Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Queueing Dynamics and State Space Collapse in Fragmented Limit Order Book Markets

Costis Maglaras , Ciamac C. Moallemi , Hua Zheng

To cite this article:
Costis Maglaras , Ciamac C. Moallemi , Hua Zheng (2021) Queueing Dynamics and State Space Collapse in Fragmented Limit Order Book Markets. Operations Research 69(4):1324-1348. https://doi.org/10.1287/opre.2020.1989

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 69, No. 4, July–August 2021, pp. 1324–1348
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Queueing Dynamics and State Space Collapse in Fragmented Limit Order Book Markets

Costis Maglaras,[a] Ciamac C. Moallemi,[a] Hua Zheng[b]

[a] Graduate School of Business, Columbia University, New York, New York 10027; [b] JP Morgan Chase, New York, New York 10029

Contact: c.maglaras@gsb.columbia.edu, https://orcid.org/0000-0002-4283-2177 (CM); ciamac@gsb.columbia.edu, https://orcid.org/0000-0002-4489-9260 (CCM); hua.zheng@jpmchase.com (HZ)

Received: July 10, 2017
Revised: May 23, 2019
Accepted: September 30, 2019
Published Online in Articles in Advance: June 14, 2021

Subject Classifications: applications: queues; brokerage/trading: financial institutions; networks: queues
Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2020.1989

Abstract. In modern equity markets, participants have a choice of many exchanges at which to trade. Exchanges typically operate as electronic limit order books under a *price-time* priority rule and, in turn, can be modeled as multiclass first-in-first-out queueing systems. A market with multiple exchanges can be thought as a decentralized, parallel queueing system. Heterogeneous traders that submit limit orders select the exchange (i.e., the queue), in which to place their orders by trading off financial considerations against anticipated delays until their orders may fill. These limit orders can be thought of as jobs waiting for service. Simultaneously, traders that submit market orders select which exchange (i.e., queue) to direct their order. These market orders trigger instantaneous service completions of queued limit orders. In this way, the *server* is the aggregation of self-interested, atomistic traders submitting market orders. Taking into account the effect of investors' order-routing decisions across exchanges, we find that the equilibrium of this decentralized market exhibits a state space collapse property whereby (a) the queue lengths at different exchanges are coupled in an intuitive manner; (b) the behavior of the market is captured through a one-dimensional process that can be viewed as a weighted aggregate queue length across all exchanges; and (c) the behavior at each exchange can be inferred via a mapping of the aggregated market depth process that takes into account the heterogeneous trader characteristics. The key driver of this coupling phenomenon is anticipated delay as opposed to the queue lengths themselves. Analyzing a trade and quote data set for a sample of stocks over a one-month period, we find empirical support for the predicted state space collapse. Separately, using the data before and after NASDAQ's natural fee-change experiment from 2015, we again find agreement between the observed market behavior and the model's predictions around the fee change.

Funding: This work was supported by the National Science Foundation [Grant CMMI-1235023].
Supplemental Material: The online supplement is available at https://doi.org/10.1287/opre.2020.1989.

Keywords: applications: queues • brokerage/trading: financial institutions • networks: queues

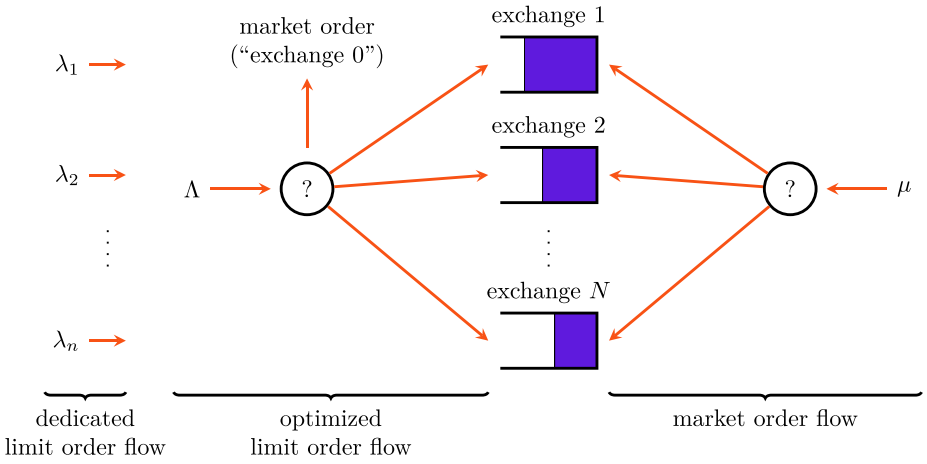
1. Introduction

1.1. Motivation

Modern equity markets are highly fragmented. In the United States alone, there are more than a dozen exchanges and about 40 alternative trading systems where investors may choose to trade. Market participants, including institutional investors, market makers, and opportunistic investors, interact within today's high-frequency, fragmented marketplace with the use of electronic algorithms that differ across participants and types of trading strategies. At a high level, they dynamically optimize where, how often, and at what price to trade, seeking to achieve their own best execution objectives while accounting for short-term differences or opportunities across the various exchanges. Exchanges function as electronic limit order books, typically operating under a price-time priority rule: resting orders are prioritized for trade first based on their respective prices and then, at a given price, according to their time of arrival (i.e., in first-in-first-out (FIFO) order). The dynamics of an exchange can be understood as that of a multiclass system of queues, where each queue is associated with a price level. Job arrivals into these queues correspond to new limit orders posted at the respective prices. Market orders trigger executions that, in queueing system parlance, correspond to service completions.

The market, consisting of multiple exchanges, can be viewed as a stochastic network that evolves as a collection of parallel, multiclass queueing systems. Figure 1 depicts one side of the market at one price level. Heterogeneous, self-interested traders optimize where to route their limit and market orders, coupling the dynamics of these parallel queues. Studying the interaction effects between market fragmentation and high-frequency, optimized order-routing decisions is

Figure 1. (Color online) A One-Sided, Top-of-Book Model of Multiple Limit Order Books



Notes. Limit orders (i.e., jobs) arrive to each exchange (modeled by the respective queues) in (a) dedicated streams and (b) optimized limit order placement decisions. Liquidity is removed through the arrival of decentralized, self-interested market orders, acting as service completions.

an important issue in understanding market behavior and trade execution and is the main focus of this paper.[1]

At a point in time, conditions at the exchanges may differ with respect to the best bid and offer[2] price levels, the market depth at various prices, recent trade activity, and so on. Exchanges publish real-time information for each security that allow investors to know or compute these quantities. These, in turn, imply differences in a number of execution metrics across exchanges, such as the probability that an order will be filled, the expected delay until such a fill, or the adverse selection associated with a fill. Exchanges also differ with respect to their underlying economics. Under the *make-take* pricing that is common, exchanges typically offer a rebate to liquidity providers (i.e., investors that submit limit orders that make markets when their orders get filled; simultaneously, exchanges charge a fee to takers of liquidity that initiate trades using marketable orders that transact against posted limit orders). Fees range in magnitude and are typically between −\$0.0010[3] and \$0.0030 per share traded. Because the typical bid-offer spread in a liquid stock is \$0.01, the fees and rebates are a significant fraction of the overall trading costs and material in optimizing over routing decisions. Most retail investors do not have access to this information, but essentially all institutional investors and market makers—that, taken together, account for almost all trading activity—have access and do make use of this information. They use so-called *smart order routers* that take into account real-time state information and formulate an order-routing problem that considers various execution metrics in order to decide whether to place a limit order or trade immediately with a market order and accordingly to which venue(s) to direct their order. Investors are heterogeneous; specifically, they differ with respect to the way that they trade off metrics such as price, rebates, and delays, primarily driven by their intrinsic patience until they fill their order.

From a modeling viewpoint, the aforementioned system consists of parallel multiclass queues (the exchanges) that differ in their economics and anticipated delays. These subsystems are decentralized. Moreover, service capacity is neither centrally controlled nor dedicated as is typical in production or service systems. Instead, it emerges by aggregating individual market orders (service completions) directed to different queues, themselves optimizing over heterogeneous tradeoffs between economics and operational metrics related to queueing effects.

1.2. Summary of Results

First, the paper offers a novel model for order routing in fragmented markets that takes into account queueing phenomena in limit order books, as well as the atomistic limit order placement and market order (service completions) routing decisions. This model explicitly leverages ideas for the economics of queues literature to capture the tradeoff between delay and rebate capture in the routing decision of limit orders. It also incorporates, in a reduced form, the self-interested routing decisions of marker orders that comprise the service completion process. The resulting model is a two-sided parallel queue system. The self-interested nature of the service completion process may be of independent interest; for example, one possible application might be in modeling personnel that work in retailing that may strategize over which customer to help next, or self-interested drivers in a ride hailing network that can select where to drive their car when they are not serving customers. The formulation of the limit order-routing problem, importantly, incorporates the heterogeneous preferences of

the various market participants with respect to the way they trade off delays (time) with the anticipated rebate (money).[4]

Second, from a methodological viewpoint, we study a deterministic and continuous fluid model associated with this system that takes into account the routing decisions of atomistic limit order placements and market orders (service completions). The key result is to characterize the structural form of the equilibrium state of this fluid model and derive a form of state space collapse (SSC) property. The market equilibrium and SSC are not the result of the price protection mechanism[5] imposed in the U.S. equities market. Rather, they arise out of order-routing decisions among exchanges that offer to trade at the same price level but at different (rebate, delay) combinations. We characterize this coupling effect that yields a strikingly simplifying property whereby the behavior of the multidimensional market reduces to that of a one-dimensional system expressed in terms of what we refer to as *workload*, which is an aggregate measure of the total available liquidity. In equilibrium, the workload is a sufficient statistic that summarizes the state of the market. The expected delay at each exchange is proportional to the workload, where the proportionality constant depends on exchange specific parameters. In equilibrium, if one exchange is experiencing long delays, then the other exchanges will also be experiencing proportionally long delays. Conversely, if (out of equilibrium) one exchange has temporarily an atypically small associated delay relative to its cost structure, the new order flow will quickly take advantage of that delay/cost opportunity and erase that difference.[6] For $N = 2$ exchanges, we use a geometric argument to prove that the fluid model transient starting from an arbitrary initial condition converges to the equilibrium state in finite time. We conjecture that a similar argument carries through when there are $N > 2$ exchanges. The specific form of our SSC result depends on our assumptions regarding the routing of limit and market orders, as is typical of such results. The parameters that describe the heterogeneity of trader preferences and the fees and rebates at the various exchanges dictate the resulting equilibrium state.

Third, we empirically verify the state space collapse property for a sample of trade and quote (TAQ) data for the month of September 2011 for the 30 securities that comprise the Dow Jones Index. Although all are liquid stocks, these securities differ in their trading volumes, price, volatility, and spread. Our methodological results suggest certain testable hypotheses, most notably regarding the effective dimensionality of the market dynamics, the linear relation between the expected delays across exchanges, and the relation between expected delays and market-wide workload. These empirical findings are summarized in Section 4 and find statistical support for the SSC prediction of our model, despite its stylized assumptions. To our knowledge, this seems to be one of the first empirical verifications of SSC in a real and complex stochastic processing system.

The one-dimensional workload characterization seems to offer a tractable model for downstream analysis of questions that pertain to exchange competition (e.g., how to set fees or associated volume tiers), policy questions that may affect the routing decision problem or impose exogenous transaction costs (e.g., a transaction tax), and market design questions (e.g., whether the coexistence of competing, differentially priced exchanges is beneficial from a welfare perspective). To that effect, one of its predictions is that if a high-rebate exchange was to lower its rebate and fee, the market response would be such that the queues and trading volumes would equilibrate in such a way to reduce the anticipated delay for limit orders placed in that exchange compared with the delays encountered in other exchanges. Lower fees would make the exchange more attractive for the submission of market orders, possibly increasing volume. On the other hand, lower rebates would reduce the attractiveness of the exchange for placing limit orders, which would, all other things kept constant, lead to a reduction in queue sizes. Small queues, in turn, discourage market order activity. Our model predicts the above opposing effects would balance out through their effect on trading delays, which should decrease, as traders will be willing to wait less to receive the smaller rebate. In 2015, NASDAQ ran a pilot experiment to precisely explore this issue for a sample of 14 stocks. In Section 4.4, we find strong statistical support for our model's prediction. Our study complements several industry reports that studied market data before and after this natural experiment (Hatheway 2015, Pearson 2015) that had primarily focused on descriptive statistics and ex ante/ex post comparisons of volume and depth comparisons.

1.3. Literature Survey

There are two strand of literature that we briefly review. The first is on market microstructure and financial engineering and focuses on the structure and behavior of limit order books. Apart from the classical market microstructure models, such as those proposed by Kyle (1985), Glosten and Milgrom (1985), and Glosten (1987), our paper is related to several strands of work. First is the set of papers that report on empirical analyses of the dynamics of exchanges that operate as electronic limit order books, such as Bouchaud et al. (2004), Griffiths et al. (2000), and Hollifield et al. (2004) and the review article of Parlour and Seppi (2008). Related to the above work, there is a

body of literature that studies the effect of adverse selection, which factors in order placement decisions (Holthausen et al. 1990, Sofianos 1995, Keim and Madhavan 1998, Dufour and Engle 2000, Huberman and Stanzl 2004, Gatheral 2010).

Second, there are several papers that study market fragmentation, exchange competition, and their effect on market outcomes dating back to the work of Hamilton (1979), Glosten (1994, 1998), and more recently, Bessembinder (2003) and Barclay et al. (2003). A number of papers, including O'Hara and Ye (2011), Jovanovic and Menkveld (2011), and Degryse et al. (2011), empirically study the impact of exchange competition on available liquidity and market efficiency. Biais et al. (2010) and Buti et al. (2011) consider the impact of differences in tick-size on exchange competition, whereas in the markets we consider, the tick-size is uniform. Foucault et al. (2005) describe a theoretical model to understand make-take pricing when monitoring the market is costly. Malinova and Park (2010) empirically study the introduction of make-take rebates and fees in a single market. Foucault and Menkveld (2008) study the impact of smart order routing on market behavior in a setting with two exchanges. However, they discuss smart order-routing decisions by traders submitting market orders aiming to optimize their execution price (i.e., in a setting where exchanges operate without a price protection mechanism, like Regulation National Market System (Reg NMS) that applies to the U.S. equities market, which would eliminate the opportunity from such routing decisions); their paper does not consider the routing decisions of limit orders and disregards queueing effects. van Kervel (2012) considers the impact of order routing in a setting where market makers place limit orders on multiple exchanges simultaneously to increase execution probabilities. Their analysis ignores economic and execution delay differences between venues. Sofianos et al. (2011) discuss smart order placement decisions in relation to their all-in cost, introducing similar considerations to the ones explored in this paper. More recently, Cont and Kukanov (2013) studied a smart order-routing control problem, where a trader decides how to split a noninfinitesimal order size across multiple venues, taking into account the delay and rebate differences across exchanges and operating under a control horizon T. Our model considers traders that submit infinitesimal order sizes, so the decision of how to split their order is not relevant, but they are heterogeneous in terms of how they trade off delay with rebates; our model also considers the routing of market orders and tries to characterize the (stylized) market equilibrium.

