stages* to this game, where the first two stages are played once. In Stage One, exchanges simultaneously set (or "post") trading and ESST fees; these are fixed for the entirety of our game. In Stage Two, trading firms simultaneously decide which exchanges to purchase ESST from. Finally, in Stage Three, a multi-exchange version of the trading game described in the previous subsection is repeated infinitely often. Parts of the timing from the trading game stage will mirror that described in the single-exchange case, but are presented here for completeness.

Formally,

[39] For example, an investor can send trades to both Nasdaq and BATS with their arrival times sufficiently synchronized that it is not possible for a high-frequency trading firm to observe the trade on their co-located server at Nasdaq, and transmit information about the trade to their co-located server at BATS fast enough that they can react on BATS before the investor's own order reached BATS in the first place (or vice versa switching BATS and Nasdaq). Mathematically, we capture this by assuming that an investor, upon arrival, or an informed trader, upon observing private information, has a chance to send messages to all exchanges before other market participants have an opportunity to respond.

[40] The geographical configuration of different exchanges' server farms in New Jersey places a lower bound on how quickly a high-frequency trading firm can react on exchange j' to an action on exchange j, on the order of 100-200 microseconds. Synchronization thus requires the ability to send messages to different exchanges such that they arrive within about 100 microseconds of each other, which we understand to be technologically straightforward and commodified now, but not in the early days of Reg NMS. 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*a*).

[41] Regarding the policy debate over Reg NMS, useful starting points are Tyc (2014) and Part III of Budish (2016*b*).

[42] 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.

1. Stage One *(Exchange Price Setting)*: All M exchanges simultaneously post 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 *(Multi-Exchange Trading Game)*: The following trading game, consisting of two periods, is repeated infinitely often. At the beginning of each trading game, the *state* is publicly observed and given by $(y, \boldsymbol{\omega})$, where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_M)$ represents the state of *all* exchanges' order books.

   (a) Period 1: Given publicly observed state $(y, \boldsymbol{\omega})$, each TF i may simultaneously send orders to all exchanges. An *order* for TF i sent to exchange j is a set of messages denoted by $o_{ij} \in \mathcal{O}$, where $\mathcal{O}$ is the set of all potential combinations of messages and, as before, may include standard limit orders, cancellations, and IOCs. Limit orders and IOCs sent to an exchange take the form (q_i, p_i), where such an order states 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. All messages are serially processed by each exchange, with speed serving as a tie-breaker. Orders submitted by TFs to each exchange may affect the state of that exchange's order book, ω_j.

   (b) Period 2: 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 send IOCs and cancellation messages to all exchanges, potentially affecting $\boldsymbol{\omega}$.

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

3.2.3 Equilibrium of the Multiple-Exchange Game

We adopt the following solution concept for our game:

Stages 1 and 2 (Exchange Price Setting and Speed Technology Adoption). In the first two stages of our game (which are played once), we rely on subgame perfect Nash equilibrium among exchanges and trading firms, given anticipated behavior in Stage 3.

Stage 3 (Multi-Exchange Trading Game). As before, we assume that all market participants employ pure-Markov strategies in each trading game, and cannot condition their actions on the entire history of play. Rather, strategies are allowed only to condition on: trading and ESST fees; the number of TFs that have purchased ESST from each exchange (determined by TF adoption decisions in Stage 2); and on the trading game's current state, $(y, \boldsymbol{\omega})$. For the Stage 3 multi-exchange trading game, we rely on our order book equilibrium solution concept (Definition A.1). With regards to period 2 of the multi-exchange trading game, if there are multiple exchanges offering liquidity to an investor at the same best price and trading fee, an investor will be indifferent over how he splits his order across those exchanges. We assume that investors break ties when indifferent over transacting on different exchanges by using what we refer to as (constant)

routing table strategies: i.e., investors choose a vector of fixed weights $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_M)$ such that if there are multiple exchanges $\mathcal{A}$ offering depth at the same (best) price net of trading fees, an investor submits $\gamma_j / \sum_{k \in \mathcal{A}} \gamma_k$ fraction of its demand (subject to availability) to each exchange $j \in \mathcal{A}$. It is without loss to assume that these weights sum to 1.[43] Thus, as with the single exchange model, it will be the case that investors, informed traders, and TFs (given the state) have essentially unique optimal period-2 strategies: investors purchase up to one unit across exchanges at the best price possible, but may break ties using some (potentially arbitrary) routing table strategy; informed traders trade wherever it is profitable to do so; and TFs, if a large enough publicly observed jump occurs, engage in a sniping race where they attempt to snipe stale quotes while liquidity providers attempt to cancel their existing orders.

Existence. We now state our main theoretical result.

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 $\boldsymbol{F}^* = (F_1^*, \ldots, F_M^*)$ such that the sum of ESST fees for exchanges with positive market shares, $\sum_{j:\sigma_j^*>0} F_j^*$, is lower than the upper bound 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^* amount 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.

The bound on ESST fees is:

$$\sum_{j:\sigma_j^*>0} F_j^* \leq \frac{\Pi^*_{continuous}}{N} - \max(0, \pi_N^{lone-wolf} - \min_j F_j^*), \tag{3.2}$$

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

The equilibria described in this Proposition have the following key properties (which we summarize here and discuss further in 3.2.4). First, all exchanges charge zero trading fees. Second, exchanges charge positive ESST fees that are bounded by (3.2) (which we will discuss in detail shortly), and all TFs purchase ESST from all exchanges with positive market shares. Last, in the multi-exchange version of our trading game, in each trading game exactly one unit of liquidity is provided at spread $s^*_{continuous}$ across all exchanges according to some vector of market shares $\boldsymbol{\sigma}^*$. That is, liquidity provision is economically the same as in the single-exchange case, but spread out across the M exchanges according to the vector $\boldsymbol{\sigma}^*$. Investors and informed traders also act essentially identically as in the single-exchange trading game, but, again, coordinated across the M exchanges also according to the vector $\boldsymbol{\sigma}^*$. Intuitively, market participants use frictionless search to

[43] Note that this allows investors to (essentially) employ lexicographic preferences over exchanges when submitting orders: e.g., if there are three exchanges with depth available at the best price, allowing $\boldsymbol{\gamma} = (1 - \varepsilon - \varepsilon^2, \varepsilon, \varepsilon^2)$ for $\varepsilon > 0$ sufficiently small approximates an investor consuming depth from exchange 1 before moving to exchange 2, and then consuming depth from exchange 2 before moving to exchange 3. If $\gamma_j = 0$ for all $j \in \mathcal{A}$, we assume that an investor splits his demand uniformly among exchanges contained in $\mathcal{A}$.

"synthesize" a single exchange from the M parallel exchanges, and then act economically the same way as before. 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 sum, the trading game outcome is economically the same as the single non-strategic exchange case analyzed in Section (3.1), with the main difference being that exchanges and TFs now split the rents generated from latency arbitrage activity.

The proof of our 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 involves a single unit of liquidity being 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, investors' routing table strategies $\boldsymbol{\gamma}^*$ coordinate the provision of liquidity across exchanges, resulting in each exchange's share of liquidity provided ("depth") matching the share transacted ("volume"). 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, on exchange j, exactly σ^*_j depth is provided and σ^*_j volume is transacted upon an investor's arrival in each trading game.

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 $\pi_N^{lone-wolf} \in (\frac{N-1}{N} \times \frac{\Pi^*_{continuous}}{N}, \frac{\Pi^*_{continuous}}{N})$ 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 in Stage 1, there is an equilibrium in which exchanges all charge zero trading fees, and levy any arbitrary vector of ESST fees that satisfy condition (3.2). We construct equilibrium strategies in which TFs only purchase ESST from exchanges with zero trading fees (which ensure that no exchange profitably can deviate and charge positive trading fees) and ESST fees no greater than $\boldsymbol{F}^*$.

3.2.4 Features of the Status Quo

We now discuss the main features of the equilibria described in Proposition 3.2. Later, in Section 4, we show that these features—in particular, that exchanges operate essentially as a single synthesized exchange, that trading fees are close to zero, and that exchanges earn significant profits from ESST fees—are consistent with patterns that we observe in the data.

Single "Synthesized" Exchange. Regulatory features of U.S. stock exchanges—in particular, Reg NMS and UTP, modeled here as frictionless search and access—support an environment where market participants can "stitch" together multiple exchanges into what is essentially a single "synthesized" exchange. This has the following implications, shared by all equilibria described in Proposition 3.2. 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: due to adverse selection, "excess" liquidity that will not be consumed by an investor should one arrive will not be offered on any exchange. Last, multiple exchanges are able to maintain positive market shares without the market tipping to any one exchange. In our model, even though exchanges are not differentiated—all active exchanges have the same trading fee, and all market participants face the same costs of trading on any exchange—there is a continuum of equilibria that differ with respect to the market shares of each exchange: indeed, as proven in Proposition 3.2, there exists an equilibrium that supports *any* arbitrary vector of market shares.

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., what we call 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 $R(q)$.[44] That is, multiple exchanges can coexist in equilibrium if their liquidity schedules add up to what a single exchange would have provided. 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.

The interior market shares result is also closely related to Ellison and Fudenberg (2003), who provide conditions under which a "plateau" of equilibria involving interior market shares can exist among platforms sharing the same seller-buyer ratio.[45] In our setting, all exchanges can maintain positive market shares in equilibrium if they share the same "depth-volume" ratio: a common ratio ensures that the marginal unit of liquidity on any exchange cannot be more profitably offered on (or demanded from) another. It is important for trading frictions to be zero—and not just close to zero—for this result. If instead investors incurred an arbitrarily small but strictly positive cost of splitting their orders across multiple exchanges relative to submitting orders to a single exchange, then any equilibrium of our trading game must be one in which *all* trading volume occurs on a single exchange.

In a sense, the specific location of liquidity and trading activity is unimportant for equilibrium outcomes;

[44] That is, for any quantity q, an investor who optimally splits his order between the two exchanges, say q_1 to exchange 1 and q_2 to exchange 2, pays a total price $R_1(q_1) + R_2(q_2)$ that equals $R(q)$.

[45] Ellison and Fudenberg (2003) study competition among platforms for single-homing buyers and sellers. They show that if competing platforms share the same seller-buyer ratio (analogous to our depth-volume ratio), then interior market shares among platforms can be sustained for an interval of market shares for which negative "market impact" effects from joining a platform offset any potential "scale" effects that otherwise favor agglomeration and market tipping. Outside of this interval, the only equilibria are those with complete tipping.

in any given trading game, all that matters is the aggregate quantity of liquidity provided, and spread at which it is offered. Thus, our model is not designed to predict equilibrium exchange market shares. Nevertheless, our model does provide an interpretation for how they might arise. In our equilibrium construction, investors break ties when indifferent across exchanges using routing table strategies. Such strategies, in turn, coordinate where TFs provide liquidity. This implies at least two things. 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 additional 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 routing table strategies are not restricted to be stationary, one may expect that they may be relatively stable over time if they represent real world algorithms, practices, or protocols that are costly to adjust. Thus, one may also expect realized market shares across exchanges for a given security to be relatively stable across time.[46]

Competitive Trading Fees. In the equilibria described in Proposition 3.2, trading fees are competitive and equal to zero on all exchanges. Since exchanges are undifferentiated, and investors face no search costs or frictions when splitting their orders across multiple exchanges, liquidity can always be provided more cheaply on exchanges with lower trading fees. Hence any exchange j, given that all other exchanges set zero trading fees, cannot levy positive trading fees and attract positive trading volume. This is true even if investors broke ties in j's favor (all else equal), and even if j charged lower ESST fees. 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."[47] 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

[46] 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.

[47] 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.

in any equilibria where TFs, whenever profitable to do so, only purchase ESST from exchanges with the lowest trading fees.[48]

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 upper bound on ESST fees given by (3.2), a condition that holds more generally across a range of equilibria (as we discuss below). However, Proposition 3.2 also establishes that there are equilibria where (3.2) does not bind, and exchanges charge lower ESST fees (including zero in total). To sustain these equilibria, TFs employ strategies that require them to coordinate with one another and not purchase ESST from any exchange that raises its ESST fee above some (arbitrary) threshold. Such coordination might not be reasonable to assume: for example, when (3.2) is not binding, 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 equilibria where an exchange could increase its ESST fee while still having all TFs purchase from it (i.e., all TFs purchasing ESST from the exchange with increased fees 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 exchange-specific speed technology (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.[49]

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*

[48] 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 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 it can be shown that 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 $\varepsilon > 0$.

[49] 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.

fees, $N \times \sum_{j \in \mathcal{M}} F_j^*$*, are strictly less than* $\frac{M}{(M-1)N} \Pi^*_{continuous}$*.*

Recall that any given TF, under the equilibria that we construct, expects to earn (gross ESST fees) $\Pi^*_{continuous}/N$. Even if it pays the maximum amount to exchanges for ESST that is consistent with condition (3.2) (i.e., each TF pays $\frac{M}{(M-1)N^2} \Pi^*_{continuous}$ in total), each TF maintains a substantial share of these rents that increases in the number of exchanges. In our empirical setting, there are 3 major exchange families; if there are 7 major trading firms (so that $M = 3$ and $N = 7$), the proposition implies that exchanges in total are able to extract at most only 3/14 of all latency arbitrage rents, with the remainder accruing to TFs. (This does not mean that exchanges and TFs keep all of these rents as economic profits; as discussed below, these rents may be dissipated further to other market participants or in the form of fixed costs.)