Third, there is a growing body of work that develops models of limit order book dynamics and studies optimal execution problems. Obizhaeva and Wang (2006), Rosu (2009), Alfonsi et al. (2010), and Parlour (1998), treat the market as one limit order book and use a model of market impact and abstracts away queueing effects. The high-frequency behavior of limit order books can probably be best modeled and understood as that of a queueing system. This connection has been explored in recent work, starting with Cont et al. (2010) (see also Blanchet and Chen 2013; Cont and De Larrard 2013; Guo et al. 2013; Lakner et al. 2016, 2017; Maglaras et al. 2014; Avellaneda et al. 2011); this set of papers does not consider fragmentation.

The second strand of literature related to our work is on stochastic modeling and relates to the asymptotic analysis tools that motivate our method of analysis and the area of queueing systems with strategic consumers. So-called equivalent workload formulations and the associated idea of state space collapse arise in stochastic network theory in the context of their approximate Brownian model formulations. This idea has been pioneered by the work of Harrison (1988, 2000). Workload fluid models were introduced in Harrison (1995). The condition that guarantees that parallel server systems exhibit SSC down to one-dimensional systems was introduced by Harrison and Lopez (1999), and two papers that establish SSC results with optimized routing of order arrivals are Stolyar (2005) and Chen et al. (2019). We model market order-routing decisions via a reduced-form state-dependent service rate process. Mandelbaum and Pats (1995) derive fluid and diffusion approximations for such queues. Our analysis is itself deterministic, building on ideas and tools from the asymptotic analysis of queues. We do not provide a limit theorem to justify the deterministic fluid model we postulate as the system model but instead focus on its analysis and implications. SSC results tend to be pathwise properties, established via an asymptotic analysis after an appropriate rescaling of time. In our system, arrival rates of limit and market orders vary stochastically over time on a slower time scale than that of the transient fluid model dynamics. An asymptotic analysis on the slower time scale of the event rate variations, in the spirit of the so-called pointwise stationary fluid models (PSFM), would establish such a pathwise SSC property by exploiting the transient fluid model results of this paper. Standard machinery for establishing such results either exploit the work by Bramson (1998) or Bassamboo et al. (2004). Our model seems to satisfy the key requirements that one would need to derive the PSFM and as a result the sample path version of the SSC property, but we will not pursue this here other than a short discussion in Section A of the online supplement.

Optimal order placement decisions are made according to an atomistic choice model as per Mendelson and Whang (1990). In the context of queueing models with pricing and service competition, there are several papers including those of Luski (1976), Levhari and Luski (1978), Li and Lee (1994), and Lederer and Li (1997). Cachon and Harker (2002) and So (2000) analyze customer choice models that divert from the lowest cost supplier under $M/M/1$ system models. Allon and Federgruen (2007) studied the competing supplier game in a setting where the offered services are partial substitutes. An extensive survey is provided in Hassin and Haviv (2003).

Most of these papers look at static rules, where consumers make decisions based on steady-state expected delays. Chen et al. (2019) considers competing suppliers and arriving consumers making decisions based on real-time information, like in our model, but where each supplier has his own dedicated processing capacity; the resulting dynamics are different and only couple through order arrivals. The nature of the service completion process that emerges as the aggregation of infinitesimal self-interested contributions appears novel viz the existing literature. Finally, Plambeck and Ward (2006) study an assemble-to-order system that involves a two-sided market fed by product requests on one side and raw materials on the other, but such systems allow queueing on both sides, and the flow of material is controlled by the system manager. Caldentey et al. (2009) and Gurvich and Ward (2014) study the dynamics of matching queues.

2. Model

We propose a stylized model of a fragmented market consisting of N distinct electronic limit order books simultaneously trading a single underlying asset. The model will take the form of a system of parallel FIFO queues; new price and delay sensitive jobs arrive over time and optimize their routing decisions, and self-interested agents arrive over time and optimize where to route their market order that triggers an instantaneous service completion at the respective queue (i.e., this routing decision happens at the end of the service time). Our focus is to understand the effect of optimized order-routing decisions on the interaction between multiple limit order books. We make a number of simplifying assumptions that aid the tractability of our model studied in Sections 2 and 3.

One-sided market: We model one side of the market, which, without loss of generality, choose to be the bid side, where investors post limit orders to buy the stock and wait to execute against market orders directed by sellers. Although our model is one sided, it may be possible to extend our equilibrium analysis to a two-sided model where both sides are simultaneously coupled through the flow of market orders. Exploring such a two-sided model is an important direction for future research.

Top-of-book only: Limit orders are distinguished by their limit price. We only consider limit orders at each exchange posted at the national best bid price, which is the highest bid price available across all exchanges: the top-of-book. A profit-maximizing seller would only choose to trade at the top of book, and, in fact, in the United States, this is enforced de jure by U.S. Securities and Exchange Commission (SEC) Regulation NMS.

Fluid model: We consider a deterministic fluid model, or *mean field* model, where the discrete and stochastic order-arrival processes are replaced by continuous and deterministic analogues, where infinitesimal orders arrive continuously over time at a rate that is equal to the instantaneous intensity of the underlying stochastic processes. This model can be justified as an asymptotic limit using the functional strong law of large numbers in settings where the rates of order arrivals grow large but the size of each individual order is small relative to the overall order volume over any interval of time. It is well suited for characterizing transient dynamics in such systems, which is the time scale over which queue lengths drain or move from one configuration to another; this is also the relevant time scale in order-routing decisions. For liquid securities, orders arrive on a time scale measured in milliseconds to seconds, whereas queueing delays are of the order of seconds to minutes.

Constant arrival rates: Market activity exhibits strong time-of-day effects, typically over longer time scales (e.g., minutes to hours) than what we focus on. The analysis of the next section assumes that arrival rates are constant and do not depend on time or the state at the exchanges.

Our model is illustrated in Figure 1. For each of the N exchanges, there is a (possibly empty) queue of resting limit orders at the national best bid price. The vector of queue lengths at time t is denoted by $Q(t) \triangleq (Q_1(t), Q_2(t), \ldots, Q_N(t)) \in \mathbb{R}_+^N$.

2.1. Limit Order Routing

A continuous and deterministic flow of investors arrives to the market with the intent of posting an infinitesimal limit order. This flow consists of two types:

Dedicated limit order flow arrives at rate $\lambda_i \geq 0$ and is destined to exchange i, independent of the state $Q(t)$ at the various exchanges. This flow could represent, for example, investors that may not have the ability to route orders to all exchanges or to make real-time order-routing decisions.

Optimized limit order flow arrives at a rate $\Lambda > 0$. Each infinitesimal investor observes the state of the

market, $Q(t)$, and optimizes over where to route the associated infinitesimal order, or, if conditions are unfavorable, not to leave a limit order and to trade instead with a market order at the offered (other) side of the market; this option is denoted by $i = 0$.

Once a limit order is posted at a particular exchange, it remains queued until it is executed against an arriving market order. This disregards order cancellations. Cancellations occur, for example, when time-sensitive orders *deplete* their patience and cancel to cross the spread and trade with a market order; when investors perceive an increased risk of adverse selection; and so on. This assumption simplifies the order-routing decision and leads to a tractable analysis.[7]

2.1.1. Expected Delay. All things being equal, an investor would prefer a shorter delay until an order gets executed. Apart from price risk considerations, this is often because of exogenous constraints on the speed at which the order needs to get filled; in many instances, a limit order may be a child order that is part of the execution plan of a larger parent order, which itself needs to be filled within a limited time horizon and under some constraints on its execution trajectory defined by its *strategy*. As will be seen in Section 4, the expected delays vary in the range of 1 to 1,000 seconds.

Given $Q_i(t)$ and a market order arrival rate $\mu_i > 0$, the expected delay in exchange i is

$$\mathsf{ED}_i(t) \triangleq \frac{Q_i(t)}{\mu_i}. \tag{1}$$

The μ_i's are assumed to be known, and, indeed, in practice, they can be approximated by observing recent real-time trading activity at each exchange. When the investor decides not to place a limit order but instead trade with a market order, the order is immediately executed and $\mathsf{ED}_0 \triangleq 0$.

2.1.2. Rebates. Exchanges provide a monetary incentive to add liquidity by providing rebates for each limit order that is executed. Over time, these have varied by exchange from −\$0.0010 (a negative liquidity rebate is, in fact, a fee charged to liquidity providers) to \$0.0030 per share traded. As mentioned earlier, they are significant in magnitude compared with the bid-ask spread of a typical liquid stock of \$0.01 per share and represent an important part of the trading costs that influence the order-routing decisions. All things being equal, investors prefer higher rebates.

We denote the liquidity rebate of exchange i by r_i. In the case where the investor chooses to take liquidity ($i = 0$), a market order will, relative to a limit order, involve both paying the bid-offer spread and paying a liquidity-taking fee. The sum of these payments is denoted by $r_0 < 0$.

In practice, order-placement decisions depend on various factors in addition to the ones described previously. For example, an investor may have explicit views on the short-term movement of prices (short-term alpha), and these can be relevant for the placement of limit orders, be sensitive to adverse selection, or the anticipated price movement after the execution of a limit order. To maintain tractability, we will focus on the direct tradeoff between financial benefits and delays. We will denote the financial benefit per share traded associated with exchange i by $\tilde{r}_i$ and refer to it as the *effective rebate*; this includes the direct exchange rebate but possibly incorporates other financial considerations. All else being equal, a higher effective rebate is preferable.

We denote the opportunity set of effective rebate and delay pairs encountered by an investor arriving at time t by $\mathscr{E}(t) \triangleq \{(\tilde{r}_i, \mathsf{ED}_i(t)) : 0 \le i \le N\}$. Investors are heterogeneous with respect to their way of trading off rebate against delay. Each investor is characterized by its type, denoted by $\gamma \ge 0$, that is assumed to be an independent identically distributed (i.i.d.) draw from a cumulative distribution function $F(\cdot)$, which is differentiable, has a continuous density function, and selects a routing decision $i^*(\gamma)$ to maximize his *utility* according to the rule[8]:

$$i^*(\gamma) \in \operatorname*{argmax}_{i \in \{0,1,\ldots,N\}} \gamma\tilde{r}_i - \mathsf{ED}_i(t). \tag{2}$$

In other words, γ is a tradeoff coefficient between price and delay, with units of time per dollar, that characterizes the type of the heterogeneous investors. Given the range of rebates and expected delays, this tradeoff coefficient should roughly be in the range of 1 to 10^4 seconds per \$0.01. Heterogeneity in γ across investors is an important feature of our model, which captures the practical reality that investors differ in their urgency to execute their orders, which, in turn, affects their patience and limit order placement behavior, implicitly or explicitly (through their choice of algorithmic trading strategy and associated parameters; cf. Endnote 4). If at the time of an order arrival the prevailing delays are long, then some investors may choose not to post a limit order altogether but instead cross the spread and execute their order with a market order (the exchange designated 0 in our model).

An equivalent formulation to (2), commonly used in the economic analysis of queues, is to convert the delay into a monetary cost by multiplying it with a delay sensitivity parameter. However, another alternative interpretation would assume that investors

differ in terms of their expected delay tolerance (i.e., the maximum length of time they are willing to wait for an order to be filled). Given an estimate of the anticipated delays, investors with relatively longer delay tolerance would try to place orders in high rebate exchanges, whereas others with shorter delay tolerance would sacrifice high rebates and place their orders in exchanges with shorter delays (and lower rebates) to maximize the probability that their order will get filled in time. Such a reformulation of (2) would still involve a fundamental tradeoff between monetary rebate weighed against measures of delay or execution risk. Overall, although (2) is a simplified criterion, it captures the fundamental tradeoff between time and money, and it will ultimately yield structural results that are consistent with our empirical analysis.

2.2. Market Order Routing

Investors arrive to the market continuously at an aggregate rate $\mu > 0$, seeking to sell an infinitesimal quantity of stock instantaneously via a market order. For an investor who arrives to the market at time t when the queue length vector is $Q(t)$, the routing decision is restricted to the set of exchanges $\{i : Q_i(t) > 0\}$. One important factor influencing this decision is that each exchange charges a fee for taking liquidity, and these fees vary across exchanges. Typically, the fee at an exchange is slightly higher than the rebate, and the exchange pockets the difference as a profit. Fee and rebate data are given in Section 4. For the purposes of this discussion, we assume that the fee on exchange i is equal to the rebate r_i. Because a market order executes without any delay, it is natural to route it to exchange i^* to minimize the fee paid:

$$i^* \in \operatorname*{argmin}_{i \in \{1,\ldots,N\}} \{r_i : Q_i(t) > 0\}. \tag{3}$$

In practice, routing decisions may differ from those predicted by fee minimization for a number of reasons: (a) Real order sizes are not infinitesimal, and to trade a significant quantity, one may need to split an order across many exchanges. (b) If an investor observes that liquidity is available at an exchange, because of latency in receiving market data information or in transmitting the market order to the exchange, that liquidity may no longer be present by the time the investor's market order reaches the exchange. This is accentuated if there are only a few limit orders posted at an exchange. Both (a) and (b) create a preference for longer queue lengths. (c) If an exchange has little available liquidity, *clearing* the queue of resting limit orders is likely to result in greater price impact. (d) There may be other considerations involved in the order-routing decision, such as different economic incentives between the agent making the order-routing decision and the end investor. All these effects point to a more nuanced decision process than the fee minimization suggested by (3), which we will capture through a reduced form *attraction* model that is often used in marketing to capture consumer choice behavior. Specifically, given $Q(t)$, the instantaneous rate at which market orders to sell arrive at exchange i is denoted by $\mu_i(Q(t))$ given by

$$\mu_i(Q(t)) \triangleq \mu \frac{f_i(Q_i(t))}{\sum_{j=1}^N f_j(Q_j(t))}. \tag{4}$$

Equation (4) specifies that the fraction of the total order flow μ that goes to exchange i is proportional to the attraction function $f_i(Q_i(t))$, with $f_i(0) = 0$, that is, market orders will not route to an exchange i with no liquidity. The previous discussion suggests that $f_i(\cdot)$ is an increasing function of the queue length Q_i and a decreasing function of the size of the fee charged by the exchange.