We emphasize that while this particular interior division of latency arbitrage rents is specific to our model, what will ultimately matter for the question posed by our paper—will alternative market designs that eliminate latency arbitrage be adopted by market participants?—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 among exchanges and TFs, as well as cooperative solution concepts for rent-splitting such as the Shapley value.[50] A strength of our approach is that it highlights that even if exchanges can post prices—which, in most 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 Inefficiency. As in BCS, in our model trading firms providing liquidity incorporate the cost of being sniped into the bid-ask spread they charge, leading to higher costs of liquidity provision. Though outside our model (as we have assumed that investor demand is inelastic), this nevertheless induces a potential inefficiency from foregone trades due to a wider spread and potentially thinner market. Furthermore, competition for latency arbitrage rents (split among TFs and exchanges in our model via ESST fees) likely leads to substantial investments by TFs (e.g., the latency-sensitive part of their human capital costs and technology costs), exchanges (e.g., the cost of co-location facilities, latency-sensitive matching engines), speed technology providers (e.g., specialized field programmable gate array (FPGA) chips, microwave links, latency-sensitive switches, etc.), and broker-dealers (e.g., technology and human capital devoted to building sophisticated routing algorithms). Such expenditures may also be exacerbated due to excess entry and standard business stealing effects (Mankiw and Whinston, 1986) and have opportunity costs (e.g., forgone innovation in other industries due to excess financial sector employment). Importantly, these costs would not be incurred under alternative market designs that eliminated the high-frequency trading arms race.

3.2.5 Asymmetric Trading Fees & Tick Size Constraints

We have abstracted away from more sophisticated trading fee structures, and assumed that exchanges charge a common per-share per-side trading fee to all market participants. In reality exchanges often charge different

[50] 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 (cf. Collard-Wexler, Gowrisankaran and Lee (2019)).

fees for "making" liquidity as opposed to "taking" liquidity.[51] However, this is without loss of generality in our model: as prices are continuous and security x can be traded at any price, and as investor demand is symmetric (equally likely to wish to buy or sell) and inelastic to prices and spreads, only the net trading fee—fees minus rebates—matters for determining equilibrium behavior. This stands in contrast to other two-sided market environments where participants cannot "unwind" the division of fees via side payments (cf. Rochet and Tirole (2006)); here, equilibrium spreads adjust to reflect underlying net trading fees.

If, instead, prices were discrete due to positive tick-sizes—the minimum price movement of any security—asymmetric fee structures could be used by exchanges to "fill in the ticks" and provide a reason for exchange fragmentation (cf. Chao, Yao and Ye (2019)). Nevertheless, as we discuss further in the next section, the main five active exchanges employ essentially the same fee structure (a take fee close to the regulatory limit, and a positive make rebate) with a net trading fee close to zero (consistent with the intuition from our model that exchanges essentially Bertrand compete on trading fees). We thus believe that asymmetric trading fees alone cannot rationalize observed interior market shares among these major exchanges.

4 Empirical Validation

In this section, we document a series of seven stylized facts regarding modern U.S. stock exchange competition, utilizing a combination of exchange-labeled trades-and-quotes data, exchange fee filings, and exchange financial filings. These facts are broadly consistent with our model of the status quo, and importantly, when taken in total, are *not* consistent with other models of exchange competition. These other models, most of which were not designed specifically to capture modern U.S. equities exchanges, include those in which participants single-home across exchanges, leading to network effects, supra-competitive trading fees, and often tipping; models in which exchanges are meaningfully differentiated, either horizontally or vertically; and models in which tick-size frictions are central.

Section 4.1 will present evidence that relates to the equilibrium characterization of the trading game (i.e., Stage 3). Section 4.2 will present evidence that relates to the equilibrium characterization of exchange trading fees (i.e., f). Section 4.3 will present evidence that relates to the equilibrium characterization of exchange-specific speed technology fees (i.e., F). Section 4.4 will provide discussion of the stylized facts taken in total, with reference both to our model of the status quo and to other models of financial exchange competition that are inconsistent with the empirical evidence.

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 3.2.4, that we will assess empirically. First, all active exchanges have the same equilibrium bid and ask, 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

[51] Some exchanges employ what is often called the *maker-taker* fee structure, where a provider of liquidity is paid a rebate while a taker of liquidity is charged a positive trading fee; other exchanges employ a *taker-maker* fee structure where the reverse is true.

are useful to document carefully both because they provide empirical evidence for our admittedly-stylized model of trading, and because they are *not* consistent with several 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 all exchanges and labels every trade and quote update by exchange. Two disadvantages of this dataset are (i) 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; and (ii) the timestamps, while of millisecond granularity, are less accurate than the timestamps provided in direct-feed data purchased directly from exchanges. Unfortunately, direct-feed data are not available for academic research from at least one major exchange family. Since our analysis does not rely on ultra-precise timestamps (unlike, e.g., a study of latency arbitrage per se), and since comprehensiveness across exchanges is critical for our purposes, TAQ data was the obvious choice.

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 takers of liquidity (i.e., the submitter of a limit order that trades against a resting bid or offer) are charged a fee, providers of liquidity (i.e., the resting bid or offer) are paid a rebate, and the exchange fee (i.e., f in our notation) is the net of the taker fee minus the maker 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 gets the rebate and the maker pays the fee (cf. Chao, Yao and Ye, 2019). These 3 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, and, anecdotally, industry participants regard them as vestiges of an earlier era of stock exchange competition. Our main analyses report results for both the "Top 8" and the "Top 5".

- Symbols. In 2015, there were 9175 stocks or exchange traded funds that traded at least once; however, most stocks and ETFs trade relatively infrequently.[52] 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 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 all symbols that satisfy these filters with at least 1 million shares of average daily trading volume.

[52] 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".

Stylized Fact #1: Many Exchanges Simultaneously at the Best Bid and Best Offer. The first feature of our Stage 3 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.[53] 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 (typically, listed on Nasdaq) the maximum is 7. Panel A presents results for the Top 8 exchanges, i.e., all exchanges with meaningful market share (dropping the regionals), and Panel B presents the results for the Top 5 exchanges, i.e., the main "maker-taker" exchanges.

As can be seen, the modal answer to the question "how many exchanges are at the best price?" is "all of them." Of the Top 8, in about 50% of milliseconds all exchanges are at the best bid (similarly, best offer), and it is rare that only one or a few exchanges are at the best price. If we look at just the Top 5 maker-taker exchanges, and ask what proportion of those exchanges are at the best price, in about 85% of milliseconds the answer is "all of them."[54]

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 8 exchanges, in about 50% of milliseconds all exchanges are at the best bid (similarly, best offer). Of the Top 5, in about 85% of milliseconds all exchanges are at the best bid (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.*

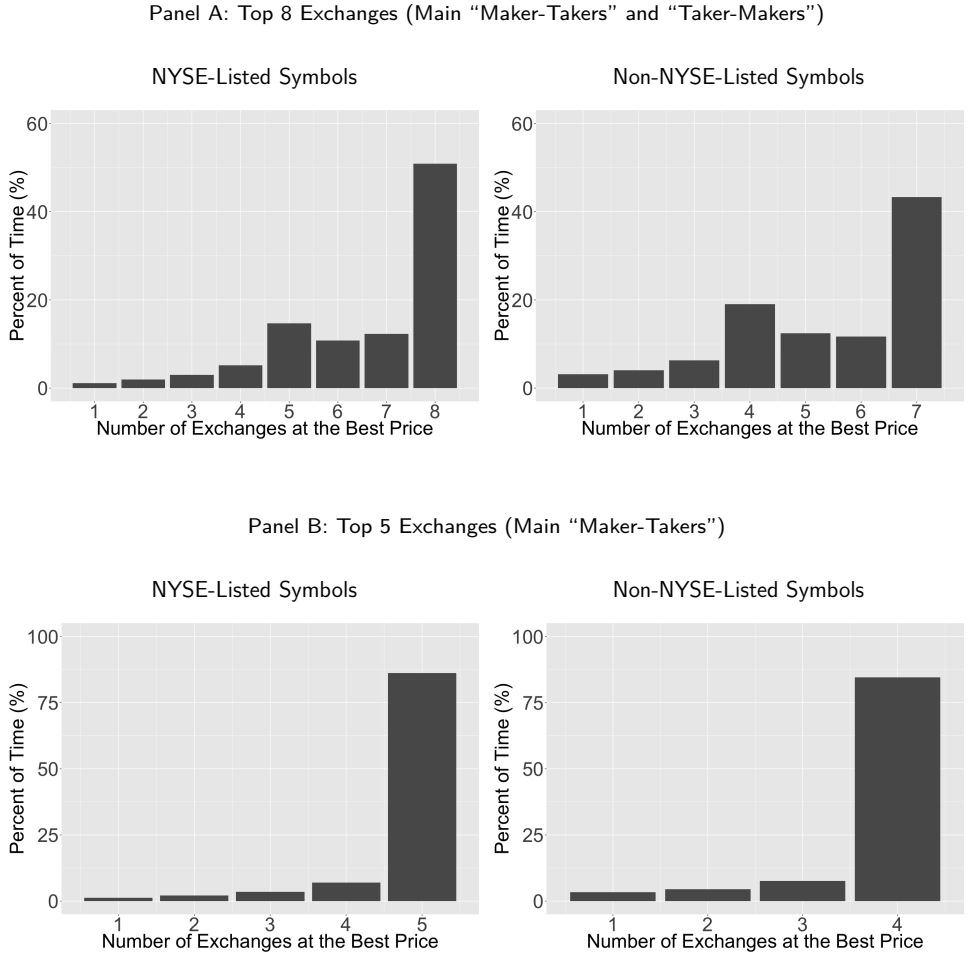
Stylized Fact #2: Linear Depth-Volume Relationship. A second feature of our Stage 3 equilibria is that "volume follows depth"—more precisely, while our analysis is silent as to what determines exchange j's share of displayed liquidity, it does state that whatever is exchange j's share of displayed liquidity 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

[53] NYSE recently (April 2018) changed this practice and began allowing non-NYSE listed stocks and ETFs to trade on NYSE.

[54] Note that if a trading firm provides liquidity at the same quoted price on a taker-maker exchange as on a maker-taker exchange, they in effect are offering a price that is roughly \$0.0045 better for the taker of liquidity and \$0.0045 worse for themselves as the provider of liquidity (see Table 4.1 in Section 4.2). 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 but not on the taker-maker exchanges. This same economic force manifests in a related way in our next stylized fact; see the text for discussion. Also see the next footnote for a worked example.

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 price was equal to N. Panel A considers Top 8 exchanges, Panel B considers Top 5 exchanges; for discussion of Top 8 and Top 5 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. The Top 5 panel omits the 0 bar for the rare case where none of the Top 5 exchanges are at the NBBO as defined by the Top 8; this occurs 0.02% of the time among NYSE listed symbols and 0.06% of the time among non-NYSE listed symbols.

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. 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 the 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 order to exchange j, if otherwise indifferent (cf. footnote 46)—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 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. Depth share we calculate by first computing depth for symbol i – exchange j at each millisecond k within the day, 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:

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

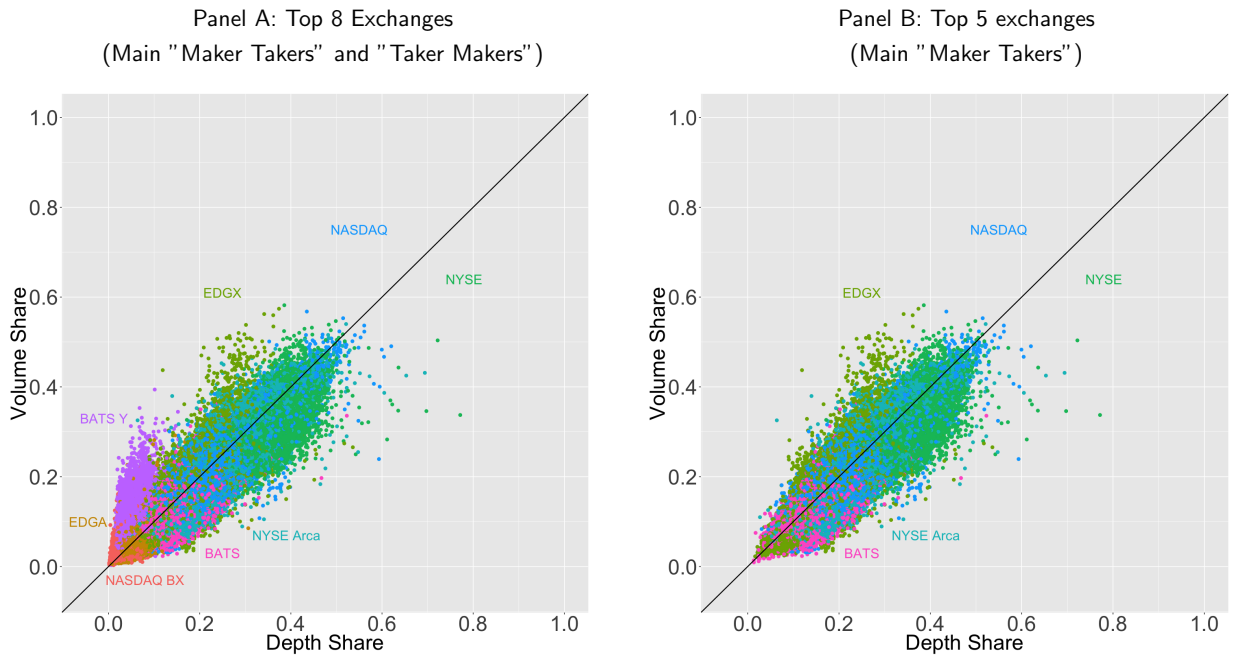
with T_{it} denoting the number of milliseconds for symbol i on date t between the symbol's first quote at or after 9:30 on that date and 16:00 (13:00 on half-days), dropping any milliseconds where the NBBO is locked or crossed. Figure 4.2 presents a scatterplot of $VolumeShare_{ijt}$ against $DepthShare_{ijt}$, 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 (cf. Stylized Fact #3), we color code by exchange and label each exchange's cluster of dots. Panel A presents results for the Top 8 exchanges, and Panel B presents results for the Top 5 exchanges.

As can be seen in the figure, the depth-volume data mostly clusters along the 45 degree line, as predicted, with the exception, at the bottom left of Panel A, of the taker-maker exchanges (EDGA, BATS Y, NASDAQ BX). The reason for this is that the taker-maker exchanges pay a rebate to the taker of liquidity, so, 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, the taker-maker exchanges have volume shares that are larger than their depth shares, i.e., their slope on Panel A is steeper than 45 degrees.[55] If we focus on just the maker-taker exchanges (Panel B), all 5 of which have essentially similar fee structures (cf. Table 4.1 and Stylized Fact #4 below), the relationship between volume and depth more cleanly tracks the 45 degree line. The slope of a regression of volume on depth is 0.991 (s.e. 0.020), and the R^2 of the relationship is 0.865.

In robustness tests, we found that this linear depth-volume relationship obtains at significantly higher

[55] 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. Then the net-of-fee price on the taker-maker exchange is about \$12.338 while the net-of-fee price on the maker-taker exchange is about \$12.343 – a difference of about half a penny. So the depth on the taker-maker exchange will get consumed before the depth on the maker-taker exchange, leading to a steeper volume-against-depth relationship on the taker-maker than the maker-taker.

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. Panel A considers Top 8 exchanges, Panel B considers Top 5 exchanges; for discussion of Top 8 and Top 5 see the text. 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.

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.[56] 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, using just the 5 maker-taker exchanges. 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.

We conclude:

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 of the regression line is 0.99 (statistically indistinguishable from 1) and the R^2 is 0.87. The relationship does not obtain for the taker-maker exchanges; they have comparatively little depth share, but what depth they do have gets disproportionately high volume share. The depth-volume relationship for the Top 5 does obtain at*

[56] For our sample of 100 symbols, focusing on just the maker-taker exchanges, the R^2 of the regression of volume-share on depth-share is 0.484 at 5 minutes, 0.594 at 10 minutes, 0.725 at 30 minutes, and 0.779 at 1 hour. The regression coefficients are 0.979, 0.982, 0.986 and 0.987 (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.

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

Stylized Fact #3: Exchange Market Shares are Interior and Relatively Stable, Both Aggregate and Within-Symbol. The third feature of the trading game equilibria that we explore is that market shares can be interior, i.e., there need not be market tipping. Furthermore, as discussed in Section 3.2.4, 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—but 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. We do this first at the aggregate level and then at the individual stock level.

Figure 4.3 presents exchange market shares since the start of the Reg NMS era in July 2007 (Panel A) and since Jan 1, 2011 (Panel B), since BATS and Direct Edge entered, with two exchanges each, in the period from late 2008 to mid 2010.[57] These exchange market shares are based on volume from all traded stocks and ETFs, though we note that the figure looks similar if we focus on the top 100 as we have done for other analyses. As can be seen in Panel B, aggregate exchange market shares are certainly interior and have been relatively stable since 2011. In this 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. The coefficient of variation of daily volume shares for the top 5 (maker-taker) exchanges, for this 2011-2015 period, ranges from .06 to .16. Exchange market shares are certainly not exactly constant over time, but the evidence suggests that they are relatively stable.

Figure 4.4 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 individual-stock 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. If we regress s_{ijt} (for the Top 8), 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.

Last, we again emphasize that both exchange and exchange-symbol market shares are consistently interior. This is *not* a market with tipping, where some exchanges control all or substantially all of trading for some stocks or ETFs. Among all symbols in our sample, the single highest exchange market share is about 40%.

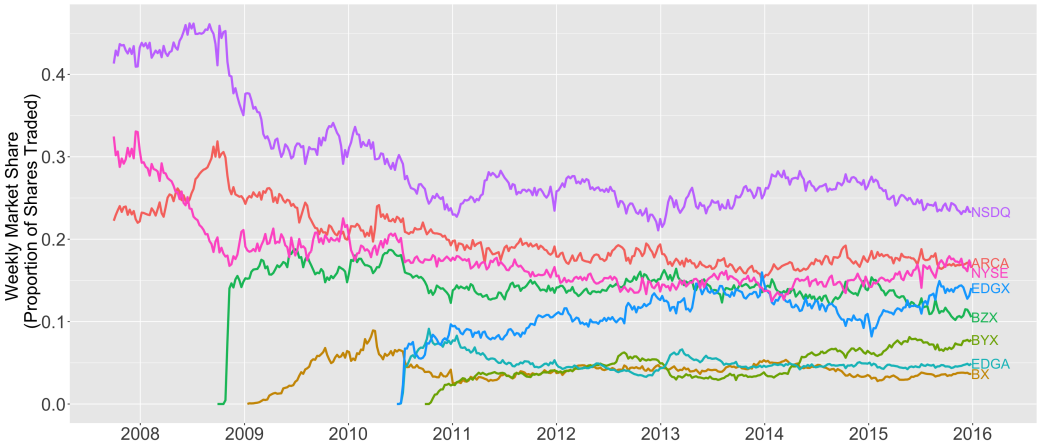
We conclude:

Stylized Fact 3. *Market shares are interior and relatively stable, at both the aggregate level and the individual-symbol level. While there is of course some variance over time and in the cross section, simple exchange fixed effects explain about 97% of the aggregate-level variation and about 76% of the individual-symbol-level variation.*

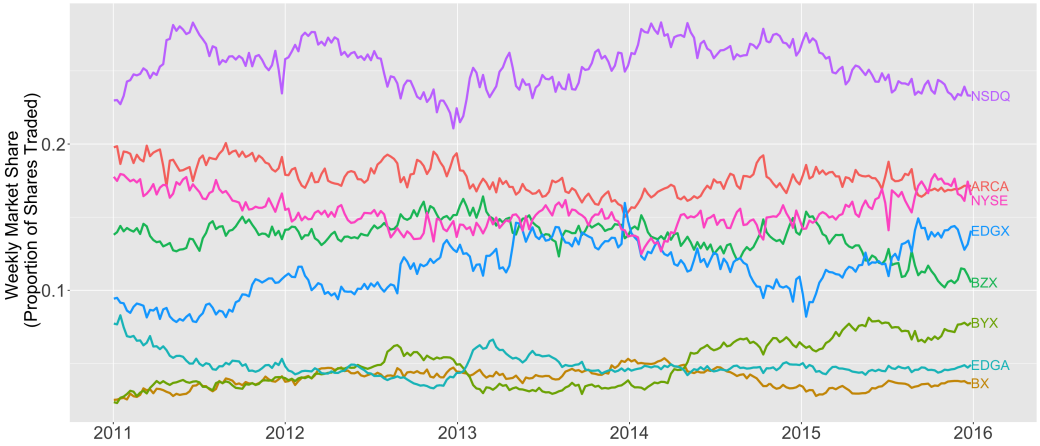
[57] Please note that before BATS and Direct Edge were approved exchanges they operated Alternative Trading Systems (ATS's, or dark pools). The large day-one market shares of these exchanges are therefore not as dramatic as they might appear, but rather likely represent trading volume on the ATS transitioning over to the newly-launched exchange.

Figure 4.3: Exchange Market Shares: Overall

Panel A: Reg NMS Era Weekly Market Shares



Panel B: 2011 - 2015 Weekly Market Shares



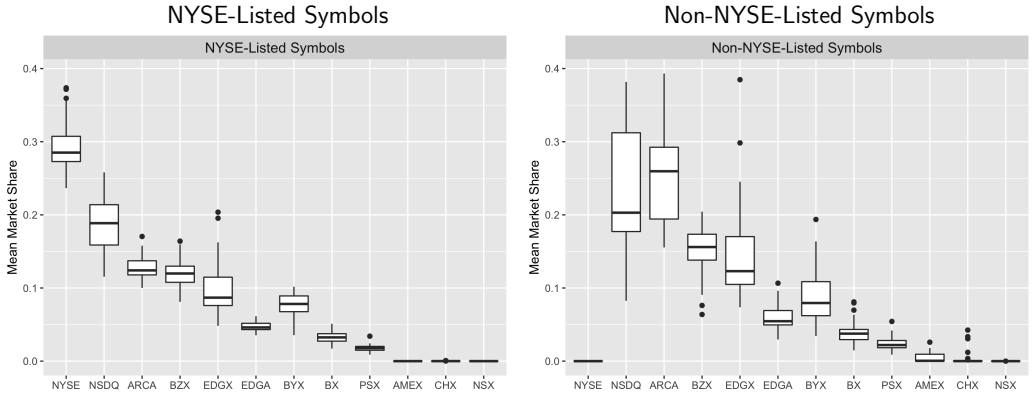
Notes: The data is from NYSE TAQ. The market shares are of on-exchange trading volume in shares. Panel A covers October 2007 through Dec 2015 and Panel B covers January 2011 to Dec 2015.

4.2 Evidence on Exchange Trading Fees (i.e., f)

Data Sources. We use three types of data sources for our analysis of trading fees. First, we use exchange fee schedules. 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. 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.

Second, we use exchange company financial filings, specifically the BATS April 2016 S-1 filing, Nasdaq's

Figure 4.4: 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.

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 filings are at the exchange level, most of the financial data in the annual report is 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).

Third, we use a variety of other institutional data sources to obtain estimates for aspects of our computation: we use a UBS ATS fee schedule for an estimate of off-exchange trading fees, we use a NYSE document called Order Type Usage for an estimate of opening/closing auction fees, and BATS documents for an estimate of routing fees and prevalence.

Stylized Fact #4: Average 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 pay a fee of \$0.0030, while makers of liquidity earn a rebate of between \$0.0020 – \$0.0032, depending on how much volume they trade on Bats BZX. BATS's net fee per-trade per-side therefore ranges 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". In addition to these fees for standard trading, there are also dozens of other fees for orders that are routed to other exchanges, executed in the opening or closing auctions on Bats BZX or elsewhere, etc. Both NYSE and Nasdaq have fee schedules that differ, at least slightly, based on whether the stock being traded is listed on NYSE or Nasdaq.

Table 4.1 presents the observed range for "f" for the top 8 exchanges. As can be seen, most of the exchanges have minimum fees that are actually slightly negative. On many of the exchanges the requirement to pay this negative fee is a pure volume threshold—these are the exchanges where the row marked "Regular"

in the program column has a negative minimum fee per share per-side—whereas on some of the exchanges, the fee only gets negative for traders who participate in special programs, like the NYSE's Supplemental Liquidity Provider or Designated Market Maker programs. The maximum observed fee per side is always strictly positive and is typically about 5 mills, though it is noticeably lower for the two BATS taker-maker exchanges (BYZ and EDGA) and higher (8 mills) for the Nasdaq taker-maker exchange.

To try to get a more precise estimate for f, we can use the major exchange families' annual financial filings. The advantage of this exercise is that we can estimate the average "f", not just the potential range. The disadvantages are (i) that we have to conduct this analysis at the level of the exchange family, not the individual exchange (as in Table 4.1), and (ii) we have to make some assumptions about fees from non-regular trading (e.g., auction trading, routed volume, etc.) to get to an estimate for per-share per-side regular-hours f. These disadvantages in mind, the results are presented as Table 4.2; supporting details are in Appendix B and an associated spreadsheet available in the online appendix. As can be seen, our estimate of 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 per year. 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. 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.[58] In other words, regular-hours trading revenues do not nearly cover exchange operating expenses. As another point of comparison, the annual revenue for StubHub, the largest secondary-market venue for concert and sports tickets, exceeds \$1 billion; that is, StubHub's revenue is over 5 times that for all U.S. regular-hours stock trading, despite the secondary market for event tickets being a tiny fraction of the secondary market for U.S. equities.[59]

We conclude:

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) is, on a per-share per-side basis, -\$0.00015 to +\$0.00075 for regular trading on the top 5 maker-taker exchanges. If we include the taker-maker exchanges as well, the observed range for regular trading is -\$0.00015 to +\$0.00080. The average per-share per-side fee paid, for regular hours on-exchange trading, 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 data and colocation. 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 f 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 slightly below zero, because of SEC Section 31 fees and, for firms that are FINRA members,

[58] 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.

[59] Whereas a trade of a \$100 share of stock yields an average total fee of about \$0.0002 to the exchange, the trade of a \$100 concert ticket yields an average total fee of about \$22 to StubHub, a difference of 100,000 times. (Other secondary market ticket venues' fees are similar in percentage terms). As a result, even though StubHub's annual dollar volume is about \$4.5 billion—less than 1/10,000th the dollar volume of U.S. regular hours equity trading—StubHub's annual revenue of just over \$1 billion is about 5 times that for U.S. regular hours equity trading. Source for StubHub numbers: eBay fiscal year 2017 10-K, pages 39-40.

Table 4.1: 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.

Table 4.2: Estimate of average trading fees ("f")
(3 major exchange families)

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

Notes: Please see Appendix B and the associated spreadsheet for supporting details for these calculations.

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 the dollar volume, the sum of SEC + 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.[60] 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, this would be 1.14 mills. 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 problem. 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 8 exchanges exhibited in* Table 4.1 *(Nasdaq, BATS BZX, EDGX, BATS BYX), the fee is negative for the highest regular volume tier, with the lowest observed fee being -\$0.00015 per share per side. For another 3 of the 8 exchanges exhibited in Table* 4.1 *(NYSE, NYSE Arca, Nasdaq BX), the fee is negative for traders with high-enough volume who satisfy additional requirements; the lowest observed such fee is -\$0.00040 per share per side. 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)

Data Sources. Our evidence on the magnitude of exchange-specific speed technology revenues comes primarily from exchange company financial filings (10-K's, S-1's and merger proxies). We describe the specific documents used as we go. We use a Consolidated Tape Association fee filing to get an estimate for the aggregate tape revenues, which we subtract for our main estimate of ESST revenues.