In the remainder of this paper, we use a basic linear model of attraction that specifies

$$f_i(Q_i) \triangleq \beta_i Q_i, \tag{5}$$

where $\beta_i > 0$ is a coefficient that captures the attraction of exchange i per unit of available liquidity. We posit (but our model does not require) that the β_i's be ordered inversely to the fees of the corresponding exchanges. We will revisit this empirically in Section 4.

2.3. Fluid Model

The deterministic fluid model equations are the following: for each exchange i,

$$\begin{aligned} Q_i(t) = Q_i(0) + \lambda_i t + \Lambda \int_0^t \chi_i(Q(s))\, ds \\ - \int_0^t \mu_i(Q(s))\, ds. \end{aligned} \tag{6}$$

The quantity $\chi_i(Q(\cdot))$ denotes the instantaneous fraction of arriving limit orders that are placed into exchange i, defined as

$$\chi_i(Q(t)) \triangleq \int_{\mathcal{G}_i(Q(t))} dF(\gamma), \tag{7}$$

where $\mathcal{G}_i(Q(t))$ denotes the set of optimizing limit order investor types γ that would prefer exchange i, that is, the set of all $\gamma \geq 0$ with $\gamma \tilde{r}_i - \mathsf{ED}_i(t) \geq \gamma \tilde{r}_j - \mathsf{ED}_j(t)$ for all $j \neq i$, given the expected delays $\mathsf{ED}_0(t) = 0$ and $\mathsf{ED}_j(t) = Q_j(t)/\mu_j(Q(t))$, for $j = 1, \ldots, N$, implied[9] by $Q(t)$.

3. Equilibrium Analysis

Suppose that at some point in time, a high rebate exchange has a very short expected delay relative to other exchanges. Then, the routing logic in (2) will direct many arriving limit orders toward this exchange, increasing delays and erasing its relative advantage over the other exchanges. This type of argument suggests that queue lengths will evolve over time and eventually converge into some equilibrium configuration where no exchange seems to have a relative advantage with respect to its rebate/delay tradeoff, taking into account the investors' heterogeneous preferences and the differences in the fees and rebates across exchanges.

Expressing (6) in differential form, we have that $\dot{Q}_i(t) = \lambda_i + \Lambda\chi_i(Q(t)) - \mu_i(Q(t))$, for $i = 1,\ldots,N$. Denoting such an equilibrium queue length vector by Q^*, we have that

$$\lambda_i + \Lambda\chi_i(Q^*) = \mu_i(Q^*), \quad i = 1,\ldots,N. \tag{8}$$

These equations are coupled through the market order rates $\mu_i(Q^*)$ and the aggregated routing decisions given by $\chi_i(Q^*)$ that take into account investor heterogeneity.

3.1. Equilibrium Definition

For each possible price-delay tradeoff coefficient $\gamma \geq 0$, $\pi_i(\gamma)$ denotes the fraction[10] of type γ investors who post limit orders to an exchange if $i \in \{1,\ldots,N\}$ or choose to use a market order if $i = 0$. We require that the routing decision vector $\pi(\gamma) \triangleq (\pi_0(\gamma), \pi_1(\gamma),\ldots,\pi_N(\gamma))$ satisfies

$$\pi_i(\gamma) \geq 0, \quad \forall\, i \in \{0,1,\ldots,N\}; \qquad \sum_{i=0}^{N} \pi_i(\gamma) = 1. \tag{9}$$

Denote by $\pi \triangleq (\pi_i(\gamma))_{\gamma\in\mathbb{R}_+}$ a set of routing decisions across all investor types and let $\mathcal{P}$ denote the set of all π where $\pi(\gamma)$ is feasible for (9), for all $\gamma \geq 0$, and where each $\pi_i(\cdot)$ is a measurable function over $\mathbb{R}_+$. We have suppressed the dependence of π on the rate parameters (λ, Λ, μ) and the queue length vector. We propose the following definition of equilibrium:

Definition 1 (Equilibrium). *An equilibrium $(\pi^*, Q^*) \in \mathcal{P} \times \mathbb{R}_+^N$ is a set of routing decisions and queue lengths that satisfies the following:*

(i) *Individual Rationality: For all $\gamma \geq 0$, the routing decision $\pi^*(\gamma)$ for type γ investors is an optimal solution for*

$$\begin{aligned} \underset{\pi(\gamma)}{\text{maximize}} \quad & \pi_0(\gamma)\gamma\tilde{r}_0 + \sum_{i=1}^{N} \pi_i(\gamma)\left(\gamma\tilde{r}_i - \frac{Q_i^*}{\mu_i(Q^*)}\right) \\ \text{subject to} \quad & \pi_i(\gamma) \geq 0, \quad \forall\, i \in \{0,1,\ldots,N\}; \\ & \sum_{i=0}^{N} \pi_i(\gamma) = 1. \end{aligned} \tag{10}$$

(ii) *Flow Balance: For each exchange $i \in \{1,\ldots,N\}$, the total flow of arriving market orders equals the flow of arriving limit orders, that is,*

$$\lambda_i + \Lambda \int_0^\infty \pi_i^*(\gamma)\, dF(\gamma) = \mu_i(Q^*). \tag{11}$$

Assuming that queue lengths are constant and given by Q^*, the expected delay on each exchange i is given by $Q_i^*/\mu_i(Q^*)$. The individual rationality condition (i) ensures that limit orders are routed in a way that is consistent with (2). The flow balance condition, (ii), ensures that inflows and outflows at each exchange are balanced and that the queue length vector Q^* remains stationary. Definition 1 is consistent[11] with the informal system of Equations (8) because $\chi_i(Q^*) = \int_0^\infty \pi_i^*(\gamma)\, dF(\gamma)$.

3.2. State Space Collapse

Given a vector of queue lengths Q, define the *workload* to be the scaled sum of queue lengths given by $W \triangleq \sum_{i=1}^N \beta_i Q_i$. The workload captures the aggregate market depth across all exchanges, weighted by the attractiveness of each exchange. Orders queued at attractive exchanges (high β_i, typically corresponding to low $\tilde{r}_i$) are weighted more because these orders have greater priority to get filled first, and, therefore, more greatly impact the delays experienced by arriving limit orders at all exchanges. In fact, from (1) and (4), the expected delay on exchange i is given by

$$\mathsf{ED}_i = \frac{W}{\mu\beta_i}. \tag{12}$$

That is, the one-dimensional workload is sufficient to determine delays at every exchange. Theorem 1 establishes something stronger: in equilibrium, the queue length vector Q^*, which is the state of the N-dimensional system, can be inferred from the equilibrium workload W^*. This is a notion of *state space collapse*.

Theorem 1 (State Space Collapse). *Suppose that the pair $(\pi^*, W^*) \in \mathcal{P} \times \mathbb{R}_+$ satisfy*

(i) *π^* is an optimal solution for*

$$\begin{aligned} \underset{\pi}{\text{maximize}} \quad & \int_0^\infty \left\{ \pi_0(\gamma)\gamma\tilde{r}_0 + \sum_{i=1}^{N} \pi_i(\gamma) \right. \\ & \qquad \left. \times \left(\gamma\tilde{r}_i - \frac{W^*}{\mu\beta_i}\right)\right\} dF(\gamma) \\ \text{subject to} \quad & \pi_i(\gamma) \geq 0, \quad \forall\, i \in \{0,1,\ldots,N\}, \forall\, \gamma \geq 0, \\ & \sum_{i=0}^{N} \pi_i(\gamma) = 1, \quad \forall\, \gamma \geq 0. \end{aligned} \tag{13}$$

(ii) π^* *satisfies*

$$\sum_{i=1}^{N}\left(\lambda_i + \Lambda\int_0^\infty \pi_i^*(\gamma)\,dF(\gamma)\right) = \mu. \tag{14}$$

Then, (π^*, Q^*) *is an equilibrium, where for each exchange* $i \neq 0$, Q^* *is defined by*

$$Q_i^* \triangleq \left(\lambda_i + \Lambda\int_0^\infty \pi_i^*(\gamma)\,dF(\gamma)\right)\frac{W^*}{\mu\beta_i}. \tag{15}$$

Conversely, if (π^*, Q^*) *is an equilibrium, define* $W^* \triangleq \beta^\top Q^*$. *Then,* (π^*, W^*) *satisfies (i) and (ii).*

Proof. Suppose that (π^*, W^*) satisfies (i) and (ii). For Q^* given by (15), we have that

$$\beta^\top Q^* = \sum_{i\neq 0}\frac{W^*}{\mu}\left(\lambda_i + \Lambda\int_0^\infty \pi_i^*(\gamma)\,dF(\gamma)\right) = W^*.$$

Thus,

$$\frac{W^*}{\mu\beta_i} = \frac{\beta^\top Q^*}{\mu\beta_i} = \frac{Q_i^*}{\mu_i(Q^*)}. \tag{16}$$

Combining this with the fact that optimization problem in (i) is separable with respect to γ (i.e., it can be optimized over each $\pi(\gamma)$ separately), it is clear that (π^*, Q^*) satisfies the individual rationality condition (10). Furthermore, rewriting (15),

$$\lambda_i + \Lambda\int_0^\infty \pi_i^*(\gamma)\,dF(\gamma) = \mu\frac{\beta_i Q_i^*}{W^*} = \mu\frac{\beta_i Q_i^*}{\beta^\top Q^*} = \mu_i(Q^*).$$

Thus, (π^*, Q^*) satisfies the flow balance condition (11), and (π^*, Q^*) is an equilibrium.

For the converse, suppose that (π^*, Q^*) is an equilibrium and $W^* \triangleq \beta^\top Q^*$. Then,

$$\frac{W^*}{\mu\beta_i} = \frac{\beta^\top Q^*}{\mu\beta_i} = \frac{Q_i^*}{\mu_i(Q^*)}.$$

Given that (π^*, Q^*) satisfies (10), this implies that (π^*, W^*) satisfies (i). Furthermore, if we sum up all N equations in (11), it is clear that (π^*, W^*) satisfy (ii). □

Condition (i) of Theorem 1 implies individual rationality when faced with delays implied by the workload W^* (cf. (10) and (12)). Condition (ii) is a market-wide flow balance equation. Given a pair (π^*, W^*) satisfying (i) and (ii), Q^* is determined as a function of workload W^* through the *lifting map* (15) that distributes the workload across exchanges in a way that takes into account rebates, delays, and investor heterogeneity through the distribution $F(\cdot)$ of the tradeoff coefficient γ. The lifting map corresponds to Little's law: each queue length is equal to the corresponding aggregate arrival rate (dedicated and optimized) times the equilibrium expected delay.

3.3. Equilibrium Characterization

Theorem 1 allows us to characterize the equilibrium behavior of N decentralized limit order books through their one-dimensional workload. The following assumption will turn out to be sufficient for the existence of an equilibrium.

Assumption 1. *Assume that*

(i) *The cumulative distribution function* $F(\cdot)$ *over the price-delay tradeoff coefficients* γ *is nonatomic with a continuous and strictly positive density on the nonnegative reals.*

(ii) *The arrival rates* (λ, Λ, μ) *satisfy* $\sum_{i=1}^N \lambda_i < \mu < \Lambda + \sum_{i=1}^N \lambda_i$.

(iii) *Each exchange* $i \in \{1, \ldots, N\}$ *satisfies* $\tilde{r}_i > \tilde{r}_0$.

The dedicated flow $\sum_{i=1}^N \lambda_i$ is not delay sensitive. Condition (iii) ensures that the queueing system is stable ($\sum_{i=1}^N \lambda_i < \mu$) and leads to a nontrivial equilibrium where queue lengths are nonzero ($\mu < \Lambda + \sum_{i=1}^N \lambda_i$). Condition (iii) says that if delays are zero, then the effective rebate of a limit order is always preferable to the cost of crossing the spread and paying a fee to trade with a market order, $\tilde{r}_0$. Returning to condition (ii), given that $\mu < \Lambda + \sum_{i=1}^N \lambda_i$, one would expect nonzero queue lengths to build up in the system to discourage some optimizing investors from placing a limit order and instead trade with a market order. Intuitively, one expects this to be the most impatient investors, that is, those of type $\gamma \le \gamma_0$, for some γ_0, chosen to satisfy (14),

$$\Lambda\big(1 - F(\gamma_0)\big) + \sum_{i=1}^N \lambda_i = \mu. \tag{17}$$

Under conditions (i) and (ii) of Assumption 1, γ_0 satisfying (17) exists and is uniquely determined by

$$\gamma_0 \triangleq F^{-1}\left(1 - \frac{\mu - \sum_{i=1}^N \lambda_i}{\Lambda}\right). \tag{18}$$

For all types $\gamma \le \gamma_0$ not to submit limit orders, the routing criterion (2) requires that

$$\max_{i\neq 0}\ \gamma(\tilde{r}_i - \tilde{r}_0) - \frac{W^*}{\mu\beta_i} \le 0, \tag{19}$$

for all $\gamma \le \gamma_0$. Under Assumption 1(iii), the left side of (19) is increasing in γ. Hence, (19) is satisfied if we ensure that type γ_0 investors are indifferent between market orders and limit orders.

Lemma 1. *Under Assumption 1, suppose that* (π^*, W^*) *is an equilibrium and define* γ_0 *by* (18). *Then,*

$$\max_{i\neq 0}\ \gamma_0(\tilde{r}_i - \tilde{r}_0) - \frac{W^*}{\mu\beta_i} = 0. \tag{20}$$

Furthermore, suppose that for a given W^*, (20) holds, and for each exchange i, define

$$\kappa_i \triangleq \beta_i(\tilde{r}_i - \tilde{r}_0). \tag{21}$$

Then, an exchange i achieves the maximum in (20) if and only if $i \in \operatorname{argmax}_{j\neq 0}\ \kappa_j$.

(The proof of Lemma 1 is provided in the online supplement.) The quantity κ_i is related to the desirability of exchange i from the perspective of a limit order investor; κ_i is high when β_i is high (resulting in low delay) or when $\tilde{r}_i$ is high (resulting in a high rebate). Lemma 1 suggests that maximizing κ_i characterizes the behavior of type γ_0 (the marginal) investors that are indifferent between choosing between a market order and a limit order. We refer to exchanges that achieve this maximum as marginal exchanges. Thus, given a marginal exchange $\bar{\imath} \in \operatorname{argmax}_{j\neq 0} \kappa_j$, according to Lemma 1,

$$\gamma_0(\tilde{r}_{\bar{\imath}} - \tilde{r}_0) - \frac{W^*}{\mu\beta_{\bar{\imath}}} = 0,$$

and therefore the equilibrium workload is $W^* = \gamma_0\mu\kappa_{\bar{\imath}}$. Theorem 2, whose proof can be found in the online supplement, summarizes the previous discussion and characterizes the equilibrium.