Stylized Fact #6: Exchanges Earn Significant Revenues from Data and Co-Location/Connectivity (i.e., ESST). 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

[60] 1.09 mills (5×0.218 mills) of SEC fees + 1.19 mills of FINRA fees.

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 can't be too large of a fraction of the total sniping prize (see Proposition 3.4).

We use exchange company financial filings to estimate ESST revenues in 2015. We focus on 2015 both for consistency with the analysis above and because, given BATS's acquisition of Direct Edge in 2014 and acquisition by CBOE in 2016, fiscal year 2015 is the first and last full year in which BATS's U.S. Equities business, inclusive of the Direct Edge acquisition, reports its accounts at the granularity we seek. At the bottom of this section we discuss the available evidence regarding growth of ESST since 2015.[61]

For BATS, our exercise is relatively straightforward. BATS's April 2016 S-1 filing (i.e., IPO prospectus) reports its U.S. Equities business as a separate reporting segment, and within this reporting segment it separately breaks out its market data business and its co-location and connectivity business. In 2015, its 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[62] were \$81.0M and its operating expenses were \$110.2M. This means that revenue from market data and co-location/connectivity were more than twice the revenues from trading fees, and, moreover, 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 without (loss of \$29.2M). A pie-chart of BATS' 2015 U.S. Equities revenue breakdown is given as Figure 4.5. For comparison, we also include the analogous pie-chart for the Chicago Mercantile Exchange, a large futures exchange that, because futures contracts are proprietary to the exchange (unlike equities which are fungible), is able to earn significant revenues from trading fees. For CME, 84% of all revenues come from trading and clearing fees, and CME would be significantly profitable on the basis of this revenue stream alone.[63]

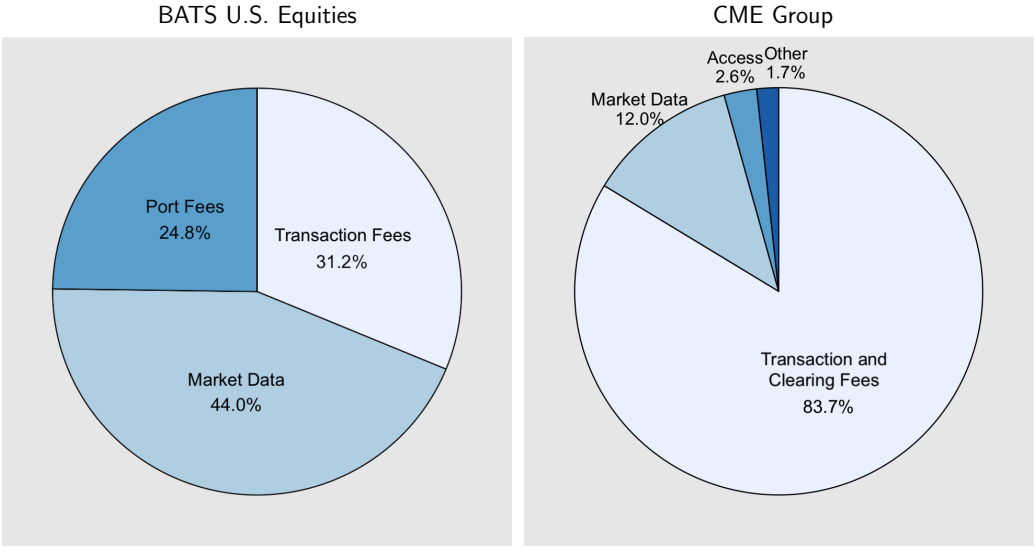
For both Nasdaq and NYSE our exercise is less straightforward because neither firm breaks out its U.S. equities business as its own reporting segment. For NYSE a further complication is its Nov 2013 acquisition by Intercontinental Exchange (ICE). Our approach for Nasdaq utilizes market data and connectivity 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 C. Our approach yields a range for Nasdaq's U.S. equities revenue of \$222.4M-\$267.3M for

[61] 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. See the next footnote for a specific example of the difference.

[62] 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). Note that the methodology in the Jones (2018) paper referenced in the previous footnote would treat BATS's US equity transactions revenues as the full \$938.8M in Transaction Fees, not the \$81.0M in Transaction Fees less Liquidity Payments and Routing and Clearing Fees. This interpretation leads to a very different economic picture from the one depicted in Figure 4.5.

[63] CME's 2015 trading and clearing revenue was \$2,784M and operating expenses were \$1,338M, for trading/clearing revenue less operating expenses of \$1,446M. Market data revenue was \$399.4M, Access & Communication Fees revenue was \$86.1, and Other revenue was \$57.4M. Note that a difference between CME and BATS (and between futures exchanges and equity exchanges more broadly) is that CME does its own clearing, and charges for this, whereas BATS and the other equity exchanges do not. For CME, we think of trading fees and clearing fees as two ways that CME earns a fee on a per-contract-traded basis, i.e., both are part of what our model calls "f".

Figure 4.5: Revenue Breakdown for BATS and CME in 2015



Notes: BATS U.S. equities data from BATS Inc. 2016 S-1 filing, page 86. BATS Transaction Fees are net of Liquidity Payments (i.e., rebates) and Routing and Clearing costs. CME data from CME Group Inc. fiscal year 2015 10-K filing, page 37 . The unabbreviated revenue categories as reported in the 10-K are "Clearing and Transaction Fees", "Market Data and Information Services", "Access and Communication Fees", and "Other". See footnote 63 regarding the difference between futures and equities exchanges with regards to clearing fees. For further details see the text.

market data, \$121.0M-\$139.0M for co-location/connectivity, and \$343.3M-\$406.4M combined. For NYSE, we obtain a range of \$218.9-\$241.5M for U.S. equities market data, \$251.6-\$281.5M for U.S. equities co-location/connectivity, and \$470.5-\$523.0M combined.

Across all three major exchange families, then, our 2015 U.S. equities estimate is \$555.4-\$623.0M for market data, \$436.8-\$484.8M for co-location/connectivity, and \$992.2-1107.8M total. These market data revenue figures include revenue from proprietary data feeds as well as from market-wide "Tape Plans", sometimes known as consortium data products or the SIP feed (cf. footnote 23). Proprietary data feeds are utilized by latency-sensitive market participants, whereas the market-wide consortium data feeds are not as fast, and therefore should likely 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 \$315M.[64] If we subtract this \$315M from the total we have proprietary market-data revenue of \$238.4-306.0M, and total ESST revenue of \$675.2-790.8M. Table 4.3 summarizes.

For context, note that our estimate for ESST revenue is several times larger than the revenue from regular-hours trading fees. If we take the lone-wolf bound from the theory seriously, and use $N = 3$ exchange families and $M = 7$ large high-frequency traders,[65] our estimated range for ESST revenues yields a lower bound on the total size of the latency-arbitrage pie of between \$3.1-\$3.7B in 2015.

[64]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 dramatically 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.

[65]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 high-frequency trading firms in the "lead lap" of the speed race in the U.S. market.

Table 4.3: 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 C. 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).

2015 is the last year in which BATS reports segment data for its U.S. equities business, and as described, 2014 is the last year for which granular NYSE data is available. We can get a rough sense of growth since 2015 by looking at growth in overall market data and co-location/connectivity revenue for BATS and Nasdaq. For BATS, growth from 2015-2017 was 8.5% per year for market data revenues and 11.7% for co-location/connectivity. For Nasdaq, growth from 2015-2017 was 6.7% for market data and 10.3% for co-location/connectivity. If we use 9% as a rough midpoint of this range, this implies 2018 ESST revenues are about 30% higher than 2015, for a range of $874M-1024M in 2018.

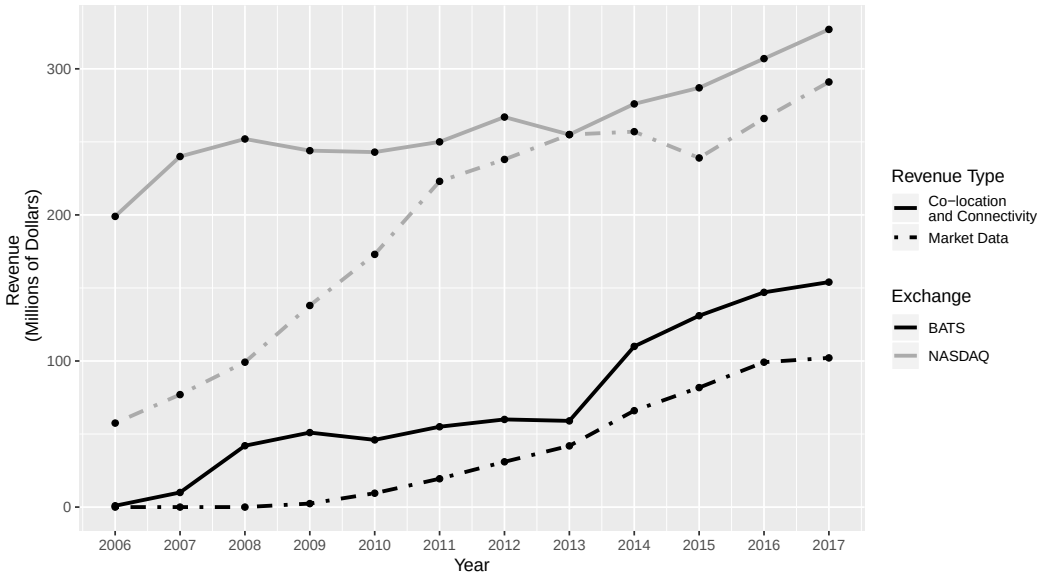
Stylized Fact 6. *Exchanges earn significant revenue from exchange-specific speed technology. While the data reported in exchange parent company financial filings are not perfect, sensible assumptions applied to that data suggest that in 2015 total ESST revenue was between $675-790M. This is several times larger than regular-hours trading revenues. For the BATS exchange, for which the data is cleanest, the U.S. Equities business is profitable inclusive of market data and co-location/connectivity (profits before tax of $149M) but loss-making without market data and co-location/connectivity (loss of $29M). Growth in ESST revenue since 2015 puts the upper bound of our estimate at about $1bn for 2018.*

Stylized Fact #7: Exchange Revenue from Data and Co-Location/Connectivity has Grown Significantly in the Reg NMS Era. While exchanges do not directly report U.S. Equities ESST revenue (as evident from all of 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 the revenue growth for the financial reporting categories that contain U.S. equities ESST. We do this for Nasdaq and BATS (see Figure 4.6), starting in 2006: 2006 was both the first year of BATS's operation and the first year the word "co-location" appears in a Nasdaq annual financial filing (it has appeared every year since).[66]

For Nasdaq co-location and connectivity, this category was called "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 years 2013-2014

[66] NYSE's financial reporting segments changed too frequently during the post Reg NMS period for the exercise to be useful.

Figure 4.6: Exchange Market Data and Co-Location Revenue 2006 - 2017



Notes: BATS data come from the 2012 S-1 filing (years 2007-2011), the 2016 S-1 filing (years 2012-2015), the CBOE/BATS Merger Proxy (Cboe Holdings, Inc. and BATS Global Markets, Inc., 2016) for 2016 and the CBOE 2017 10-K. Nasdaq data come from 2006-2017 10-K filings. For the specific reporting categories used in the figure please see the descriptions in the text. For Nasdaq Market Data, 2006-2012 represents the sum of U.S. Tape Plans revenue and U.S. Market Data Products revenue; 2013 is U.S. Market Data Products revenue which starting in 2013 includes both Tape Plans revenue and proprietary data products; and 2014-2017 is global "Data Products" revenue multiplied by a factor of 0.72 which is the 2013 ratio of U.S. Data revenue to Global Data revenue. For BATS's 2017 Co-location and Connectivity, we use the sum of BATS's contribution to CBOE Access Fees plus BATS's contribution to CBOE "Exchange Services and Other fees" (see pg. 68), each annualized by multiplying by a factor of 12/10 since BATS contributed to CBOE revenues for 10 months that year (March-Dec). For BATS's 2017 Market Data, we use BATS's contribution to CBOE "Market Data Fees" multiplied by an annualization factor of 12/10.

versus the years before and after, so we view the periods 2006-2012 and 2015-2017 as yielding reliable apples-to-apples growth rates. The growth rate in the period 2006-2012 was 26.7%, and in 2015-2017 it has been 10.3%. Revenues look flat in the period 2012-2015, but it is hard to be conclusive since there were some modest segment reporting changes.

For Nasdaq market data, in 2006-2012 they separately broke out revenue from U.S. Tape Plans from revenue for U.S. market data products – notably, in 2006, Tape revenue was nearly double Proprietary revenue ($129M Tape vs. $69.6M Proprietary) whereas by 2012 Proprietary revenue was about 30% higher than Tape ($150M Proprietary vs. $117M Tape). Starting in 2013 Nasdaq reported only combined data revenue and made some other segment reporting changes. Starting in 2014, they reported only global data products. We use the 2013 U.S. to Global ratio of 72% on the 2014-2017 numbers to get an apples-to-apples growth rate. Revenue growth in 2006-2012 was 5.1% annually (though Proprietary data products revenue grew by 13.7% annually, versus modest reduction in Tape revenue). Revenue growth in 2013-2017 has been 6.4% annually.

For BATS co-location and connectivity, the 2012 S-1 filing[67] reports that they began charging for co-location/connectivity in Q4 2009. Initially the revenue was reported in a segment called "Other Revenues,"[68]

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

[68]Described as "Other revenues consist of port fees, which represent fees paid for connectivity to our markets, and, more

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 growth was 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.

For BATS market data, these revenues are consistently reported as "Market Data Fees" throughout the 2006-2017 period; the only difficulty in interpreting the time series is the Direct Edge acquisition which contributed to a doubling of data revenues from 2013 to 2015. In the period 2006-2013, growth was dramatic from 2006 to 2008, and then modest from 2008 to 2013. Since 2015, growth has been about 8.5% per year.

Focusing especially on the co-location/connectivity business, the data are suggestive of the exchanges "discovering a new pot of gold" in the period from 2006 through 2012/2013 or so—that is, discovering that they could charge significant money for something they used to not charge for—and then growth leveling off in the years since. Nasdaq's access services revenues increased nearly four-fold during the 2006-2012 period, and BATS's increased four-fold from 2010-2013.

Stylized Fact 7. *Exchanges' revenues from exchange-specific speed technology appear to have grown significantly in the Reg NMS era. Particularly dramatic is the growth in co-location/connectivity revenues in the period from 2006 to 2012/2013 or so. Nasdaq's co-location/connectivity revenue quadrupled in this period, and BATS's quadrupled from 2010-2013. Since 2015, data revenues have grown about 7.5% per year (average of Nasdaq and BATS's growth rates) and co-location/connectivity revenues have grown about 11% 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. The fit is not perfect of course, but overall we hope the reader comes away from the empirical evidence believing that the model is credible. Specifically, that the model is sufficiently credible to consider, in the following section, what it implies for our motivating question of whether the market will fix the market.

We now discuss other potential models of exchange competition that are at odds with important aspects of the data and regulatory environment. The first class of models are those in which some market participants "single home", thereby generating exchange-specific network effects. Such models include the classic 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)—and the modern example of Pagnotta and Philippon (2018).[69] Relatedly, Cantillon and Yin (2008) consider a model in which participants can multi-home

recently, additional value-added products revenues [likely, co-location]."

[69] Pagnotta and Philippon (2018) 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, and that we document in Stylized Facts #6-#7, is speed that enables some market participants to be faster than other market participants on the same

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 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—for example, in Pagano (1989) tipping is the only equilibrium if transactions fees are the same across exchanges, and Cantillon and Yin (2008) is motivated by a famous empirical example of market tipping, the "Battle of the Bund"—which stands in contrast to Stylized Facts #1-#3. Last, these models are directly at odds with important 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 consider 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 (2018*b*), in which exchanges are horizontally differentiated.[70] Such differentiation allows exchanges to charge supra-competitive trading fees, which is inconsistent with Stylized Facts #4-#5. Also, such models suggest segmentation across venues, both with respect to which market participants choose which exchanges, and which securities trade where. Stylized Facts #1-#3 seem difficult to square with such segmentation. Such segmentation is also at odds with 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—the key point is that exchanges can use differential fee structures to enable trading firms to provide liquidity at slightly different net-of-fee prices across exchanges, which both "fills in the penny" for the market (i.e., makes tick-size constraints less binding) and gives exchanges market power. However, this model is inconsistent with the fact that the Top 5 exchanges, which control 83% of volume, all use essentially the same fee structure (cf. 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.[71] That said, while inconsistent with several aspects of our data, this model does seem first-order 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 does still seem consistent with many aspects of modern futures exchanges. The crucial difference between futures and equities, as emphasized earlier, is that

exchange. See also Cespa and Vives (2019) who model speed in a similar fashion to Pagnotta and Philippon (2018) by allowing for 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 period. 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.

[70] Baldauf and Mollner (2018*b*) consider a model, which they fit to data from the Australian stock market, in which exchanges are located on a Salop circle (to capture horizontal differentiation) and, interestingly, the size of the latency arbitrage pie increases with the number of exchanges. 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. 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 the model of Baldauf and Mollner (2018*b*) as well as some empirical evidence cited therein from non-Reg NMS settings, often in cases where the number of exchanges went from one to multiple, such as Foucault and Menkveld (2008)); 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*a*)). Unfortunately, our empirical evidence is not able to speak to this issue. If Baldauf and Mollner (2018*a*,*b*) are correct that the latency-arbitrage pie grows with the number of exchanges, that should only strengthen the arguments we make in Section 5.

[71] 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.

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 (cf., Figure 4.5 and the surrounding discussion). 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 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: all exchanges maintain positive market shares (i.e., interior as opposed to tipping), exchange trading fees are competitive, and exchanges obtain 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 of organizing the data. In this section we now use the model to examine the private and social incentives for market design innovation.

For our discussion, we focus on a particular alternative to the continuous limit order book design: *discrete-time frequent batch auctions* (FBAs).[72] As proposed and examined in BCS, an exchange using FBAs differs from one using the continuous limit order book in the way that it processes and prioritizes orders. Just as in the continuous limit order book market, (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. However, unlike the continuous market, orders are processed in discrete-time "batches": specifically, at the end of each pre-specified time interval (e.g., one millisecond), the FBA exchange conducts a uniform price double auction among all new orders that have arrived during that interval and those that remain in the order book from previous intervals. In the case that there are multiple orders at the same price, priority in the FBA is price-then-time, like in the continuous market, but treating time as discrete: that is, orders that are outstanding from earlier intervals have higher priority than orders at the same price that arrive at later intervals, but orders that arrive within the same interval have the same priority, with random tie-breaking if necessary. Information policy is also the discrete-time analogue of the continuous limit order book: the same information is disseminated as in the continuous market (e.g., new orders, cancels, trades, etc.) but with this information disseminated in discrete time, at the conclusion of each interval.

BCS analyze equilibrium behavior on a single FBA exchange in isolation, and show that by processing orders in batches—as opposed to serially—FBAs eliminate latency arbitrage, and in so doing enhance liquidity provision and stop the latency arms race. Intuitively, in the event of symmetric public information, discrete-time batch auctions transform competition on speed into competition on price, which eliminates the rents.

Our Approach. Despite the theoretical benefits of FBAs, it is not obvious that an FBA-exchange competing head-to-head with continuous limit order book exchanges would necessarily gain traction and obtain

[72] Our subsequent analysis is not specific to frequent batch auctions, per se. Rather, our results regarding incentives for market design innovation apply for *any* design that eliminates latency arbitrage rents. One specific alternative is a market design in which the exchange processes cancellation orders immediately upon receipt, but processes aggressive orders with some delay. This approach, sometimes called "asymmetric delay" or "asymmetric speed bump", also reduces the ability of trading firms to engage in latency arbitrage activity, though it has some weaknesses relative to FBAs. Please see Section VIII.C-D of Budish, Cramton and Shim (2015) for a discussion of asymmetric speed bumps and some potential weaknesses relative to FBAs, and please see Baldauf and Mollner (2018*a*) for a detailed theoretical analysis of asymmetric speed bumps. See also footnotes 3 and 21 for more details regarding the Chicago Stock Exchange's unsuccessful effort to adopt an asymmetric speed bump.

significant trading volume. Nor is it clear that any exchange has an incentive to incur the cost and effort of adopting a new market design.

To inform these issues, we employ the following approach. First, we build on the theoretical model of exchange competition and trading behavior developed in Section 3 to examine competition among multiple exchanges that each use one of two potential market designs: either a continuous-time limit order book (which we refer to as "Continuous"), or FBAs ("Discrete"). BCS, in their original analysis, only examined trading behavior on either a single Continuous exchange or on a single Discrete exchange. Here, we examine outcomes when one or more Continuous exchanges compete with a single Discrete exchange, and when multiple Discrete exchanges compete with one another. Combined with earlier results derived when multiple Continuous exchanges compete with one another (Section 3), this exercise "fills out" the cells of a matrix corresponding to the payoffs that exchanges obtain under different combinations of market designs. We then use this matrix, alongside parameters that capture the costs of adopting or imitating an alternative market design, to examine the incentives for a new or existing exchange to adopt Discrete.

Our key result is that private incentives to adopt a market design that eliminates latency arbitrage are dramatically lower than social incentives, and potentially even negative. We establish this claim without explicitly taking a stand on the nature or timing of adoption and entry decisions. If one exchange is the first to adopt Discrete while all its rivals employ Continuous, that exchange will capture a large share of trading volume and large economic rents. Intuitively, the innovator gets compensated for eliminating the latency arbitrage tax. However, any subsequent adoption by other exchanges of Discrete—which we argue is likely and potentially quite rapid—leads to the dissipation of industry rents, potentially rendering the initial adoption of Discrete unprofitable. Existing incumbent exchanges have even weaker incentives to adopt, since the source of their status-quo rents is precisely the inefficiencies that FBAs and alternative designs seek to eliminate. Formally, the market design adoption game among incumbent exchanges constitutes a *repeated prisoners' 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.

5.1 Competing Market Designs

To analyze exchange competition with potentially different market designs, we maintain the following timing conventions and assumptions from our multiple competing exchange model introduced in Section 3.2: given the market designs that each exchange operates, in Stage 1, exchanges post trading fees and (if relevant, in a manner made more explicit below) ESST fees; next in Stage 2, the N "fast" trading firms (TFs) with access to general-purpose speed technology choose which exchanges (if any) to purchase ESST from; and finally, in Stage 3, a repeated multi-exchange version of the trading game occurs.

5.1.1 Modeling Frequent Batch Auctions

In the Stage 3 trading game, at the end of each period, the frequent batch auction exchange ("Discrete") first processes all cancellation orders received in that period (reflecting that in an FBA orders can be canceled at any moment during the batch interval), and then aggregates all outstanding orders to buy and sell—both new orders submitted in that period and orders that remain outstanding from previous periods—into demand and supply curves, respectively. If demand and supply cross, then trades are executed at the market-clearing

price;[73] in the event that it is 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. All orders that are not canceled or executed in that period remain in the order book for the next period, and, as such, are part of the publicly observable state. The Continuous exchange is modeled as in Section 3; the key difference versus Discrete, as emphasized, is that if multiple messages are received in the same period, they are processed serially.

We make the following assumptions about the FBA batch interval and the way that Discrete and Continuous operate in parallel. First, we assume that the batch interval is sufficiently short so that investors do not care per se about waiting the small amount of additional time it takes to trade on Discrete rather than Continuous (i.e., they care about the price they pay, but not per se about delay), while long enough to enable "genuine" batch processing in the event that there is public news and multiple TFs respond. That is, if multiple TFs with essentially the same technology respond at essentially the same time to the same piece of public news, their responses can be processed by the exchange in the same batch interval. How long of a batch interval is long enough to meaningfully batch process is an engineering question as opposed to an economic one, but to give a sense of magnitudes, some industry participants have suggested to us that as little as 50 microseconds may suffice (i.e., 0.000050 seconds), and our sense from aggregating many similar conversations is that 0.5-1.0 milliseconds (i.e., 0.0005-0.0010 seconds) would be comfortably more than sufficient to meaningfully batch process. Second, and relatedly, we assume that an informed trader can profitably trade on both Discrete and Continuous before their information is revealed or inferable, and similarly that an investor can trade on both Discrete and Continuous before TFs can respond. This can be interpreted as allowing the informed trader and the investor to "synchronize" their orders across exchanges, as is now commonplace through broker-dealer routing algorithms.[74] Third, we assume that Discrete does not sell ESST; practically, we have in mind that a Discrete exchange would allow market participants to co-locate their servers, but would not be able to charge prices commensurate with their role, on Continuous exchanges, in extracting sniping rents.[75]

Last, we restrict attention to order book equilibria (as defined in Section 3) that are also robust to profitable deviations by the continuum of "slow" trading firms. This restriction has no effect on our previous results: with only Continuous exchanges, slow trading firms did not play any role in equilibrium.[76] However, on Discrete, slow trading firms can provide liquidity without being sniped, so this assumption captures the idea that such firms may now act as a competitive fringe of liquidity providers on Discrete. 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 (i.e., "fast" by non-high-frequency trading standards).

[73] 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.

[74] For this synchronization to be technologically feasible requires that the batch interval is long relative to the randomness in order transmission times. Again, this is an engineering quantity as opposed to an economic one, but our sense is that the randomness is comfortably less than 100 microseconds (0.0001 seconds) for sophisticated market participants. Note as well that end investors need not be able to do this synchronization themselves, rather they would do so through their broker-dealer.

[75] 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.

[76] Following the arrival of a publicly observed jump, no slow firm would ever win the sniping race, and any slow trading firm providing liquidity would be sniped for certain (as opposed to with probability $(N-1)/N$ in the case for fast TFs). Hence, slow trading firms would not be able to profitably provide liquidity at or below the equilibrium spread given by (3.1).

5.1.2 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).[77] Recall that if there was only a single Continuous exchange in operation charging zero trading fees (see Section 3.1), 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 any equilibrium at a spread given by $s^*_{discrete} < s^*_{continuous}$, which solves:

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

Why? Liquidity providers need only worry about adverse selection costs ($\lambda_{private}\cdot L(s^*_{discrete})$) when setting their spreads—Discrete eliminates the ability for trading firms to engage in stale quote sniping. This is for two reasons. First, any cancellations by liquidity providers—as long as they are received in the same batch period—are processed immediately whereas liquidity taking orders (and other new orders) are processed in batch at the end of the interval. Second, even if an outstanding liquidity providing order that is mis-priced following the arrival of public news is not cancelled, Discrete protects the liquidity provider from incurring losses: the price that is ultimately paid will, in equilibrium, be bid up or down to a price that reflects the new public information due to competition among all trading firms in the uniform price auction.

Now consider the multi-exchange trading game between Continuous and Discrete. Suppose initially that trading fees on both exchanges were zero, and all TFs 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 any Stage 3 trading game with a single Continuous and single Discrete exchange, where all trading firms have purchased exchange-specific speed technology from Continuous, and trading fees on both exchanges are zero. Then in any equilibrium of the trading game given state (y,ω), 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 elsewhere. Such an equilibrium of the trading game exists.*

To understand why liquidity cannot be offered on Continuous 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 $\lambda_{invest}\frac{\bar{s}_{continuous}}{2} - (\frac{N-1}{N}\lambda_{public}+\lambda_{private})\cdot L(\bar{s}_{continuous}) = 0$.[78] 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

[77] As discussed in Section 3.2, 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.