Theorem 2 (Equilibrium Characterization). *Under Assumption 1, define γ_0 by (18). Suppose that the pair $(\pi^*, W^*) \in \mathcal{P} \times \mathbb{R}_+$ satisfy*

$$W^* \triangleq \gamma_0\mu \max_{i\neq 0}\ \kappa_i, \tag{22}$$

and

$$\begin{aligned} \pi_0^*(\gamma) &= 1, \quad \text{for all } \gamma < \gamma_0, \\ \pi_i^*(\gamma_0) &= 0, \quad \text{for all } i \notin \mathcal{A}^*(\gamma_0) \cup \{0\}, \\ \pi_i^*(\gamma) &= 0, \quad \text{for all } \gamma > \gamma_0,\ i \notin \mathcal{A}^*(\gamma), \end{aligned} \tag{23}$$

where $\mathcal{A}^*(\gamma) \triangleq \operatorname{argmax}_{i\neq 0}\ \gamma\tilde{r}_i - W^*/\mu\beta_i$. Then, (π^*, W^*) is an equilibrium, that is, satisfies (13) and (14).

Conversely, suppose that $(\pi^*, W^*) \in \mathcal{P} \times \mathbb{R}_+$ is an equilibrium, that is, satisfies (13) and (14). Then, W^* must satisfy (22) and π^* must satisfy (23), except possibly for γ in a set of F-measure zero.

This characterization of the workload process and its dependence on model parameters can be used as a point of departure to analyze market structure and market design issues and competition and welfare implications of the presence of many differentiated exchanges. Theorem 2 implies that the equilibrium workload is unique and that equilibrium routing policies are unique up to ties.

We can establish uniqueness of the equilibrium queue length vector Q^* in the Theorem 3 (its proof is available in the online supplement), under the following mild assumption:

Assumption 2. *Assume that the effective rebates $\{\tilde{r}_i,\ i \neq 0\}$ are distinct and, without loss of generality, that the exchanges are labeled in an increasing order, that is, $\tilde{r}_0 < \tilde{r}_1 < \cdots < \tilde{r}_N$.*

Theorem 3 (Uniqueness of Equilibria). *Under Assumptions 1 and 2, there is a unique equilibrium queue length vector Q^*.*

In the online supplement, we consider the question of whether the fluid model queue length vector $Q(t)$ converges to the unique equilibrium vector Q^* as $t \to \infty$. For $N = 2$ exchanges, we use a geometric argument to prove that the fluid model transient starting from an arbitrary initial condition converges to the equilibrium state in finite time. We conjecture that a similar argument carries through when there are $N > 2$ exchanges.

3.4. Discussion

The state-space collapse result and its functional form hinge on the formulation of the order-routing models described in Sections 2.1 and 2.2. The primary drivers of the dimension reduction are (a) the desirability to place an order at a given queue is decreasing in its anticipated delay and (b) that the attractiveness of an exchange for an incoming market order is increasing in its queue length. Both drivers seem plausible even under different models of order-routing optimization logic on both sides of the market, and one might expect these to lead to some form of state-space collapse: long queues would discourage new orders from joining while attracting more service completions, thus reducing queue size; small queues would attract more arrivals but fewer service completions, thus increasing queue size. For example, the same rationale holds if we replace the market order-routing model (4) with a model of the form $\mu_i(Q) \triangleq M_i + f_i(Q)$, for each exchange i. Here, each $M_i \geq 0$ represents *dedicated* market order flow to exchange i that does not react to the state of the system, whereas the $f_i(Q)$ term captures optimized order flow. The detailed form of the equilibrium of such a system would not coincide with the one derived here; however, at a high level, one would expect similar results under different modeling assumptions that satisfy (a) and (b).

4. Empirical Results

Motivated by our analysis and the fact that for liquid securities the markets experience high volumes of flow per unit time, one would expect the market to behave as if it is near its equilibrium state most of the time, which would manifest itself as a strong coupling between the quote depths and dynamics of competing

exchanges. More precisely, the expected delay trajectories across exchanges and over time should exhibit strong linear relationships and behave like a lower dimensional process. Our model suggests the coupling of the dynamics across exchanges should be best explained through the relation of the respective expected delays as opposed to the queue lengths themselves. The expected delay in exchange i is of the form $Q_i/\mu_i(Q)$, from which we see that the queue length affects the delay in a nonlinear way that should likely result in a worse fit in the data according to our model. Moreover, the workload process (a measure of weighted aggregate depth) should offer accurate estimates of delays and queue depths at different exchanges, as stated in (12).

The precise form of our predictions is, of course, predicated on the structure of (2), (4), and (5) and the deterministic and stationary nature of the model we studied. In the sequel, we will explore whether our predictions are supported through empirical evidence from a representative sample of market data that incorporates actual trading behaviors that are more complex, and its dynamics are stochastic and nonstationary. We will also examine data of long periods of time, thus empirically exploring the SSC predictions in a pathwise sense (see the short discussion in Section 1 earlier and in Section A of the online supplement).

The first few subsections will estimate our model primitives and empirically explore the predicted SSC result on a data set for all 30 constituent stocks of the Dow Jones Industrial Average (DJIA) over the duration of a one-month period in 2011. In Section 4.4, we will explore a more recent sample of data in the beginning of 2015 where the NASDAQ runs a natural experiment of reducing the rebates and fees for a sample of stocks. In that context, we will verify the validity of our model predictions around this exogenous parameter change and illustrate how our model could prove useful in studying such market design and policy questions.

4.1. Overview of the Data Set

We use TAQ data, which consists of sequences of quotes (price and total available size, expressed in number of shares, at the best bid and offer on each exchange) and trades (price and size of all market transactions, again expressed in number of shares), with millisecond timestamps. Our trade and quote data are from the nationally consolidated data feeds. We treat the depth at the bid or the ask at each exchange as if it is made up of individual infinitesimal orders, and we ignore the fact that the quantity actually arises from a collection of discrete, noninfinitesimal orders.

We consider the 30 component stocks of the DJIA over the 21 trading days in the month of September 2011. A list of the stocks and some basic descriptive statistics are given in Table 1. In Section 4.4, we will study a more recent, different data set.

We restrict attention to the $N = 6$ most liquid U.S. equity exchanges: NASDAQ, New York Stock Exchange (NYSE),[12] ARCA, DirectEdge X (EDGX), BATS, and DirectEdge A (EDGA). Smaller, regional exchanges were excluded because they account for a small fraction of the composite daily volume and are often not quoting at the NBBO level. The associated fees and rebates during the observation period of September 2011 are given in Table 2.

Throughout the observation period of our data set, the exchange fees and rebates were constant, and similarly we will assume in our subsequent analysis that the effective rebates $\{\tilde{r}_i\}$ and attraction coefficients $\{\beta_i\}$ for each stock were also constant throughout.

In contrast, the arrival rates (λ, Λ, μ) are time varying. We will estimate these rates for each stock by averaging the event activity over one-hour time intervals between 9:45 a.m. and 3:45 p.m. (i.e., excluding the opening 15 minutes and the closing 15 minutes).[13]

This yields $T = 126$ time slots over the 21-day horizon of our data set. For each time slot t, exchange i, stock j, and side $s \in \{\mathsf{BID}, \mathsf{ASK}\}$, we estimated the corresponding queue length as the average number of shares available at the NBBO, denoted by $Q_i^{(s,j)}(t)$. Similarly, denote by $\mu_i^{(s,j)}(t)$ the arrival rate of market orders to side s on exchange i for security j, in time slot t. The rates $\mu_i^{s,j}(t)$ are estimated by classifying trades to be bid or ask side of the market, by matching trade time stamps with the prevailing quote at the same time, that is, using a zero time shift in the context of the well-known Lee-Ready algorithm. Given these parameters, we compute the following measure of expected delay:

$$\mathsf{ED}_i^{(s,j)}(t) \triangleq \frac{Q_i^{(s,j)}(t)}{\mu_i^{(s,j)}(t)}. \tag{24}$$

This expression disregards the effect of order cancellations from the bid and ask queues, as well as the noninfinitesimal nature of the order flow (compare with the remarks in Endnote 6 on cancellations). It serves as a practical proxy for expected delay that is commonly used in trading systems. For each stock and each exchange, Figure 2(a) shows the expected delay, averaged across time slots and the bid and ask sides of the market. Delays range from five seconds to about five minutes across the 30 stocks we studied, and we observe two to three times the variation in the delay estimates at different exchanges for the same security. Similarly, for each stock and each exchange, Figure 2(b) shows the average queue lengths, or the number of shares available at the NBBO, averaged across time slots and the bid and ask sides of the

Table 1. Descriptive Statistics for the 30 Stocks over the 21 Trading Days of September 2011

	Symbol	Listing exchange	Price		Average bid-ask spread ($)	Volatility (daily)	Average daily volume(shares, × 10^6)
			Low ($)	High ($)			
Alcoa	AA	NYSE	9.56	12.88	0.0099	2.2%	27.8
American Express	AXP	NYSE	44.87	50.53	0.0135	1.9%	8.6
Boeing	BA	NYSE	57.53	67.73	0.0170	1.8%	5.9
Bank of America	BAC	NYSE	6.00	8.18	0.0098	3.0%	258.8
Caterpillar	CAT	NYSE	72.60	92.83	0.0286	2.3%	11.0
Cisco	CSCO	NASDAQ	14.96	16.84	0.0098	1.7%	64.5
Chevron	CVX	NYSE	88.56	100.58	0.0181	1.7%	11.1
DuPont	DD	NYSE	39.94	48.86	0.0110	1.7%	10.2
Disney	DIS	NYSE	29.05	34.33	0.0102	1.6%	13.3
General Electric	GE	NYSE	14.72	16.45	0.0098	1.9%	84.6
Home Depot	HD	NYSE	31.08	35.33	0.0101	1.6%	13.4
Hewlett-Packard	HPQ	NYSE	21.50	26.46	0.0099	2.2%	32.5
IBM	IBM	NYSE	158.76	180.91	0.0596	1.5%	6.6
Intel	INTC	NASDAQ	19.16	22.98	0.0097	1.5%	63.6
Johnson & Johnson	JNJ	NYSE	61.00	66.14	0.0114	1.2%	12.6
JPMorgan	JPM	NYSE	28.53	37.82	0.0099	2.2%	49.1
Kraft	KFT	NYSE	32.70	35.52	0.0100	1.1%	10.9
Coca-Cola	KO	NYSE	66.62	71.77	0.0108	1.1%	12.3
McDonalds	MCD	NYSE	83.65	91.09	0.0135	1.2%	7.9
3M	MMM	NYSE	71.71	83.95	0.0181	1.6%	5.5
Merck	MRK	NYSE	30.71	33.49	0.0098	1.3%	17.6
Microsoft	MSFT	NASDAQ	24.60	27.50	0.0097	1.5%	61.0
Pfizer	PFE	NYSE	17.30	19.15	0.0099	1.5%	47.7
Procter & Gamble	PG	NYSE	60.30	64.70	0.0107	1.0%	11.2
AT&T	T	NYSE	27.29	29.18	0.0099	1.2%	37.6
Travelers	TRV	NYSE	46.64	51.54	0.0128	1.6%	4.8
United Tech	UTX	NYSE	67.32	77.58	0.0182	1.7%	6.2
Verizon	VZ	NYSE	34.65	37.39	0.0099	1.2%	18.4
Wal-Mart	WMT	NYSE	49.94	53.55	0.0103	1.1%	13.1
Exxon Mobil	XOM	NYSE	67.93	74.98	0.0109	1.6%	26.2

Notes. The average bid-ask spread is a time average computed from our TAQ data set. The volatility is an average of daily volatilities over September 2011. All the other statistics were retrieved from Yahoo Finance.

market. Queue lengths range from 10 to 100,000 shares across securities, and exhibit about a 10 times variation in the queue sizes across exchanges for the same security. Deeper queues correspond to longer delays.

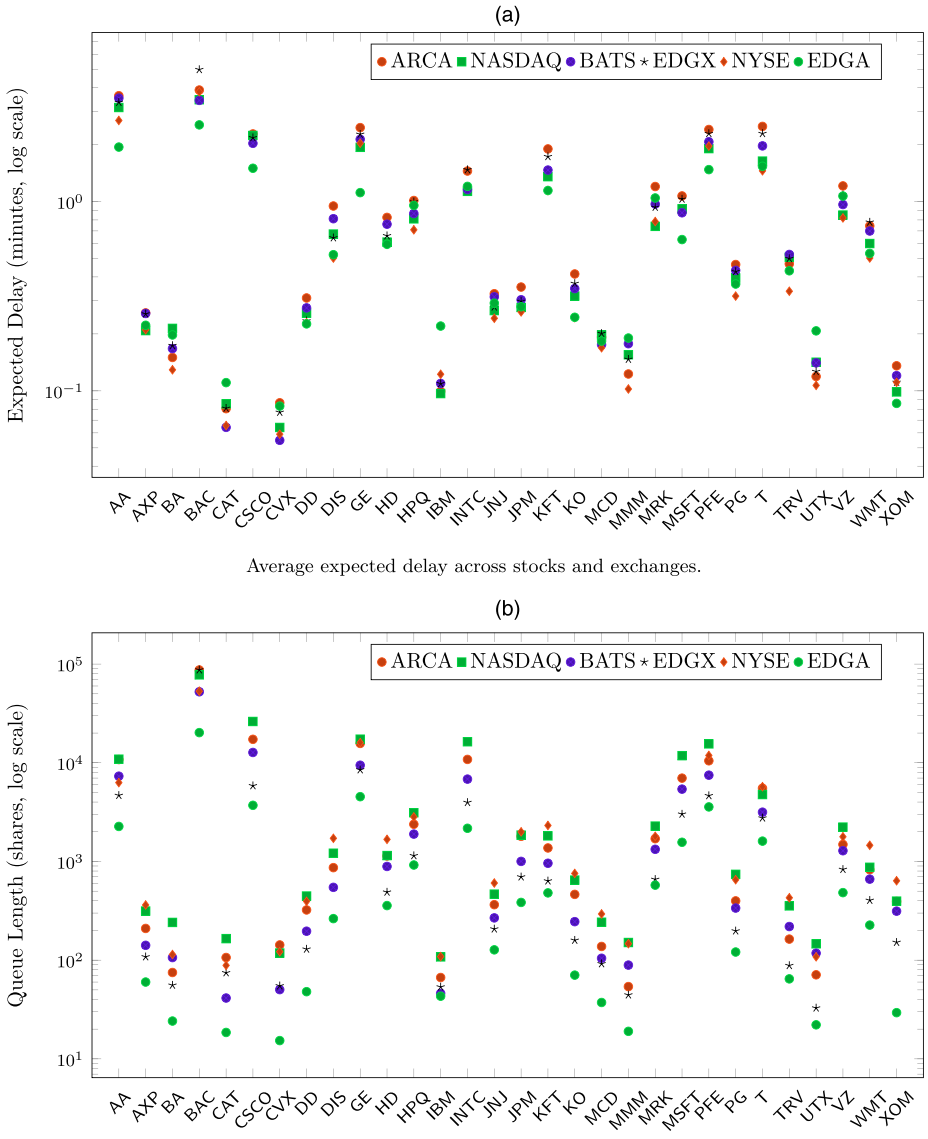
4.1.1. Principal Component Analysis. The state-space collapse result of our model predicts that delays are coupled across exchanges and are restricted to a one-dimensional subspace. Define the empirically observed expected delay vector trajectories $\{\mathsf{ED}^{(s,j)}(t) : t = 1, \ldots, T;\ s = \mathsf{BID}, \mathsf{ASK}\}$, where $\mathsf{ED}^{(s,j)}(t)$ was estimated in (24), and the trajectories consider all one hour time slots in the 21 days of our observation period. A natural way to test the effective dimensionality of this vector of trajectories is via principle component analysis (PCA) by examining the number of principle components necessary to explain the