[78] 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 $f > 0$.[79]

Given these results, we establish the following:

Proposition 5.2. *In any equilibrium of the full multi-exchange game (i.e., Stages 1-3) among a single Continuous exchange and single Discrete exchange, (i) Discrete charges strictly positive trading fees; (ii) in every iteration of the trading game given state (y, ω), 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.*

When a single Discrete exchange competes against Continuous, it completely "tips" the market. Furthermore, Discrete earns economic profits that exceed $\frac{N-1}{N}$ proportion of the speed-race pie, $\Pi^*_{continuous}$, via trading fees. 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, 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.[80]

Caveat: Tick-Sizes and Sniping Costs. In our model, with continuous prices (i.e., no tick-size constraints), a single Discrete exchange will be able to completely tip the market away from Continuous exchanges and capture all trading volume. However, due to tick-size constraints (also discussed in Section 3.2.5), we do not think complete tipping is a realistic prediction. Specifically, in the empirically likely case that the elimination of latency arbitrage results in a per-share savings of less than the penny tick size (i.e., $|s^*_{continuous} - s^*_{discrete}| < \$0.01$),[81] then there may exist equilibria whereby positive volume remains on Continuous exchanges, because, intuitively, liquidity sometimes will not be able to be profitably provided on Discrete at a strictly better tick than on Continuous.[82] This suggests that a relevant consideration for the extent to which Discrete can "win" against Continuous is the size of sniping costs per share relative to the

[79] Let $\bar{s}_{discrete}(f)$ (defined formally in the Appendix in (A.6)) 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.

[80] 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 (cf. Kyle, Obizhaeva and Wang (2018)). That said, as we emphasize in the conclusion, incorporating investors with multi-unit trading needs, 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.

[81] For instance, in BCS, the annual latency arbitrage profits in SPY from ES-SPY arbitrage were about \$75M per year, against annual regular-hours on-exchange trading volume of just under 30B shares in the time period of the BCS data, for profits per share of about \$0.0026.

[82] Similar to observed behavior among maker-taker and taker-market exchanges (cf. Chao, Yao and Ye (2019)), we conjecture that with positive tick-sizes, there may be an alternative equilibrium with a single Discrete and multiple Continuous exchanges whereby volume can fluctuate across trading games between exchanges depending on the true value of the security, y, and its proximity to the nearest tick. A formal analysis is outside the scope of this paper.

tick-size.[83]

5.1.3 Multiple Discrete Exchanges

Now we consider the case of multiple Discrete exchanges. With at least two Discrete exchanges (and potentially another Continuous exchange), the resulting equilibrium has similar features to the one derived in Proposition 3.2 with multiple Continuous exchanges:

Proposition 5.3. *In any equilibrium of the 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 given state (y, ω), 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.*

As before, because features of the regulatory environment (particularly Reg NMS and UTP) imply that costs of searching and splitting orders across exchanges are zero, in equilibrium multiple Discrete exchanges also operate as a single synthesized exchange: a single unit of liquidity is always provided in each trading game, and equilibria differ from one another only in exchange market shares. Importantly, no trading volume occurs on any Continuous exchange, and competition among multiple Discrete exchanges operating the same market design leads to zero trading fees in equilibrium. However, there is a key difference. As opposed to the case with multiple Continuous exchanges, there is no longer latency arbitrage and the associated rents for exchanges.

Caveat: Additional Revenue Sources Beyond f and F. In reality, exchanges earn revenues from sources beyond what we have modeled, i.e., beyond trading fees f and exchange-specific speed technology fees F. One important source of such revenue is non-latency sensitive data revenues, sometimes called Tape or SIP revenues, as discussed in the previous section. A second source of such revenue is exchange listing fees. Hence, trading game profits would not be *zero* when multiple exchanges employ Discrete.

For simplicity, we model these ancillary revenues as being a constant amount $R > 0$ per trading game, of which each exchange j earns a proportion equal to its market share. That is, if exchange j has market share σ_j, it earns ancillary revenues of $\sigma_j R$. A more realistic model of ancillary revenues would have some ancillary revenues that depend strictly on market share (e.g., Tape revenues) and some ancillary revenues that depend on both market share and additional exchange-specific factors. For example, listing revenues are large for NYSE and Nasdaq yet negligible for BATS, and might be small for an entrant even if that entrant wins significant market share. The main point that is economically important for the discussion of adoption incentives below is just that incumbent exchanges, all else equal, prefer larger market share to smaller market share even when trading fee revenues and ESST revenues are both zero.

5.2 Adoption Incentives

Per-Trading-Game Payoffs. Summarizing our results so far, Figure 5.1 presents the per-trading-game payoffs for two exchanges employing potentially different market designs. In cell (C,C)—when both exchanges

[83] A similar concern arises if investor demand is not infinitely elastic with respect to prices and trading fees. While of course, in practice, demand is not likely to be infinitely elastic, the fact that observed trading fees are about \$0.0001 per share per side and sniping costs are likely an order of magnitude larger (see footnote 81) suggests that this potential issue is less empirically relevant than the tick-size friction.

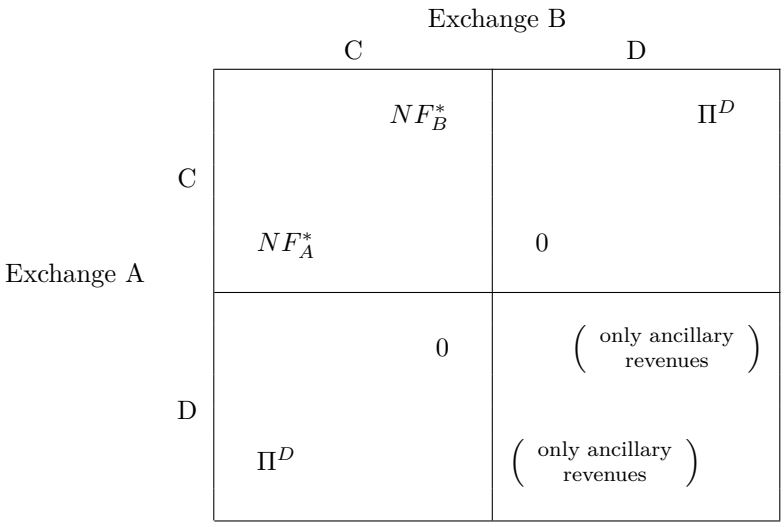


Figure 5.1: Per-trading game payoffs for two exchanges, each employing either the continuous-time limit order book (Continuous, or C) or discrete-time frequent batch auctions (Discrete, or D). For clarity, the diagram does not include ancillary revenues R, which each exchange earns in proportion to its market share. These payoffs constitute a prisoner's dilemma because $\Pi^D > NF_j^*$ for all j, and ancillary revenues are strictly positive for exchanges with positive market share. See the text for discussion.

employ Continuous, representing what we refer to as the *status quo*—each exchange j earns strictly positive economic profits from ESST fees (as well as its share of ancillary revenues, omitted from the Figure for clarity). In cell (D,D)—when both exchanges employ discrete—each exchange obtains only its fraction of ancillary revenues. Importantly, Discrete exchanges do not earn any ESST fees.

In cell (C,D) or (D,C)—where exchanges employ different designs—the Continuous exchange earns zero while the Discrete exchange earns positive rents. In particular, Discrete's trading fee revenues, denoted Π^D, represent a substantial share of sniping rents generated under the status quo (by Proposition 5.2) and must exceed what any individual exchange would be able to earn from ESST under the status quo (combining Proposition 3.4 and Proposition 5.2). That is, $\Pi^D > NF_j^*$ for all exchanges $j \in \mathcal{M}$.

Note that Figure 5.1 also represents normal-form game payoffs in a standard prisoner's dilemma: C and D represent the analogous "cooperate" and "defect" actions, D is a strictly dominant strategy for each exchange, and each exchange prefers payoffs under (C,C) to payoffs under (D,D).

Adoption Costs and Delay. The per-trading-game payoffs in Figure 5.1 are not sufficient to analyze adoption incentives on their own. In particular, there may be fixed costs related to being the first adopter of a new market design. Such costs arise from myriad sources, and include those related to: winning regulatory approval from the SEC, hiring additional engineers and support staff to implement and support the new design, the costs of explaining and marketing the new design to market participants, and so forth. We denote these costs by c_{adopt}. In addition, if the first Discrete exchange is a new *de novo* exchange, 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 (in addition to regulatory approval for the new rules), etc. These adoption and entry costs can be substantial. IEX is estimated to have raised just 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;[84] that is, costs that are part of what we call c_{entry}.

We also assume that after any exchange adopts Discrete, existing incumbent exchanges may be able to *imitate* Discrete by incurring costs $c_{imitate}$. Such imitation costs are likely to be much lower than initial adoption costs ($c_{imitate} < c_{adopt}$) for several reasons, including that the SEC would have already approved the market design, thereby establishing precedent; and the first adopter would have already incurred the costs of educating market participants on the new market design.

However, critically, imitation is not likely to be immediate. Rather than explicitly model a formal dynamic entry or adoption game, we instead allow any incumbent exchange to imitate the adoption of Discrete after some fixed amount of time. Formally, we make the following assumptions. First, we assume exchanges' per-trading game discount factor is equal to $\delta < 1$, and interpret c_{entry}, c_{adopt}, and $c_{imitate}$ as per-trading game costs paid in perpetuity: e.g., if the fixed cost of entry is \$100M, then $(\sum_{t=0}^{\infty} \delta^t) c_{entry} = \$100M$. Second, after T iterations of the trading game (starting from the time Discrete entered), all incumbent exchanges can adopt Discrete by paying the imitation costs $c_{imitate}$. Last, we assume that such imitation is profitable—that is, for at least one incumbent exchange j, imitation costs $c_{imitate}$ are less than its share of ancillary revenues; we believe that this is reasonable because imitation costs seem likely to be quite low relative to incumbents' share of \$315 million of annual tape revenue.

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 $\Pi^D + R$ per-trading game as long as all other exchanges remain with Continuous. However, given that (at least) one incumbent exchange would imitate Discrete when able to do so, after T periods Discrete would only earn its share of ancillary revenues. 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, or:

$$\rho(\Pi^D + R) + (1 - \rho)(\sigma_{entrant} R) \geq c_{adopt} + c_{entry}, \tag{5.2}$$

where $\rho \equiv (\sum_{t=0}^{T} \delta^t)/(\sum_{t=0}^{\infty} \delta^t)$ represents the *share of net present value* represented by the first T iterations out of an infinitely repeated series of trading games, and $\sigma_{entrant}$ denotes the anticipated market share of the entrant following adoption of Discrete by other incumbent exchanges. Profitable entry thus depends not only on whether the profitability of a standalone Discrete exchange, $\Pi^D + R$, is large relative to adoption and entry costs, but also on the term ρ which captures how quickly the entrant is imitated. As one anecdote that relates to the speed of imitation, consider that NYSE filed to imitate IEX's speed bump within around 7 months of its initial approval (under its NYSE American exchange license) and received SEC approval within about 1 year of IEX's initial approval.

Adoption Incentives: An Incumbent Exchange. We next ask whether any incumbent exchange, when all exchanges are operating as Continuous, would wish to adopt Discrete. Clearly exchanges prefer the outcome (C,C) to (D,D) because of ESST fees. Yet, if adoption costs of Discrete are sufficiently low

[84] 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.

and as long as rival exchanges cannot imitate without substantial delay, an incumbent exchange may find it worthwhile to adopt. The relevant condition for exchange j is:

$$\rho(\Pi^D + R) + (1-\rho)(\sigma'_j R) \geq c_{adopt} + \underbrace{(NF^*_j + \sigma_j R)}_{\text{opportunity cost (status-quo rents)}} . \tag{5.3}$$

The left-hand-side of (5.3) is essentially the same as that for the entrant, (5.2); the only difference is that the incumbent's anticipated market share when there are multiple discrete exchanges, denoted σ'_j, may differ from the entrant's, $\sigma_{entrant}$. The right-hand-side differs in that it does not include entry costs, c_{entry}, but instead includes the forgone status-quo profits from remaining a Continuous exchange, $NF^*_j + \sigma_j R$ (where σ_j represents the incumbent's market share in the status quo). Since an incumbent's ESST profits, NF^*_j, are likely to be significantly larger than c_{entry},[85] incumbents with existing profits to protect are even less likely than de novo entrants to adopt.

That incumbent exchanges continue to operate Continuous is thus consistent with them maintaining the "cooperative" (C,C) outcome of the prisoners' dilemma in Figure 5.1. 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].*"[86]

(Emphasis added). The quote suggests that industry participants believe that adoption costs of a new market design are substantial, and—more importantly—if a new design turns out to be successful, it would be swiftly imitated without any benefit to the first-mover. However, the quote does not mention the additional wedge between Nasdaq's (private) incentive to adopt a new design and social preferences that we have highlighted in Figure 5.1: by adopting a new market design that eliminates latency arbitrage (and by being copied by others), an incumbent exchange would lose its economic rents from the sale of exchange-specific speed technology.

5.3 Discussion: What Kind of Entry and Innovation Has Been Observed?

Several recent actions by financial exchanges can be understood through the lens of this section's analysis. In equities, only one exchange, the Chicago Stock Exchange (CHX), has to-date proposed a market-design innovation that would address sniping. Notably, CHX was technically an incumbent exchange, dating all the way back to 1882, but had negligible market share (<1%) and, to our understanding, did not have revenue from proprietary data or co-location. Examining the right-hand-sides of equations (5.2) and (5.3),

[85] The relevant comparison is the net present value of status-quo ESST rents against the one-off cost of entry. Even if the one-off cost of entry is $100M, which is substantial, the net present value of status-quo ESST rents, for the large incumbent exchanges, is likely to be at least an order of magnitude higher based on the evidence presented in Section 4.3.

[86] The event was a Workshop of The Program in the Law and Economics of Capital Markets at Columbia which featured a presentation of BCS 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).

this means that CHX's cost of adoption was perhaps uniquely low—it already had an exchange license, but had no rents to lose from exchange-specific speed technology. However, CHX's proposal was widely opposed by incumbents and was officially "stayed" by the SEC (cf. footnote 21), and CHX ultimately was acquired by the New York Stock Exchange in 2018 and withdrew its proposal.

Also in U.S. equities, a new *fourteenth* exchange was announced in early 2019. This potential entrant—called the Members' Exchange (MEMX) and owned by a consortium of nine major trading firms and broker-dealers—is *not* innovating on market design, but rather seems motivated in large part by concern over rising fees for proprietary data, co-location and connectivity, documented in Stylized Fact #7 (Osipovich, 2019*b*; Levine, 2019). In this regard, this entry can be viewed as a combination of business-stealing in the sense of Mankiw and Whinston (1986) and an attempt to gain bargaining leverage, rather than innovation on welfare-enhancing dimensions.

MEMX's entry strategy is reminiscent of the entry strategies employed by BATS and Direct EDGE, who both entered as exchanges in the years immediately following the implementation of Reg NMS. Indeed, one industry analyst noted in reference to MEMX that, "we've seen this playbook before—it's BATS 2.0" (Stafford, 2019). In our model of the status quo, since exchanges earn positive economic profits from ESST, there is an incentive to enter as another continuous exchange if the entrant has a way of obtaining market share. While, as emphasized, our model does not characterize the source of exchange market shares, it seems plausible that both BATS and Direct EDGE, by being owned by market participants that controlled significant order flow and liquidity provision, could reasonably have expected to obtain meaningful market share and hence economic profits.[87] MEMX may be motivated by similar reasoning.

In futures markets—which as emphasized above differ from equities in that contracts are proprietary to each exchange (i.e., there is no equivalent of Unlisted Trading Privileges), implying that an exchange may be better able to capture the returns from increasing their efficiency—there has in fact been some modest activity by the major exchanges towards addressing latency arbitrage. The Chicago Mercantile Exchange (CME) filed for patent protection on a variant of frequent batch auctions (Hosman et al., 2017), though, to date, it has not been introduced in the market. The Intercontinental Exchange (ICE) recently filed for regulatory approval for an asymmetric speed bump (Osipovich, 2019*a*), "designed to reduce latency advantages between traders engaged in arbitrage strategies against related markets,"(ICE, 2019) i.e., to reduce latency arbitrage. Predictably, the ICE proposal has been opposed by several high-frequency trading firms (U.S. Commodity Futures Trading Commission, 2019). More interesting is that ICE is the parent company of NYSE — so the same company appears to support market design innovations that reduce latency arbitrage in futures markets, yet oppose them in equities markets.[88]

6 Policy Responses

The basic question for policy is whether there will be a private-market solution to latency arbitrage and the arms race (i.e., "will the market fix the market"), or would some sort of regulatory intervention be required; and if the latter, of what form.

Our analysis in Section 5 suggests that private-market incentives may not be sufficient. For a de novo

[87] At the time of the initial public offering for BATS (who merged with Direct EDGE in 2014), there were thirteen market participants who were principal investors and together owned 93.3% of the stock (as measured by voting power), and comprised 45.8% of BATS's trading volume (BATS Global Markets, Inc., 2016).

[88] A good topic for future research would be to build a model of competing futures exchanges in which the exchanges compete in both the traditional manner (Pagano, 1989; Ellison and Fudenberg, 2003; Cantillon and Yin, 2008) and also decide what market design to adopt.

entrant, inequality (5.2) shows that entry will not occur if either imitation is too rapid or the costs of entry and adoption are too large, relative to the fees the entrant can charge as compensation for eliminating latency arbitrage. For an incumbent, inequality (5.3) shows that the incumbent will not adopt under similar conditions, with the difference that instead of paying entry costs to become a new exchange they pay the opportunity cost of foregone speed-technology rents. It seems empirically plausible, given the speed of imitation, the high costs of starting an exchange, and the high revenues incumbents enjoy from speed technology, that in practice neither entrants nor incumbents have sufficient incentive.

This same analysis also suggests, however, that a modest policy intervention might be sufficient to tip the balance of incentives and facilitate adoption. Specifically, any policy that gets either (5.2) or (5.3) to obtain would get a first-mover to adopt, which the model suggests would in turn help spur broader adoption, i.e., move the industry from the equilibrium with latency arbitrage to the equilibrium without latency arbitrage. The intuition for why such a "push" may suffice, as opposed to needing a market-design mandate, is that *investors* strictly prefer to trade on markets without the latency-arbitrage tax, and investors are ultimately who exchanges and trading firms make their money from—so, once such a market enters, private-market forces can then take over.[89] In this section, we discuss two sets of such policies that could accomplish this goal.

Policy Response 1: Reduce entry and adoption costs. The costs of starting a new stock exchange are substantial, and the risk of a new stock exchange design not getting approved are substantial. As evidence of the significant costs of entry, consider that the Investors' Exchange (IEX) has had to raise about $100M to date of venture capital, and that the Chicago Stock Exchange (CHX) was purchased for a reported $70M, when the main asset of CHX was thought to be its "medallion," i.e., its stock exchange license. 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*a*), was essentially rejected by the SEC, and then instead sold its exchange to NYSE. Both of these considerations raise the risk-adjusted cost of entering as a new exchange with a new market design (i.e., $c_{entry} + c_{adopt}$). Examining equation (5.2), it is immediate that, if policy could lower these costs, it would increase the incentives for a de novo entrant to innovate.

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 (cf. Budish, 2016*c*). With respect to frequent batch auctions specifically, it remains somewhat ambiguous whether quotes in a frequent batch auction market with a $\leq$1ms batch interval would be considered "immediately and automatically accessible" under the June 2016 SEC rules guidance. Such proactive clarification would certainly reduce risk, and would likely also reduce costs per se (e.g., legal costs).

In principle, if the social returns to a new market design are large but the private returns are negative,

[89] 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 continuous-time trading with latency arbitrage and the arms race, to discrete-time trading without latency arbitrage and the arms race. Understanding the tradeoffs between these two kinds of policy responses is outside the scope of this paper.

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. The maximum necessary subsidy would be $c_{entry} + c_{adopt}$, in the case of $\rho = 0$ (immediate imitation) and $\sigma_{entrant} = 0$ (entrant will gain no share once incumbents imitate).

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 speed bump was imitated (within 7 months by a new exchange under the NYSE family called NYSE American) also speaks to ρ being small in practice.

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 (even for one that 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 (cf. 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 FBA is in the public domain. Second, even if FBA 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.[90]

A potential alternative way to create market exclusivity would be to have the SEC grant a modest period of exclusivity to the innovator, during which time other exchanges would not be allowed to imitate the design (either identically or with designs judged by the SEC to be essentially similar). This idea is somewhat analogous to a practice of the Food and Drug Administration, wherein it grants a period of market exclusivity for certain kinds of drugs that, for various reasons, are not patentable (see 21 CFR § 314.108 (2018) and Food and Drug Administration (2015)). The purpose of the FDA policy is to induce drug companies to go through the effort and expense of the FDA clinical trials necessary to bring a new drug to market. Analogously, the purpose of the SEC exclusivity period would be to induce an exchange company to go through the effort and expense of the SEC approval process, and the other costs associated with developing and implementing a new market design.

[90]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).

7 Concluding Remarks

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 given sufficient opportunity and flexibility to do so, trading venues may not wish to innovate: in the current status quo, they—alongside high-frequency trading firms and speed technology providers—profit from the speed race generated by the existing market design, and stand to lose if it is eliminated. 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 proposals put forth in the last section are designed with this perspective in mind. Rather than mandate a particular market design, these proposals—borrowing simple economic insights from the innovation literature—attempt to alter the incentives for private innovation in a way that better align them with social preferences.

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 IO, finance and market design, with theory and empirical work guided by careful attention to institutional and regulatory detail—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.[91] 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).[92]

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 discussed 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*).

[91] 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.

[92] We underscore that each of these markets is large. U.S. stock exchange volume is on the order of $50 trillion per year. The U.S. treasury primary market is nearly $10 trillion per year (Hortaçsu, Kastl and Zhang, 2018) and secondary market volume exceeds $100 trillion (SIFMA, 2019). The CME S&P 500 index futures contract alone has annual trading volume of nearly $100 trillion (CME Group, Inc., 2019).

References

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

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

60 Minutes. 2014. "Is the U.S. Stock Market Rigged?" Aired on March 30, 2014. Script retrieved February 26, 2019 from https://www.cbsnews.com/news/michael-lewis-stock-market-rigged-flash-boys-60-minutes/.

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.

Armstrong, Mark. 2006. "Competition in Two-Sided Markets." *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.

Baldauf, Markus, and Joshua Mollner. 2018*a*. "High-Frequency Trading and Market Performance." Working Paper.

Baldauf, Markus, and Joshua Mollner. 2018*b*. "Trading in Fragmented Markets." Working Paper.

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.

Bloomberg Editorial Board. 2014. "Slowing Down the Stock Market." *Bloomberg Opinion*, June 18. Retrieved from `https://www.bloomberg.com/opinion/articles/2014-06-18/slowing-down-the-stock-market`.

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. 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. 2017. "Will the Market Fix the Market?" January 6. AEA/AFA Joint Luncheon Address. Retrieved April 18, 2019 from `https://www.aeaweb.org/webcasts/2017/luncheon`. Transcript available at `http://faculty.chicagobooth.edu/eric.budish/research/aeaafa-address-jan2017-Transcript.pdf`.

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." *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.

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." Octobeer 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.

CME Group, Inc. 2019. "Fiscal Year 2018 10-K." Retrieved May 6, 2019 from http://investor.cmegroup.com/node/43571/html.

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." Stanford University Graduate School of Business Research Paper No. 15-56.

Duffie, Darrell, and Piotr Dworczak. 2018. "Robust Mechanism 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.

eBay, Inc. 2018. "Fiscal Year 2017 10-K." Retrieved February 25, 2019 from http://d18rn0p25nwr6d.cloudfront.net/CIK-0001065088/fe207097-9fbf-4f30-9555-9e500aa3eefa.pdf.

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?" *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`.

Foucault, Thierry, and Albert J. Menkveld. 2008. "Competition for Order Flow and Smart Order Routing Systems." *The Journal of Finance*, 63(1): 119–158.

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.

Glode, Vincent, and Christian Opp. 2016. "Asymmetric Information and Intermediation Chains." *American Economic Review*, 106(9): 2699–2721.

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.

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.

ICE. 2019. "Re: Amendments to Rule 4.26 Order Execution (New Passive Order Protection Functionality) Submission Pursuant to Section 5c(c)(1) of the Act and Regulation 40.6(a)." Retrieved April 16, 2019 from `https://www.cftc.gov/sites/default/files/2019-02/ICEFuturessPassiveOrder020119.pdf`.

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.

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 from `https://www.bloomberg.com/opinion/articles/2019-01-07/traders-want-their-own-exchange-too`.

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." *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.

Menkveld, Albert J., Bart Z. Yueshen, and Haoxiang Zhu. 2017. "Shades of Darkness: A Pecking Order of Trading Venues." *Journal of Financial Economics*, 124(3): 503–534.

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

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.

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 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 from https://www.wsj.com/articles/wall-street-firms-plan-new-exchange-to-challenge-nyse-nasdaq-11546866121?mod=searchresults&page=1&pos=1.

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.

Rochet, Jean-Charles, and Jean Tirole. 2006. "Two-sided Markets: a Progress Report." *RAND Journal of Economics*, 37(3): 645–667.

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.

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

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

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`.

SIFMA. 2019. "US Treasury Trading Volume." Retrieved May 6, 2019 from `https://www.sifma.org/resources/research/us-treasury-trading-volume/`.

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

Tyc, Stephane. 2014. "A Technological Solution to Best Execution and Excessive Market Complexity." Quincy Data, LLC.

U.S. Commodity Futures Trading Commission. 2019. "Comments for Industry Filing IF 19-001." Retrieved April 15, 2019 from `https://comments.cftc.gov/PublicComments/CommentList.aspx?id=2946&ctl00_ctl00_cphContentMain_MainContent_gvCommentListChangePage=1_50`.

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. 2017. "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. 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.

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 Definition and Proofs

A.1 Order Book Equilibrium (OBE)

In the following definition and proofs, we denote by o_{ij} the *order* sent by TF i to exchange j, where an *order* is a set of messages that may include standard limit orders, cancellations, and IOCs. Denote by $\boldsymbol{o}_i \equiv \{o_{ij}\}_{j\in\mathcal{M}}$ the set of orders sent by TF i to all exchanges. Limit orders and IOCs sent to an exchange take the form (q_i, p_i), which states 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. We say that o_{ij} provides liquidity if it contains any limit order that offers to buy (or sell) some positive quantity at a price less than (or greater than) y. 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 it 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 in $\boldsymbol{o}_i$ 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 $\boldsymbol{o}_i$. We denote this relation among orders by $\boldsymbol{o}_i' \subseteq \boldsymbol{o}_i$. We say that $\boldsymbol{o}_i'$ is a (strict) *price improvement* over $\boldsymbol{o}_i$ if, for any $l \in (0, 1]$ and any exchange $j \in \mathcal{M}$, transacting l units on exchange j is weakly cheaper under o_{ij}' than under o_{ij}, and there exists some quantity $l \in (0, 1]$ and exchange j for which it is strictly cheaper to transact under o_{ij}' than under o_{ij}. (If strictly less than l units of liquidity is provided at any finite price under o_{ij}, then the cost of transacting l units under o_{ij} is assumed to be infinite.) Note that if $\boldsymbol{o}_i$ involves the provision of no liquidity, then any $\boldsymbol{o}_i'$ offering positive liquidity at any (finite) price on any exchange represents a price improvement.