Table 2. Rebates and Fees of the Six Major U.S. Stock Exchanges During the September 2011 Period per Share Traded

	Exchange code	Rebate ($ per share, × 10^{-4})	Fee ($ per share, × 10^{-4})
BATS	Z	27.0	28.0
DirectEdge X (EDGX)	K	23.0	30.0
ARCA	P	21.0[a]	30.0
NASDAQ OMX	T	20.0[a]	30.0
NYSE	N	17.0	21.0
DirectEdge A (EDGA)	J	5.0	6.0

[a]Rebates on NASDAQ and ARCA are subject to tiering: higher rebates than the ones quoted may be available to traders that contribute significant volume to the respective exchange.

Figure 2. (Color online) Averages of (a) Hourly Estimates of the Expected Delays and (b) Queue Lengths for the Dow 30 Stocks on the Six Exchanges During September 2011



Notes. Results are averaged over the bid and ask sides of the market for each stock. Queues do not include estimates of hidden liquidity at each of the exchanges.

variability of the expected delay trajectories across exchanges and over time. The output of the PCA analysis is summarized in Table 3: the first principle component explains approximately 80% of the variability of the expected delays across exchanges, and the first two principle components explain approximately 90%. This is consistent with the hypothesis of low effective dimension. In contrast, when we conduct PCA for the vector trajectories of observed queue lengths $\{Q^{(s,j)}(t) : t = 1, \ldots, T;\ s = \mathsf{BID}, \mathsf{ASK}\}$, we find relatively weaker evidence for a low effective dimensionality. In this test, the first principle component explains approximately 65% of the variability of the queue lengths across exchanges, and the first two principle components explain less than 80%. A detailed report of the results can be found in Table 19 in the online supplement.

Intuitively, in the high-flow environment of our observation universe, that is, where Λ and μ are large, expected delay deviations from the equilibrium configuration would be quickly erased by optimized arriving limit and market orders. The equilibrium state itself changes over time as the rates of events change, but the coupling across exchanges remains

Table 3. Results of PCA: How Much Variance in the Data Can the First Two Principle Components Explain

	Percentage of variance explained	
	One factor	Two factor
Alcoa	80%	88%
American Express	78%	88%
Boeing	81%	87%
Bank of America	85%	93%
Caterpillar	71%	83%
Cisco	88%	93%
Chevron	78%	87%
DuPont	86%	92%
Disney	87%	91%
General Electric	87%	94%
Home Depot	89%	94%
Hewlett-Packard	87%	92%
IBM	73%	84%
Intel	89%	93%
Johnson & Johnson	87%	91%
JPMorgan	90%	94%
Kraft	86%	92%
Coca-Cola	87%	93%
McDonalds	81%	89%
3M	71%	81%
Merck	83%	91%
Microsoft	87%	95%
Pfizer	83%	89%
Procter & Gamble	85%	92%
AT&T	82%	89%
Travelers	80%	88%
United Tech	75%	88%
Verizon	85%	91%
Wal-Mart	89%	93%
Exxon Mobil	86%	92%

strong and persists even if we shorten the time period over which market statistics are averaged from 1 hour down to 15 minutes.[14]

4.2. Estimation of the Market Order-Routing Model

Define $\mu_i^{(s,j)}(t)$ to be the total arrival rate of market orders for security j and side $s \in \{\mathsf{BID}, \mathsf{ASK}\}$ in time slot t directed to exchange i, and let $\mu^{(s,j)}(t)$ be the total arrival rate across all exchanges for (s,j) in time t. The attraction model of Section 2.2 for market orders suggests the following relationship:

$$\mu_i^{(s,j)}(t) = \mu^{(s,j)}(t)\ \frac{\beta_i^{(j)} Q_i^{(s,j)}(t)}{\sum_{i'=1}^{N} \beta_{i'}^{(j)} Q_{i'}^{(s,j)}(t)}, \tag{25}$$

where $\beta_i^{(j)}$ is the attraction coefficient for security j on exchange i. Our market order-routing model is invariant to scaling of the attraction coefficients; hence, we normalize so that the attraction coefficient for each stock on its listing exchange is 1. Given that $\{\mu_i^{(s,j)}(t)\}$, $\{\mu^{(s,j)}(t)\}$, and $\{Q_i^{(s,j)}(t)\}$ are observable, we estimated the $\beta_i^{(j)}$'s using a nonlinear regression in (25). The results are given in Table 4. All attraction coefficient estimates are statistically significant.

4.3. Empirical Evidence of State-Space Collapse

Our model postulates the investors make order-placement decisions by trading off delay against effective rebates and concludes that delays across exchanges, as measured by $Q_i^{(s,j)}/\mu_i^{(s,j)}$, are linearly related. It gives an expression for estimating delays in each exchange in terms of an aggregate measure of market depth, which we call workload.

4.3.1. Verification of linear dependence of expected delays via regression analysis. Denote by $W^{(s,j)}(t)$ the workload for side s of security j in time slot t, that is,

$$W^{(s,j)}(t) \triangleq \sum_{i=1}^{N} \beta_i^{(j)} Q_i^{(s,j)}(t), \tag{26}$$

and observe that using (25), the vector of expected delays can be written as

$$\mathsf{ED}^{(s,j)}(t) = \frac{W^{(s,j)}(t)}{\mu^{(s,j)}(t)} \left(\frac{1}{\beta_1^{(j)}}, \ldots, \frac{1}{\beta_N^{(j)}} \right). \tag{27}$$

In other words, the expected delays across different exchanges are linearly related, and specifically, for each security j, exchanges i, i', and market side s,

$$\mathsf{ED}_i^{(s,j)}(t) = \frac{\beta_{i'}^{(j)}}{\beta_i^{(j)}}\ \mathsf{ED}_{i'}^{(s,j)}(t), \tag{28}$$

for each time slot t. Testing the pairwise linear relations in (28) explores whether $(\mathsf{ED}_1, \ldots, \mathsf{ED}_N)$ live on a one-dimensional space; this statistical test is based on the expected delay measurements in (24), obtained by dividing the average observed queue size in each exchange with its respective observed rate of trading, for all time slots, both sides of the market, and all the 30 component stocks of the Dow Jones Industrial Average. For each stock and each exchange, we have 126 measurements in the respective time series per side of the market. The quality of the fit of these linear regressions will be an indirect indication of the goodness of fit in (25). In more detail, we will perform a cross-sectional regression. We will normalize the expected delay measurements at each exchange by dividing them by the median expected delay of that security on a benchmark exchange (ARCA) across all time slots and both sides of the market as follows:

$$\overline{\mathsf{ED}}_i^{(s,j)}(t) \triangleq \frac{\mathsf{ED}_i^{(s,j)}(t)}{\underset{\tau=1,\ldots,T;\ s=\mathsf{BID},\mathsf{ASK}}{\mathrm{median}} \left(\mathsf{ED}_{\mathrm{ARCA}}^{(s,j)}(\tau) \right)}, \tag{29}$$

Table 4. Estimates of the Attraction Coefficients β_i from Nonlinear Regression

	Attraction coefficient					
	ARCA	NASDAQ	BATS	EDGX	NYSE	EDGA
Alcoa	0.73	0.87	0.76	0.81	1.00	1.33
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.03)
American Express	1.19	1.08	0.99	0.94	1.00	0.94
	(0.02)	(0.02)	(0.04)	(0.03)	(0.00)	(0.06)
Boeing	0.95	0.67	0.81	0.74	1.00	0.73
	(0.02)	(0.01)	(0.01)	(0.02)	(0.00)	(0.04)
Bank of America	0.94	1.04	1.01	0.77	1.00	1.43
	(0.01)	(0.02)	(0.02)	(0.01)	(0.00)	(0.04)
Caterpillar	0.82	0.78	1.13	0.70	1.00	0.58
	(0.01)	(0.01)	(0.03)	(0.02)	(0.00)	(0.04)
Cisco	0.95	1.00	1.06	0.98	—	1.45
	(0.01)	(0.00)	(0.01)	(0.02)	—	(0.03)
Chevron	0.70	0.93	1.17	0.65	1.00	0.75
	(0.01)	(0.01)	(0.02)	(0.01)	(0.00)	(0.05)
DuPont	0.90	0.98	0.98	1.03	1.00	1.00
	(0.01)	(0.01)	(0.02)	(0.02)	(0.00)	(0.06)
Disney	0.69	0.88	0.78	0.88	1.00	1.04
	(0.01)	(0.01)	(0.02)	(0.03)	(0.00)	(0.03)
General Electric	0.79	1.01	0.94	0.73	1.00	1.63
	(0.01)	(0.01)	(0.02)	(0.01)	(0.00)	(0.03)
Home Depot	0.76	0.98	0.79	0.84	1.00	1.02
	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)	(0.03)
Hewlett-Packard	1.04	1.04	1.02	0.68	1.00	0.82
	(0.02)	(0.01)	(0.02)	(0.02)	(0.00)	(0.03)
IBM	1.25	1.20	1.20	1.05	1.00	0.54
	(0.02)	(0.02)	(0.03)	(0.02)	(0.00)	(0.02)
Intel	0.83	1.00	0.96	0.84	—	1.04
	(0.01)	(0.00)	(0.01)	(0.02)	—	(0.03)
Johnson & Johnson	0.80	0.94	0.86	0.92	1.00	0.77
	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)	(0.03)
JPMorgan	0.78	0.99	0.93	0.84	1.00	0.91
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.02)
Kraft	0.72	0.89	0.83	0.73	1.00	1.06
	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)	(0.03)
Coca-Cola	0.68	0.84	0.79	0.76	1.00	0.88
	(0.01)	(0.01)	(0.02)	(0.02)	(0.00)	(0.05)
McDonalds	0.90	0.86	1.03	0.82	1.00	0.82
	(0.01)	(0.01)	(0.02)	(0.02)	(0.00)	(0.04)
3M	0.89	0.67	0.62	0.66	1.00	0.57
	(0.02)	(0.01)	(0.01)	(0.02)	(0.00)	(0.04)
Merck	0.68	1.01	0.83	0.90	1.00	0.81
	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)	(0.02)
Microsoft	0.83	1.00	1.02	0.95	—	1.41
	(0.01)	(0.00)	(0.01)	(0.02)	—	(0.03)
Pfizer	0.84	1.01	0.96	0.87	1.00	1.29
	(0.01)	(0.01)	(0.02)	(0.02)	(0.00)	(0.03)
Procter & Gamble	0.79	0.89	0.88	0.89	1.00	0.89
	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)	(0.03)
AT&T	0.62	0.94	0.75	0.59	1.00	1.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.03)
Travelers	0.80	0.69	0.69	0.84	1.00	0.80
	(0.01)	(0.01)	(0.01)	(0.03)	(0.00)	(0.03)
United Tech	1.18	0.89	0.79	0.87	1.00	0.53
	(0.02)	(0.01)	(0.01)	(0.03)	(0.00)	(0.04)
Verizon	0.77	0.95	0.88	0.72	1.00	0.85
	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)	(0.03)
Wal-Mart	0.72	0.88	0.79	0.71	1.00	0.91
	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)	(0.03)

Table 4. (Continued)

	Attraction coefficient					
	ARCA	NASDAQ	BATS	EDGX	NYSE	EDGA
Exxon Mobil	0.89 (0.01)	1.13 (0.01)	0.97 (0.01)	0.89 (0.02)	1.00 (0.00)	1.35 (0.10)

Notes. The attraction coefficient of the listing exchange is normalized to be 1. We note that exchanges with lower fees (and rebates) tend to have higher attraction coefficients β.

where $\mathsf{ED}_i^{(s,j)}(t)$ was estimated in (24). We will perform a linear regression of the normalized left side of (28) as a function of the normalized right side of (28), rescaled by the ratio of the attraction coefficients of the two exchanges.

The results of these regressions are summarized in Table 5. The R^2 varies between 52% and 70% across the five exchanges. The results are statistically significant, and we are able to reject the null hypothesis that the delay on a particular exchange has a zero regression coefficient relative to the rescaled delay on ARCA. These results statistically verify the linear dependence of delays across different exchanges suggested by (28). Note that (28) further predicts that the regression should have a zero intercept, and the slope of the rescaled $\mathsf{ED}_{\mathrm{ARCA}}$ term should be 1. These are not born in the regressions—the intercept is statistically different from 0, and the slope is statistically different from 1. Nevertheless, the intercept and slope are, respectively, quite close to 0 and 1. This is remarkable given the stylized nature of the routing model of Section 4.2 and the noise in the extensive market data sample.

Although the regressions in Table 5 were performed cross-sectionally across all securities, similar results hold if the analysis is performed on a security by security basis. Figure 3 depicts the delay relationships in the case of Bank of America. It illustrates the strong linear relationship across all exchanges over time and across significant variations in prevailing market conditions; the latter is manifested in the roughly two orders of magnitude variation in estimated expected delays.