Let $E\pi_i(\boldsymbol{o}_i, \boldsymbol{o}_{-i})$ represent TF i's expected profits from a trading game given its orders submitted to all exchanges and those submitted by all other trading firms, denoted $\boldsymbol{o}_{-i} \equiv \{o_{kj}\}_{k\neq i, j\in\mathcal{M}}$; let $\boldsymbol{o}_{-ik}$ denote orders for all TFs other than TF i and TF k. Expectations condition on the state $(y, \boldsymbol{\omega})$ at the beginning of Period 1 of each trading game, and are taken over the potentially random sequence in which orders are processed by exchanges in Period 1, and actions by nature and other market participants in Period 2.[93]

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—i.e., 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.[94] 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 game deviations and responses.

We now provide the formal definition of our order book equilibrium concept:

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 strictly 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

[93] Given our modeling assumptions, we assume that each continuous limit order book exchange serially processes messages sent in the same period from TFs with the same speed technology in a random sequence, with the queue priority of liquidity on an exchange's limit order book determined first by price and then by the time it is added (i.e., processed).

[94] 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.

there is some other TF $k \neq i$ and strictly profitable withdrawal of liquidity $\boldsymbol{o}'_k(\cdot) \subseteq \boldsymbol{o}^*_k$—which potentially conditions on the interim state $\tilde{\boldsymbol{\omega}}$ arising from the processing of Period-1 orders $(\boldsymbol{o}'_i, \boldsymbol{o}^*_{-i})$—that renders TF i's deviation no longer strictly profitable (i.e., $E\pi_i(\boldsymbol{o}'_i, (\boldsymbol{o}'_k(\tilde{\boldsymbol{\omega}}), \boldsymbol{o}^*_{-ik})) \leq E\pi_i(\boldsymbol{o}^*_i, \boldsymbol{o}^*_{-i})$).

(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 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 strictly profitable reaction $\boldsymbol{o}'_k$ that renders TF i's deviation no longer strictly profitable, and either: (a) $\boldsymbol{o}'_k$ withdraws liquidity from $\boldsymbol{o}^*_k$ (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, engages in a strictly profitable withdrawal of liquidity in response.

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 uniquely optimal strategies and described in the statement of the proposition. Consider the following set of 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, 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, 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.
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 uniquely 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 on all exchanges contained in non-empty set $\mathcal{J} \subseteq \mathcal{M}$, and no TF has purchased ESST on any exchange $m \notin \mathcal{J}$. Further assume that trading fees are zero for all exchanges $j \in \mathcal{J}$ and weakly 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 $\sum_{j \in \mathcal{M}} \sigma_j^* = 1$ and $\sigma_m^* = 0$ if $f_m > 0$.

(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 orders 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., $\sum_{j \in \mathcal{M}} l_j^* = 1$ and $l_m^* = 0$ if $f_m > 0$) 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 with zero trading fees involves a single unit of liquidity provided across all exchanges with zero trading fees at spread $s^*_{continuous}$.

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 $\pi_N^{lone-wolf} \in (\frac{N-1}{N} \times \frac{\Pi^*_{continuous}}{N}, \frac{\Pi^*_{continuous}}{N})$. 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:

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

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 $\frac{1}{N}\lambda_{public}L(\cdot)$. We first prove that $\bar{s}_N < \tilde{s}_N < s^*_{continuous}$, where $\tilde{s}_N$ is the solution to (A.1). The first inequality 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 $\tilde{s}_N > \bar{s}_N$. Similarly, it follows that $\tilde{s}_N < s^*_{continuous}$: 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 $\pi_N^{lone-wolf} \in (\Pi^*_{continuous} \times (N-1)/N^2, \Pi^*_{continuous}/N)$ (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 TF i submits an order to provide one unit of liquidity at spread $\tilde{s}_N$ across exchanges in $\mathcal{J}$ according to $\boldsymbol{\sigma}^*$, and other TFs only snipe (i.e., no orders are submitted by TFs $k \neq i$ in Period 1 of each trading game); and investors route 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 $\pi_N^{lone-wolf} > \frac{N-1}{N^2}\Pi^*_{continuous}$. In the described equilibrium, TF i earns $\pi_N^{lone-wolf}$ while the other $N-1$ TFs that snipe him earn $\pi_N^{non-lone-wolf} \equiv \frac{1}{N}\lambda_{public} L(\tilde{s}_N) = \lambda_{invest} \frac{\tilde{s}_N}{2} - (\frac{N-2}{N-1}\lambda_{public} + \lambda_{private}) L(\tilde{s}_N)$ (which follows from the definition of $\tilde{s}_N$ provided in (A.1)). At this spread, the other $N-1$ TFs prefer sniping at $\tilde{s}_N$ to providing at a strictly narrower spread on another exchange $m \notin \mathcal{J}$, and hence do not want to undercut TF i. From these two

expressions, we can compute the difference in profits between the other $N-1$ TFs and TF i:

$$\pi_N^{non-lone-wolf} - \pi_N^{lone-wolf} = (\frac{N-1}{N} - \frac{N-2}{N-1})\lambda_{public}L(\tilde{s}_N) = \frac{1}{N(N-1)}\lambda_{public}L(\tilde{s}_N) \tag{A.4}$$

which is strictly positive for any $N \geq 2$: i.e., the lone-wolf liquidity provider at spread $\tilde{s}$ makes less than what the other TFs earn by sniping (recall that $\tilde{s}_N < s^*_{continuous}$, and at $s^*_{continuous}$ TFs are indifferent between liquidity provision and sniping). Next, note that the profits that i and the $N-1$ TFs that snipe i earn must by definition equal total sniping rents at the lone-wolf spread $\tilde{s}_N$: i.e., $\pi_N^{lone-wolf} + (N-1)\pi_N^{non-lone-wolf} = \lambda_{public}L(\tilde{s}_N)$. Substituting in for $\pi_N^{non-lone-wolf}$ using (A.4) yields:

$$\begin{aligned}
\pi_N^{lone-wolf} + (N-1)(\pi_N^{lone-wolf} + \frac{1}{N(N-1)}\lambda_{public}L(\tilde{s}_N)) &= \lambda_{public}L(\tilde{s}_N) \\
(\Leftrightarrow) \qquad \pi_N^{lone-wolf} &= \frac{N-1}{N^2}\lambda_{public}L(\tilde{s}_N) \\
(\Leftrightarrow) \qquad \pi_N^{lone-wolf} &> \frac{N-1}{N^2}\Pi^*_{continuous}
\end{aligned}$$

where the last inequality follows since $\Pi^*_{continuous} = \lambda_{public}L(s^*_{continuous})$ and $\tilde{s}_N < s^*_{continuous}$ implies that $L(\tilde{s}_N) > L(s^*_{continuous})$.

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 no orders to 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 messages to provide one unit of liquidity on some exchange $j \in \mathcal{J}'$ at spread $\tilde{s}_N$ (defined in (A.1)) around y, all other TFs submit no orders to 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 and all other TFs submit no orders to 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 $\mathcal{J}' = \mathcal{J} \cap \mathcal{K}$.

- 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 (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 yield 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 orders are submitted 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 (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). In such a vector, ESST fees must be equal across exchanges, and there must be a constant $\tilde{F}$ such that $F'_j = \tilde{F}$ for all $j \in \mathcal{M}$: if not, then the lowest ESST fee, $min_j F'_j$, can be increased by some amount $\varepsilon > 0$ without violating condition (3.2) (as there would be at least one exchange with strictly higher fees), and $\boldsymbol{F}'$ would then not be a maximizer over the sum of ESST fees. 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 $M \times \tilde{F} \leq (\Pi^*_{continuous}/N) - (\pi_N^{lone-wolf} - \tilde{F})$ in ESST fees across all exchanges if condition (3.2) holds. Since $\pi_N^{lone-wolf} > \frac{(N-1)}{N^2}\Pi^*_{continuous}$ by Lemma A.2, it follows that $\tilde{F} < \frac{1}{(M-1)N^2}\Pi^*_{continuous}$ and $M \times N \times \tilde{F} < \frac{M}{(M-1)N}\Pi^*_{continuous}$, 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. For this section, we assume that the jump size distribution is absolutely continuous.

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:

$$\lambda_{invest}(\frac{\bar{s}_{discrete}(f)}{2} - f) - \lambda_{private}L(\bar{s}_{discrete}(f), f) = 0\ , \tag{A.5}$$

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.5 represents the revenues a liquidity provider earns when a 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.

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, $f^*_{discrete}$ exists and is unique.

Lemma A.3. *There exists a unique solution $f^*_{discrete}$ to:*

$$\frac{\bar{s}_{discrete}(f^*_{discrete})}{2} + f^*_{discrete} = \frac{\bar{s}_{continuous}}{2}\ . \tag{A.6}$$

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, $\bar{s}_{discrete}(f) = s(f)$); by the implicit 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.[95] Since $\frac{\bar{s}_{discrete}(f)}{2} + f$ is thus

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

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.6). The rest of the statement directly follows. □

A.3.1 Proof of Proposition 5.1

We prove that a more general version of Proposition 5.1 holds when Discrete charges a weakly positive trading fee: i.e., in any equilibrium of any Stage 3 trading game given state (y, ω) 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 $f \in [0, f^*_{discrete})$ on Discrete, exactly one unit of liquidity is provided on Discrete at bid-ask spread $\bar{s}_{discrete}(f)$ 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 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 $\bar{s}_{discrete}(f)$ 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 $\bar{s}_{discrete}(f)$ 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 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 $\bar{s}_{discrete}(f) + \varepsilon$ 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 $\bar{s}_{discrete}(f) + \varepsilon/2$ 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 $\bar{s}_{discrete}(f)$, and no liquidity is provided elsewhere. By the same arguments in Lemma 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), else the liquidity provider would be better off withdrawing its order; at any spread weakly greater than $\bar{s}_{continuous}$, there is a strictly profitable deviation by any slow TF to provide the same amount of liquidity on Discrete at some spread $s' \in (\bar{s}_{discrete}(f), \bar{s}_{continuous})$, implying these strategies cannot be an equilibrium. Last, consider next a candidate equilibrium where any amount of liquidity on Discrete is provided at some spread $\tilde{s} \neq \bar{s}_{discrete}(f)$: 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 strictly profitable deviation by any slow TF to provide the same amount of liquidity at any spread $s' \in (\bar{s}_{discrete}(f), \tilde{s})$.

---

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.3.2 Proof of Proposition 5.2

Existence. First, we establish that if Discrete charged any trading fee $f' > f^*_{discrete}$ 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.[96] 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' 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

[96] 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.

elsewhere, Discrete earns (in expectation) at least $\frac{N-1}{N}\Pi^*$ per trading game. Recall that $\bar{s}_{continuous}$ is given by the solution to $\lambda_{invest}\frac{\bar{s}_{continuous}}{2} - (\lambda_{private} + \frac{N-1}{N}\lambda_{public})L(\bar{s}_{continuous}) = 0$. Using (A.6), this expression can be rewritten as: $\lambda_{invest}(\frac{\bar{s}_{discrete}(f^*_{discrete})}{2} + f^*_{discrete}) - \lambda_{private} \cdot L(\bar{s}_{discrete}(f^*_{discrete}) + 2f^*_{discrete}) - \frac{N-1}{N}\lambda_{public}L(\bar{s}_{continuous}) = 0$. Subtracting from this the definition of the zero-profit spread on Discrete (given by $\lambda_{invest}(\frac{\bar{s}_{discrete}(f^*_{discrete})}{2} - f^*_{discrete}) - \lambda_{private}L(\bar{s}_{discrete}(f^*_{discrete}) + 2f^*_{discrete}) = 0$) 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. 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 $f > 0$. 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 (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).

B Supporting Details for Average Trading Fee Calculations in Table 4.2

In this appendix we provide supporting details for the calculation of average per-share per-side trading fees as reported in Table 4.2. 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.

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.

C Details for NASDAQ and NYSE Data and Co-Location Revenue Estimates

In this appendix we provide details for our calculations of market data and co-location/connectivity revenues for NASDAQ and NYSE, as indicated in the text of Stylized Fact #6.

NASDAQ's fiscal year 2015 10-K reports market data and co-location/connectivity revenue only at the global level – $399M and $239M, respectively.[97] To get from global to the US, for market data, we utilize information in its 2013 10-K filing that breaks out its market data business geographically: US 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 US:global revenue ratio, of 71%. Last, we need to separate out NASDAQ's US Equities business from its US Options business. We take two approaches. First, we assume that NASDAQ's market data and co-location revenue from US Equities vs. US Options is proportional to its trading volume in US Equities vs. US Options. Second, we assume that NASDAQ's US Options business generates the same market data and co-location revenue revenue as BATS's US Options business, scaled up for NASDAQ's larger US 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 US 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 US 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 US equities and US 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

[97] 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.

corresponds to "NYSE connectivity fees and colocation service revenues".[98] 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%.[99] For comparison, BATS's US equities growth rate for the 2014 to 2015 period was 19.2% for co-location/connectivity and 12.4% for market data,[100] 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 overall US business, and then utilize the same two methods described above for NASDAQ to obtain estimates for NYSE's US Equities business. This yields a range of $218.9-$241.5M for US equities market data, $251.6-$281.5M for US equities co-lo/connectivity, and $470.5-$523.0M combined.

[98] 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.

[99] 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.

[100] 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 colo/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.