A competing hypothesis is that queue lengths across exchanges are linearly related, that is, for each security j, exchanges i, i', and market side s,

$$Q_i^{(s,j)}(t) = c_{ii'} Q_{i'}^{(s,j)}(t), \tag{30}$$

for each time slot t. The following test explores such an alternative hypothesis. According to (30), predicated on queue length estimates obtained in Section 4.1, that is, $Q_i^{(s,j)}(t)$ as the average number of shares available at the NBBO for time slot t, exchange i, stock j, and side $s \in \{\mathsf{BID}, \mathsf{ASK}\}$, we perform a cross-sectional linear regression of the queue length of each security on a particular exchange, as a function of that on a benchmark exchange (ARCA). As before, we normalize the queue lengths by dividing them by the median queue length of that security on a benchmark exchange (ARCA) across all time slots and both sides of the market; that is, we use

$$\overline{Q}_i^{(s,j)}(t) \triangleq \frac{Q_i^{(s,j)}(t)}{\underset{\tau=1,\ldots,T;\ s=\mathsf{BID},\mathsf{ASK}}{\mathrm{median}} \left(Q_{\mathrm{ARCA}}^{(s,j)}(\tau) \right)}$$

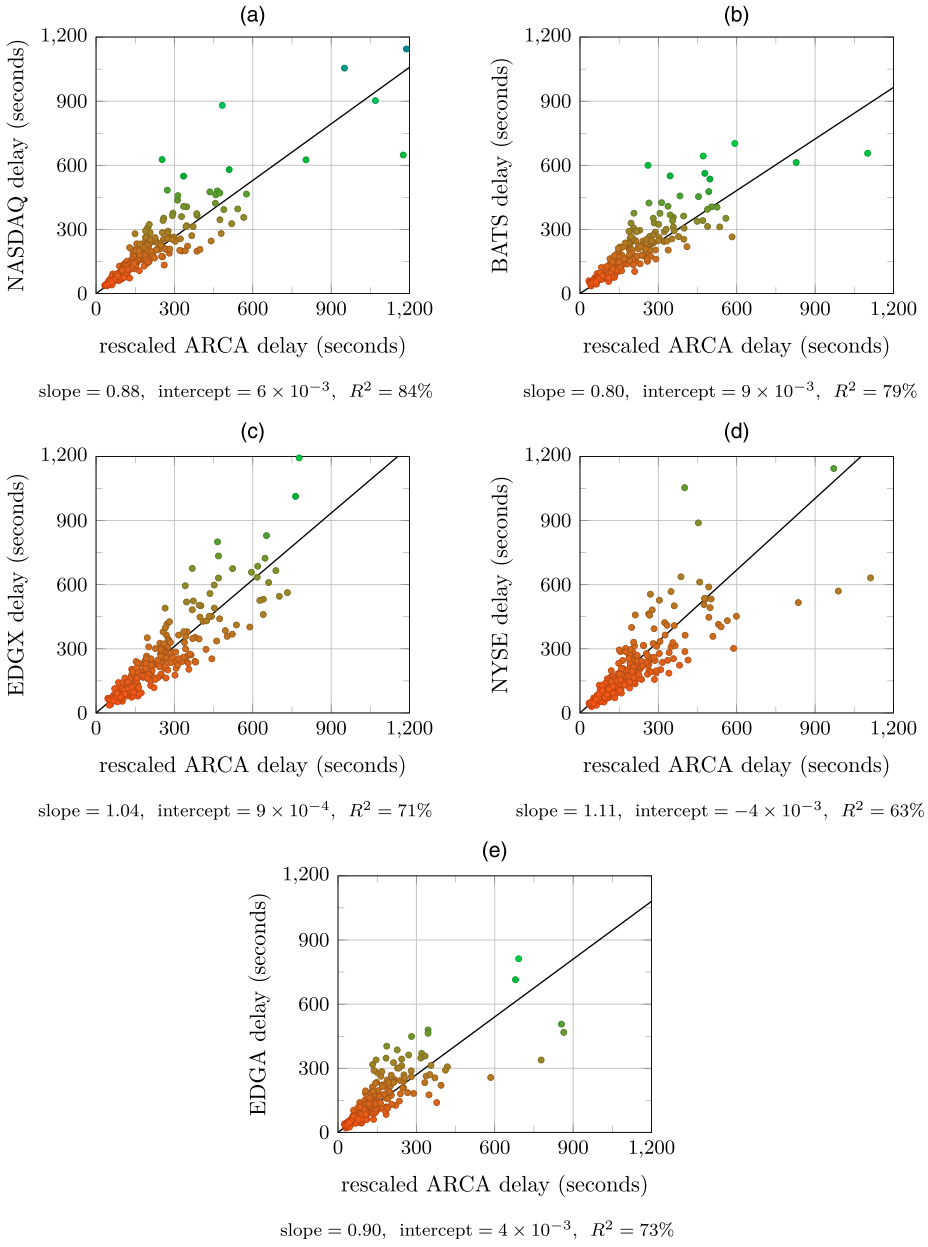
as the queue length measure in the regression. Our model would predict that queue lengths will not exhibit such a strong linear dependence as we show earlier in terms of delays. Indeed, the results provided in Table 6 show that the R^2 we found was significantly

Table 5. Linear Regressions of the Normalized Expected Delay on a Particular Exchange vs. That of the Benchmark Exchange (ARCA) Rescaled by the Ratio of the Attraction Coefficients of the Two Exchanges

	Dependent variable: $\overline{\mathsf{ED}}_{\mathrm{exchange}}$				
	NASDAQ OMX	BATS	DirectEdge X	NYSE	DirectEdge A
Intercept	0.27*** (0.01)	0.28*** (0.01)	0.24*** (0.01)	0.28*** (0.01)	0.36*** (0.01)
Rescaled $\overline{\mathsf{ED}}_{\mathrm{ARCA}}$	0.70*** (0.01)	0.72*** (0.01)	0.72*** (0.01)	0.63*** (0.01)	0.60*** (0.01)
R^2	70%	70%	52%	60%	52%

***$p < 0.01$.

Figure 3. (Color online) Scatter Plots of the Expected Delay for Bank of America (BAC) on Each Exchange vs. the Delay on ARCA Rescaled by the Ratio of the Attraction Coefficients of the Two Exchanges



Notes. (a) Slope = 0.88, intercept = 6×10^{-3}, $R^2 = 84\%$; (b) slope = 0.80, intercept = 9×10^{-3}, $R^2 = 79\%$; (c) slope = 1.04, intercept = 9×10^{-4}, $R^2 = 71\%$; (d) slope = 1.11, intercept = -4×10^{-3}, $R^2 = 63\%$; (e) slope = 0.90, intercept = 4×10^{-3}, $R^2 = 73\%$. Black lines correspond to linear regressions with intercept.

lower than that in Table 5, varying between 13% and 26%.

4.3.2. Residual Analysis and Accuracy of Delay Estimates Based on the Aggregate Workload. The SSC result culminated in relationship (27) that makes expected delay predictions in each exchange based on the one-dimensional aggregated workload process. Specifically, given the market model coefficients $\beta_i^{(j)}$ and a measurement of the queue sizes at the various exchanges, $Q_i^{(s,j)}(t)$, one can compute the workload via (26) and then construct estimates for the expected delays at the various exchanges via (27). We denote the resulting delay estimates by $\widehat{\mathsf{ED}}^{(s,j)}(t)$, where the $\widehat{\ }$ notation denotes in this context the estimate obtained via the one-dimensional workload process, as opposed to measuring the actual expected delay $\mathsf{ED}^{(s,j)}(t)$ via (24). This prediction can be tested again through a set of linear regressions between the workload delay estimate and the delay estimate that uses information

Table 6. Linear Regressions of the Normalized Queue Length on a Particular Exchange vs. That of the Benchmark Exchange (ARCA)

	Dependent variable: $\overline{Q}_{\text{exchange}}$				
	NASDAQ OMX	BATS	DirectEdge X	NYSE	DirectEdge A
Intercept	0.84***	0.39***	0.25***	0.57***	0.05***
	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)
$\overline{Q}_{\text{ARCA}}$	0.74***	0.45***	0.29***	0.96***	0.24***
	(0.02)	(0.01)	(0.01)	(0.02)	(0.00)
R^2	19%	20%	13%	26%	26%

***$p < 0.01$.

about the state of the exchange (queue length and trading rate). All these regressions are statistically significant and are accompanied with high R^2 values. We do not report on these results; instead, we pursue a more detailed analysis of the residuals, that is, the errors between the workload and exchange-specific delay estimates, $\mathsf{ED}^{(s,j)}(t) - \hat{\mathsf{ED}}^{(s,j)}(t)$. We define the quantity

$$R^2_* \triangleq 1 - \frac{\mathrm{Var}\left(\left\|\mathsf{ED}^{(s,j)}(t) - \hat{\mathsf{ED}}^{(s,j)}(t)\right\|\right)}{\mathrm{Var}\left(\left\|\mathsf{ED}^{(s,j)}(t)\right\|\right)},$$

for each security j. Here, $\mathrm{Var}(\cdot)$ is the sample variance, averaged over all time slots t and both sides of the market s. The quantity R^2_* measures the variability of the residuals unexplained by the relationship (27), relative to the variability of the underlying expected delays. By its definition, when R^2_* is close to 1, most of the variability of expected delays is explained by the relationship (27). Numerical results for R^2_* across securities are given in Table 7. Typical values for R^2_* are approximately 80%, highlighting the predictive power of the one-dimensional workload model as a means of capturing the state of the decentralized fragmented market.

Table 7. The Reduction of Variability in Expected Delays Explained by the Workload Relationship

	R^2_*
Alcoa	75%
American Express	64%
Boeing	75%
Bank of America	80%
Caterpillar	58%
Cisco	87%
Chevron	67%
DuPont	82%
Disney	78%
General Electric	82%
Home Depot	87%
Hewlett-Packard	77%
IBM	63%
Intel	82%
Johnson & Johnson	83%
JPMorgan	88%
Kraft	79%
Coca-Cola	81%
McDonalds	74%
3M	62%
Merck	78%
Microsoft	80%
Pfizer	79%
Procter & Gamble	80%
AT&T	77%
Travelers	67%
United Tech	47%
Verizon	79%
Wal-Mart	85%
Exxon Mobil	81%

4.4. Effects of Fee Change: Evidence from the NASDAQ Fee Experiment

We conclude with a separate verification of the predictions of our model in the context of a natural experiment done by the NASDAQ exchange, whereby they made a significant reduction of its fee and rebate schedule for a subset of 14 stocks between February and May of 2015.[15] NASDAQ lowered the fees charged to liquidity takers from \$0.0030 per share to \$0.0005 per share, and correspondingly, lowered the rebates rewarded to liquidity providers from \$0.0029 per share to \$0.0004 per share.

To test the impact of this significant reduction in the make-take fee on NASDAQ, we analyze and compare TAQ data of the 14 tested symbols in two separate time periods: the *preperiod* of January 12–30, 2015 and the *postperiod* of February 9–27, 2015 (three weeks before and after the initiation of the program at the beginning of February 2015, respectively). Table 8 contains the fees charged on the six major exchanges in the tested periods, during which only that of NASDAQ changed.

A change in the per share fee and rebate will affect the attractiveness of the exchange for traders placing limit orders and traders sending aggressive market orders. Our model of market and limit order routing makes some direct predictions on market outcomes, specifically suggesting that the fee change will affect both the trading rate and displayed depth at the

Table 8. Fees of the Six Major U.S. Stock Exchanges, per Share Traded, in January–February 2015 Around the Time of the NASDAQ Access Fee Experiment

Exchange	Fee ($ per share, $\times 10^{-4}$)
NASDAQ OMX: January 2015	30.0
February 2015	5.0
BATS	30.0
DirectEdge X (EDGX)	30.0
ARCA	30.0
NYSE	27.0
DirectEdge A (EDGA)	−2.0

Note. The fees here are different from previous numbers in Table 2 because they are in different time periods.

exchange through their impact on the expected delay experienced by limit orders.

In what follows, we will first estimate the exchange attraction coefficients before and after the fee change. We will propose a structural model for the attractiveness of each exchange that explicitly incorporates its prevailing fee. From this model, we expect that the attractiveness of NASDAQ is increased after the fee reduction, and we verify this prediction. We then study the effect of the fee change in the routing of limit orders. In this case, our model predicts that the equilibrium expected delay for limit orders to get filled on NASDAQ will decrease after the fee change. The empirical analysis will again verify this prediction.

Separately, we randomly construct a control group of 100 securities that were constituents of the S&P500 index during the January–February 2015 period and not included in the access fee experiment. We empirically estimate the attraction coefficients of the market order-routing model and the expected delays for these securities in the pre- and postfee change periods and then perform a difference-in-differences analysis that verifies that there are statistically significant differences between the control and test groups.

Together, these observations suggest that the impact of the fee change is best understood through its structural impact to limit and market order-routing policies and their impact on trading delays. The agreement of our predictions with the observed market response suggests that our model could be useful in addressing either policy related questions or questions of exchange competition that often involves changes in pricing (fee/rebate) decisions. Our findings complement those reported by Hatheway (2015) and Pearson (2015) on the impact of this fee change; these studies are not predicated on an underlying model of order routing and are primarily focused on market share and depth comparisons, before and after the fee change.

4.4.1. Attraction Coefficient β_{NASDAQ}. The discussion in Section 2.2 suggested that the attractiveness of an exchange for market orders is a decreasing function of its fee. Our earlier analysis focused on an observation period where fees were constant, which implied that the attractiveness coefficients of the exchanges were themselves constant throughout that time period. The NASDAQ fee experiment allows us to proceed with a more nuanced analysis to examine the effect of the exchange fees on market order flow. We postulate the following structural model for the routing of market orders:

$$\mu_i^{(s,j)}(t) = \mu^{(s,j)}(t)\,\frac{\left[e^{a_i^{(j)}+r_{i,t}b^{(j)}}\right]Q_i^{(s,j)}(t)}{\sum_{k=1}^{N}\left[e^{a_k^{(j)}+r_{k,t}b^{(j)}}\right]Q_k^{(s,j)}(t)}. \tag{31}$$

In other words, we are postulating the attractiveness coefficients take the form $\beta_i^{(j)} = e^{a_i^{(j)}+r_{i,t}b^{(j)}}$, given parameters $a_i^{(j)}, b^{(j)}$ that we estimate using three weeks of data before and after the fee change. Our hypothesis is that the parameter $b^{(j)}$ is negative, that is, an higher fee makes an exchange less desirable, all other things being equal. The corresponding $\{\mu_i^{(s,j)}(t)\}$, $\{\mu^{(s,j)}(t)\}$, and $\{Q_i^{(s,j)}(t)\}$ are estimated from the two trade and quote data samples as outlined in Section 4.1.

We normalize the results so that the parameter $a_{\text{ARCA}}^{(j)}$ of the benchmark exchange ARCA is 0. Finally, $b^{(j)}$ is estimated using nonlinear regressions on (31), based on the combined sample for each security. Results are in Table 9. Indeed, the estimated $\{b^{(j)}\}$ coefficients are negative for all 14 tested securities; 12 of these estimated coefficients are statistically significant at the 5% level of the corresponding one-sided test. This agrees with our hypothesis.

We also directly estimate the β_{NASDAQ} coefficients before and after the fee change and compare. This estimation is nonparametric in the sense that it is not predicated on (31) and estimates β_{NASDAQ} via nonlinear regressions the following:

$$\mu_i^{(s,j)}(t) = \mu^{(s,j)}(t)\,\frac{\beta_i^{(j)}Q_i^{(s,j)}}{\sum_{i'=1}^{N}\beta_{i'}^{(j)}Q_{i'}^{(s,j)}}, \tag{32}$$

based on the preperiod sample and postperiod sample, respectively. Our market order-routing model is invariant to scaling of the attraction coefficients, and in this section, we normalize so that the attraction coefficient for each stock on the benchmark exchange ARCA is 1. Table 10 reports and compares the estimated attraction coefficients before and after the NASDAQ fee experiment for individual stocks. Note that $\beta_{\text{NASDAQ}}^{(j)}$ - post is greater than $\beta_{\text{NASDAQ}}^{(j)}$ - pre for

Table 9. Estimates of b and a_i in the Attraction Model (31) and Hypothesis Testing Results on Whether the Coefficient b Is Negative

	$b^{(j)}$	$a^{(j)}_{\text{NASDAQ}}$	$a^{(j)}_{\text{EDGX}}$	$a^{(j)}_{\text{BATZ}}$	$a^{(j)}_{\text{NYSE}}$	$a^{(j)}_{\text{EDGA}}$	b Negative?	One-sided 5% test
AAL	−4.92	0.11	−0.02	0.27		0.09	Yes	No
BAC	−179.50	0.42	0.20	0.26	0.13	0.46	Yes	Yes
FEYE	−98.65	−0.30	−0.20	0.10		−0.32	Yes	Yes
GE	−93.50	0.15	−0.04	0.13	−0.01	0.27	Yes	Yes
GPRO	−50.15	−0.07	−0.21	0.07		−0.07	Yes	Yes
GRPN	−94.41	0.12	−0.07	0.25		0.00	Yes	Yes
KMI	−113.10	0.09	0.06	0.13	−0.18	−0.02	Yes	Yes
MU	−89.02	0.06	−0.02	0.14		0.19	Yes	Yes
RAD	−138.46	−0.05	−0.22	0.07	−0.29	−0.09	Yes	Yes
RIG	−76.56	−0.02	−0.02	0.04	−0.14	0.10	Yes	Yes
S	−162.59	−0.22	−0.29	0.04	−0.26	−0.38	Yes	Yes
SIRI	−69.31	0.13	−0.13	0.17		0.33	Yes	Yes
TWTR	−87.33	−0.14	−0.16	−0.01	−0.27	−0.23	Yes	Yes
ZNGA	−17.66	0.14	−0.28	0.14		0.41	Yes	No

Note. For each stock, the results are based on a combined sample that includes both the preperiod and the postperiod of the fee experiment.

12 of the 14 tested securities. For all these 12 names, the increments are statistically significant under a one-sided test at the 5% level. This is again consistent with our prediction.

4.4.2. Expected Delay $\mathsf{ED}_{\text{NASDAQ}}$. Our limit order-routing model suggests that traders tradeoff expected delay with rebate and that, in equilibrium, exchanges that offer lower rebates will also offer lower expected delays for limit orders placed in the back of the queue at the best bid (top of book) until they get filled.

As stated in (12), the expected delay satisfies $\mathsf{ED}_i = \frac{W}{\mu\beta_i}$. We expect the workload W to remain the same after NASDAQ reduces its make-take fee, as the equilibrium value of W depends on the *marginal* exchange, which is likely to be the one with the lowest rebate, which is not NASDAQ; we are assuming that the remaining model parameters remain the same. As described previously, we anticipate the attraction coefficient β_{NASDAQ} to increase after the fee change, which would result in a lower expected delay $\mathsf{ED}_{\text{NASDAQ}}$ after NASDAQ reduced its make-take fee. An alternative justification is that traders submitting orders into NASDAQ would expect lower expected delays given that they are compensated with a lower rebate when their orders trade. In equilibrium, patient traders will submit orders to higher rebate exchanges, which would result in a lower equilibrium delay at NASDAQ after the fee change.

To test this hypothesis, we will compare the normalized expected delay at NASDAQ before and after the fee change. We first compute the measure of expected delay along the lines of Section 4.1, as

$$\mathsf{ED}_i^{(s,j)}(t) \triangleq \frac{Q_i^{(s,j)}(t)}{\mu_i^{(s,j)}(t)}, \tag{33}$$

Table 10. Estimates of the Attraction Coefficient of Individual Stocks on NASDAQ Before and After the Fee Experiment and Hypothesis Testing Results on Whether the Attraction Coefficient Increases Under the Fee Change

	$\beta^{(j)}_{\text{NASDAQ}}$ - pre	Standard deviation	$\beta^{(j)}_{\text{NASDAQ}}$ - post	Standard deviation	Increase?	One-sided 5% test
AAL	1.1478	0.0152	1.0787	0.0189	No	No
BAC	1.4516	0.0297	2.5617	0.0647	Yes	Yes
FEYE	0.6958	0.0136	0.9543	0.0156	Yes	Yes
GE	1.1339	0.0244	1.5250	0.0386	Yes	Yes
GPRO	0.9285	0.0196	1.0557	0.0250	Yes	Yes
GRPN	1.1035	0.0253	1.4288	0.0258	Yes	Yes
KMI	1.0814	0.0163	1.4780	0.0267	Yes	Yes
MU	1.0314	0.0124	1.3862	0.0186	Yes	Yes
RAD	0.9515	0.0302	1.3530	0.0368	Yes	Yes
RIG	0.9775	0.0230	1.1925	0.0239	Yes	Yes
S	0.9905	0.0453	1.0957	0.0429	Yes	Yes
SIRI	1.1787	0.0265	1.3045	0.0347	Yes	Yes
TWTR	0.9093	0.0235	1.0623	0.0269	Yes	Yes
ZNGA	1.2111	0.0337	1.1939	0.0332	No	No

Table 11. Estimates of the Normalized Expected Delay on NASDAQ of Individual Stocks Before and After the Fee Experiment and Hypothesis Testing Results on Whether the Normalized Expected Delay Decreases Under the Fee Change

	$\tilde{\mathrm{ED}}^{(j)}_{\mathrm{NASDAQ}}$(pre)	Standard error	$\tilde{\mathrm{ED}}^{(j)}_{\mathrm{NASDAQ}}$(post)	Standard error	Decrease?	One-sided 5% test
AAL	0.1978	0.0023	0.1886	0.0027	Yes	Yes
BAC	0.1556	0.0022	0.0963	0.0020	Yes	Yes
FEYE	0.2282	0.0037	0.1964	0.0031	Yes	Yes
GE	0.1688	0.0026	0.1265	0.0023	Yes	Yes
GPRO	0.2262	0.0053	0.2145	0.0057	Yes	No
GRPN	0.2030	0.0034	0.1700	0.0035	Yes	Yes
KMI	0.1646	0.0018	0.1265	0.0018	Yes	Yes
MU	0.2062	0.0023	0.1683	0.0024	Yes	Yes
RAD	0.1750	0.0038	0.1027	0.0030	Yes	Yes
RIG	0.1721	0.0022	0.1401	0.0024	Yes	Yes
S	0.1765	0.0063	0.1305	0.0034	Yes	Yes
SIRI	0.2150	0.0072	0.1558	0.0077	Yes	Yes
TWTR	0.1813	0.0026	0.1453	0.0035	Yes	Yes
ZNGA	0.1831	0.0062	0.1515	0.0054	Yes	Yes

for side s, security j, on exchange i, in time slot t. We then use these measures to calculate an aggregate, normalized estimate of the expected delay on NASDAQ, as follows:

$$\tilde{\mathrm{ED}}^{(j)}_{\mathrm{NASDAQ}} = \frac{1}{2T} \sum_{s\in\{\mathsf{BID},\mathsf{ASK}\}} \sum_{t=1}^{T} \frac{\mathrm{ED}^{(s,j)}_{\mathrm{NASDAQ}}(t)}{\sum_{k=1}^{N} \mathrm{ED}^{(s,j)}_{k}(t)}. \quad (34)$$

For each stock, we can obtain two estimates $\tilde{\mathrm{ED}}^{(j)}_{\mathrm{NASDAQ}}$(pre) and $\tilde{\mathrm{ED}}^{(j)}_{\mathrm{NASDAQ}}$(post) based on the preperiod sample and postperiod sample, respectively. Table 11 reports on these two statistics for individual securities. We observe that the normalized expected delay on NASDAQ decreases for all 14 tested securities; in 13 of these 14 cases, the reduction is statistically significant. This agrees with the prediction of our model. Table 12 illustrates that this result is robust to normalizing the delay by the median (rather than the sum) of delays.

4.4.3. Linear Relation $\mathrm{ED}_{\mathrm{NASDAQ}} = \beta_{\mathrm{ARCA}}/\beta_{\mathrm{NASDAQ}} \cdot \mathrm{ED}_{\mathrm{ARCA}}$. Last, we examine how the fee change affects the linear relation (28), which is one of the major conclusions arising from our model:

$$\mathrm{ED}^{(s,j)}_{i}(t) = \frac{\beta^{(j)}_{i'}}{\beta^{(j)}_{i}}\, \mathrm{ED}^{(s,j)}_{i'}(t). \quad (35)$$

Specifically, we want to test that when considering the linear relation between NASDAQ and the benchmark exchange, ARCA, the slope of that linear relation before and after the fee change will decrease, because we expect that the attraction coefficient β_{NASDAQ} should increase in response to that change. We perform linear regressions for each security between the expected delays on NASDAQ against that on the benchmark exchange ARCA before and after the fee change and report the results in Table 13. We observe that the resulting slopes decrease for all 14 tested securities, among which 8 are statistically significant under a one-sided test at the 5% level. In addition, in a cross-sectional linear regression, the slope before the fee change was $0.78564^{***}(0.01571)$, $R^2 = 58\%$, and $0.66384^{***}(0.01221)$, $R^2 = 62\%$ after the fee change; the decrease in the slope is statistically significant. Table 14 illustrates that these results are robust to performing the regression with no intercept.

4.4.4. Difference-in-Differences Analyses. We randomly selected a control group of 100 securities that were constituents of the S&P500 index during our study period January–February 2015 and were not included in the NASDAQ fee experiment; the control group is denoted by $\mathscr{C}$, and the text group will be denoted by $\mathscr{T}$. For each security $j \in \mathscr{C}$, we used market data to first estimate the market order-routing model as in (25), denoted by $\beta^{(j)}_{\mathrm{NASDAQ}}$(pre) and $\beta^{(j)}_{\mathrm{NASDAQ}}$(post), and to empirically measure the normalized average expected delay encountered, denoted by $\tilde{\mathrm{ED}}^{(j)}_{\mathrm{NASDAQ}}$(pre) and $\tilde{\mathrm{ED}}^{(j)}_{\mathrm{NASDAQ}}$(post), using (33) and (34). We define $Y^j = \beta^{(j)}_{\mathrm{NASDAQ}}$(post) $- \beta^{(j)}_{\mathrm{NASDAQ}}$(pre) and $Z^j = \tilde{\mathrm{ED}}^{(j)}_{\mathrm{NASDAQ}}$(post) $- \tilde{\mathrm{ED}}^{(j)}_{\mathrm{NASDAQ}}$(pre) and regress Y^j and Z^j against the indicator variables $I^j = 1$ if $j \in \mathscr{T}$ and $= 0$, otherwise.

The results of these regressions are shown in Tables 15 and 16. After the fee reduction, it became more attractive to route market orders to NASDAQ as reflected in the positive and statistically significant change in its attractiveness coefficient, $\beta^{(j)}_{\mathrm{NASDAQ}}$, relative to the corresponding change in the control set.

Table 12. Results in Parallel to Those in Table 11 when the Expected Delays Are Normalized by Median Delay Instead of by Sum of Delays

	$\tilde{ED}^{(j)}_{NASDAQ}$(pre)	Standard error	$\tilde{ED}^{(j)}_{NASDAQ}$(post)	Standard error	Decrease?	One-sided 5% test
AAL	1.0029	0.0109	0.9681	0.0139	Yes	Yes
BAC	0.9293	0.0121	0.5895	0.0137	Yes	Yes
FEYE	1.2431	0.0202	1.0567	0.0164	Yes	Yes
GE	1.0267	0.0160	0.7863	0.0169	Yes	Yes
GPRO	1.1619	0.0336	1.0842	0.0360	Yes	No
GRPN	1.0468	0.0191	0.8941	0.0168	Yes	Yes
KMI	0.9938	0.0100	0.7622	0.0108	Yes	Yes
MU	1.0318	0.0114	0.8699	0.0127	Yes	Yes
RAD	1.1165	0.0303	0.6757	0.0195	Yes	Yes
RIG	1.0361	0.0128	0.8614	0.0142	Yes	Yes
S	1.2665	0.0711	0.8699	0.0218	Yes	Yes
SIRI	1.4101	0.1685	1.4559	0.4852	No	No
TWTR	1.1077	0.0166	0.9034	0.0260	Yes	Yes
ZNGA	1.0337	0.0404	0.8984	0.0308	Yes	Yes

Table 13. Linear Regression Results of Equation (35) and Hypothesis Testing Results on Whether the Slope Decreases After the Fee Change on NASDAQ

	Slope (before)	Standard error	Slope (after)	Standard error	Decrease?	One-sided 5% test
AAL	0.6742	0.0308	0.5981	0.0392	Yes	No
BAC	0.6176	0.0299	0.3245	0.0269	Yes	Yes
FEYE	0.8773	0.0761	0.7950	0.0543	Yes	No
GE	0.7851	0.0384	0.4859	0.0349	Yes	Yes
GPRO	0.3744	0.0458	0.3314	0.0624	Yes	No
GRPN	0.9570	0.0418	0.8223	0.0234	Yes	Yes
KMI	0.8764	0.0289	0.5400	0.0255	Yes	Yes
MU	0.8313	0.0277	0.5741	0.0274	Yes	Yes
RAD	0.9439	0.0611	0.3090	0.0482	Yes	Yes
RIG	0.6571	0.0286	0.5360	0.0282	Yes	Yes
S	0.5740	0.1151	0.5406	0.0395	Yes	No
SIRI	1.3337	0.2686	(0.0001)	0.0135	Yes	Yes
TWTR	0.8406	0.0679	0.7470	0.0724	Yes	No
ZNGA	0.0546	0.0087	0.0425	0.0174	Yes	No

Table 14. Results in Parallel to Those in Table 13 when the Linear Regressions Are Performed Without Intercept

	Slope (before)	Standard error	Slope (after)	Standard error	Decrease?	One-sided 5% test
AAL	0.7960	0.0149	0.8121	0.0188	No	No
BAC	0.6774	0.0154	0.3789	0.0138	Yes	Yes
FEYE	1.3292	0.0428	1.0828	0.0288	Yes	Yes
GE	0.8285	0.0221	0.5942	0.0194	Yes	Yes
GPRO	0.7147	0.0375	0.6852	0.0453	Yes	No
GRPN	0.9706	0.0294	0.8040	0.0207	Yes	Yes
KMI	0.9253	0.0150	0.6436	0.0123	Yes	Yes
MU	0.9034	0.0152	0.6440	0.0140	Yes	Yes
RAD	1.0516	0.0440	0.4285	0.0366	Yes	Yes
RIG	0.7851	0.0199	0.6453	0.0177	Yes	Yes
S	0.6519	0.1113	0.6650	0.0299	No	No
SIRI	1.4071	0.2435	0.0017	0.0134	Yes	Yes
TWTR	1.1070	0.0354	0.8970	0.0390	Yes	Yes
ZNGA	0.0584	0.0091	0.0550	0.0178	Yes	No

Table 15. Difference-in-Differences Regression for the Change in Attractiveness of NASDAQ for Routing Market Orders Pre- and Postfee Change

Intercept	0.0143
	(0.0338)
Test group indicator I^j	0.3001***
	(0.0935)
R^2	8%

***$p < 0.01$.

Table 16. Difference-in-Differences Regression for the Change in Expected Delay Pre- and Postfee Change

Intercept	−0.0007
	(0.00006)
Test group indicator I^j	−0.039***
	(0.0016)
R^2	4%

***$p < 0.01$.

The magnitudes of the estimated coefficients for the test set indicator variable for the two regressions correspond to roughly a 30% increase in attractiveness and a 24% reduction in (normalized) delay. Finally, the fractions of the control set securities that experienced an increase in attractiveness and a reduction in their expected delays from pre- to postfee change periods were 48% and 40%, respectively.

Endnotes

[1] This paper will adopt the terminology encountered in financial markets, both to help describe this domain that may be of independent interest to the stochastic modeling community and to highlight the close connection between the model, the associated results, and the underlying application.

[2] The *bid* is the highest price level at which limit orders to buy stock of a particular security are represented at an exchange; the *offer* or the *ask* is the lowest price level at which limit order to sell stock are represented at the exchange; the bid price is less than the offered price. The difference between the offer and the bid is referred to as the *spread*. Exchanges may differ in their bid and offer price levels, and at any point in time, the highest bid and the lowest offer among all exchanges comprise the national best bid and offer (NBBO).

[3] Negative fees occur on *inverted* exchanges where payments are made to liquidity takers.

[4] Investors differ in the urgency with which they seek to execute their orders, which is, in turn, captured by the parameters of their selected execution algorithm, for example, an algorithm with a target participation rate, perhaps 5%, 10%, or 20% of the market volume. Such an urgency parameter affects how long will a trader be willing to wait until a limit order would be filled, which would affect the order placement decision.

[5] Regulation NMS, see http://www.sec.gov/spotlight/regnms.htm.

[6] A simpler version of this effect is the familiar picture we encounter in highway toll booths or supermarket checkout lines, where people join the shortest queue; in our model choice behavior is more intricate, and depends on economics, anticipated delays, and trader heterogeneity.

[7] Cancellations are common. Typical models of cancellations assume either that orders cancel according to an exponential alarm clock leading to a cancelation process that is proportional to the queue length or that there is constant drift out of the bid queue because of cancelations, independent of the queue length. The first offers a reasonable model for orders generated by algorithmic trading strategies used by institutional investors, such as Volume Weighted Average Price (VWAP), Percent of Volume (POV), etc., but it is not a good way to model the behavior of orders posted by market makers. The latter account for most of the orders in the queue, and indeed, they tend to cancel using a state-dependent criterion as opposed to a time-based one. The simple cancellation models described previously would underestimate the expected delay until an order will get filled in liquid securities. The incorporation of the different cancellation behaviors, timer based and state dependent, complicates the dynamics of the queue but leads to better agreement with data (Kukanov and Maglaras 2015).

[8] Criterion (2) is *static*. In practice, order-routing decisions are *dynamic*, that is, done and updated over the lifetime of the order in the market.

[9] Here, we use a snapshot estimate of expected delays that is consistent with our definition (1) and is often used in practice. This disregards the fact that $Q(t)$, and as a result, $\mu_i(Q(t))$ may change over time, which would naturally affect the delay estimate. In what follows, we will mainly be concerned with the behavior of the system in equilibrium, where $Q(t)$ is constant and this distinction is not relevant.

[10] We will typically expect that $\pi_i(\gamma) \in \{0, 1\}$, that is, all type γ investors will prefer a single exchange, unless there are ties between exchanges.

[11] Strictly speaking, the informal definition (8) may not deal properly with situations where agents are indifferent between multiple routing decisions, whereas the formal Definition 1 handles this correctly. Under mild technical conditions we will adopt shortly (Assumption 1 and the hypothesis of Theorem 3), however, the mass of such agents is zero and the two definitions coincide.

[12] The NASDAQ listed stocks in our sample (CSCO, INTC, MSFT) do not trade on the NYSE; hence, for these stocks, only $N = 5$ exchanges were considered.

[13] The time intervals should be sufficiently long to get reliable estimates of the event rates and also long compared with the event interarrival times, so that one could expect that the transient dynamics of the market because changes in these rates settle down during these time intervals.

[14] For example, with 15-minute periods, the first principle component still explains 69% of the overall variability of the vector of delay trajectories (that are themselves four times longer), whereas the first two principle components explains 82% of the variability.

[15] The subset of stocks participating in the experiment are as follows: AAL, BAC, FEYE, GE, GPRO, GRPN, KMI, MU, RAD, RIG, S, SIRI, TWTR, ZNGA.

References

Alfonsi A, Fruth A, Schied A (2010) Optimal execution strategies in limit order books with general shape functions. *Quant. Finance* 10:143–157.

Allon G, Federgruen A (2007) Competition in service industries. *Oper. Res.* 55(1):37–55.

Avellaneda M, Reed J, Stoikov S (2011) Forecasting prices from level-i quotes in the presence of hidden liquidity. *Algorithmic Finance* 1(1):35–43.

Barclay MJ, Hendershott T, McCormick TD (2003) Competition among trading venues: Information and trading on electronic communications networks. *J. Finance* 58:2637–2666.

Bassamboo A, Harrison JM, Zeevi A (2004) Dynamic routing in large call centers: Asymptotic analysis of an LP-based method. *Oper. Res.* 10:1074–1099.

Besbes O, Maglaras C (2009) Revenue optimization for a make-to-order queue in an uncertain market environment. *Oper. Res.* 57(6):1438–1450.

Bessembinder H (2003) Quote-based competition and trade execution costs in NYSE listed stocks. *J. Financial Econom.* 70:385–422.

Biais B, Bisière C, Spatt C (2010) Imperfect competition in financial markets: An empirical study of Island and Nasdaq. *Management Sci.* 56(12):2237–2250.

Blanchet J, Chen X (2013) Continuous-time modeling of bid-ask spread and price dynamics in limit order books. Working paper, Columbia University, New York, NY.

Bouchaud J-P, Gefen Y, Potters M, Wyart M (2004) Fluctuations and response in financial markets: The subtle nature of 'random' price changes. *Quant. Finance* 4:176–190.

Bramson M (1998) State space collapse with applications to heavy-traffic limits for multiclass queueing networks. *Queueing Systems* 30:89–148.

Buti S, Rindi B, Wen Y, Werner IM (2011) Tick size regulation, intermarket competition and sub-penny trading. Working paper, Ohio State University, Columbus, OH.

Cachon G, Harker P (2002) Competition and outsourcing with scale economies. *Management Sci.* 48:1314–1333.

Caldentey R, Kaplan E, Weiss G (2009) Fcfs infinite bipartite matching of servers and customers. *Adv. Appl. Probabilities* 41(3):695–730.

Chen Y-J, Maglaras C, Vulcano G (2019) Design of an aggregated marketplace under congestion effects: Asymptotic analysis and equilibrium characterization. *Sharing Economy* (Springer, Cham), 217–248.

Cont R, De Larrard A (2013) Price dynamics in a markovian limit order market. *SIAM J. Financial Math.* 4(1):1–25.

Cont R, Kukanov A (2013) Optimal order placement in limit order markets. Working paper, Columbia University, New York, NY.

Cont R, Stoikov S, Talreja R (2010) A stochastic model for order book dynamics. *Oper. Res.* 58:549–563.

Degryse H, de Jong F, van Kervel V (2011) The impact of dark trading and visible fragmentation on market quality. Working paper.

Dufour A, Engle RF (2000) Time and the price impact of a trade. *J. Finance* 55:2467–2498.

Foucault T, Menkveld AJ (2008) Competition for order flow and smart order routing systems. *J. Finance* 63:119–158.

Foucault T, Kadan O, Kandel E (2005) Limit order book as a market for liquidity. *Rev. Financial Stud.* 18:1171–1217.

Gatheral J (2010) No-dynamic-arbitrage and market impact. *Quant. Finance* 10:749–759.

Glosten L (1994) Is the electronic order book inevitable? *J. Finance* 49:1127–1161.

Glosten L (1998) Competition, design of exchanges and welfare. Working paper, Columbia University, New York, NY.

Glosten LR (1987) Components of the bid/ask spread and the statistical properties of transaction prices. *J. Finance* 42:1293–1307.

Glosten LR, Milgrom PR (1985) Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. *J. Financial Econom.* 14:71–100.

Griffiths MD, Smith BF, Turnbull DAS, White RW (2000) The costs and the determinants of order aggressiveness. *J. Financial Econom.* 56:65–88.

Guo X, De Larrard A, Ruan Z (2013) Optimal placement in a limit order book. *Theory Driven by Influential Applications*, 191–200.

Gurvich I, Ward A (2014) On the dynamic control of matching queues. *Stochastic Systems* 4(2):479–523.

Hamilton JL (1979) Marketplace fragmentation, competition, and the efficiency of the stock exchange. *J. Finance* 34:171–187.

Harrison JM (1988) Brownian models of queueing networks with heterogeneous customer populations. Fleming W, Lions PL, eds. *Stochastic Differential Systems, Stochastic Control Theory and Applications*, vol. 10 of *Proc. IMA* (Springer-Verlag, New York), 147–186.

Harrison JM (1995) Balanced fluid models of multiclass queueing networks: a heavy traffic conjecture. Kelly F, Williams R, eds. *Stochastic Networks*, vol. 71 of *Proc. IMA* (Springer-Verlag, New York), 1–20.

Harrison JM (2000) Brownian models of open processing networks: Canonical representation of workload. *Ann. Appl. Probabilities* 10:75–103.

Harrison JM, Lopez MJ (1999) Heavy traffic resource pooling in parallel-server systems. *Queueing Systems* 33:339–368.

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems* (Kluwer Academic Publishers, Boston).

Hatheway F (2015) Nasdaq access fee experiment. Technical report, Nasdaq, New York.

Hollifield B, Millerz RA, Sandas P (2004) Empirical analysis of limit order markets. *Rev. Econom. Stud.* 71:1027–1063.

Holthausen RW, Leftwich RW, Mayers D (1990) Large-block transactions, the speed of response, and temporary and permanent stock-price effects. *J. Financial Econom.* 26:71–95.

Huberman G, Stanzl W (2004) Price manipulation and quasi-arbitrage. *Econometrica* 72:1247–1275.

Jovanovic B, Menkveld AJ (2011) Middlemen in limit-order markets. Working paper, New York University, New York, NY.

Keim DB, Madhavan A (1998) The cost of institutional equity trades. *Financial Anal. J.* 54:50–59.

Kukanov A, Maglaras C (2015) A limit order queue model with traders with heterogeneous behaviors. Working paper, Columbia University, New York.

Kurtz TG (1977/78) Strong approximation theorems for density dependent Markov chains. *Stochastic Processes Appl.* 6(3): 223–240.

Kyle AS (1985) Continuous auctions and insider trading. *Econometrica* 53:1315–1335.

Lakner P, Reed J, Simatos F (2017) Scaling limit of a limit order book model via the regenerative characterization of lévy trees. *Stochastic Systems* 7(2):342–373.

Lakner P, Reed J, Stoikov S (2016) High frequency asymptotics for the limit order book. *Market Microstructure and Liquidity* 2(1): 1650004.

Lariviere MA (2006) A note on probability distributions with increasing generalized failure rates. *Oper. Res.* 54(3):602–604.

Lederer P, Li L (1997) Pricing, production, scheduling and delivery-time competition. *Oper. Res.* 45:407–420.

Levhari D, Luski I (1978) Duopoly pricing and waiting lines. *Eur. Econom. Rev.* 11:17–35.

Li L, Lee Y (1994) Pricing and delivery-time performance in a competitive environment. *Management Sci.* 40:633–646.

Luski I (1976) On partial equilibrium in a queueing system with two servers. *Rev. Econom. Stud.* 43:519–525.

Maglaras C, Moallemi C, Zheng H (2014) Optimal execution in a limit order book and an associated microstructure market impact model. Working paper, Columbia University, New York, NY.

Malinova K, Park A (2010) Liquidity, volume, and price behavior: The impact of order vs. quote based trading. Working paper, University of Toronto, Toronto, Canada.

Mandelbaum A, Pats G (1995) State-dependent queues: Approximations and applications. Kelly F, Williams R, eds. *Stochastic*

Networks, vol. 71 in *Proc. IMA* (Springer-Verlag, New York), 239–282.

Mendelson H, Whang S (1990) Optimal incentive-compatible priority pricing for the m/m/1 queue. *Oper. Res.* 38(5):870–883.

O'Hara M, Ye M (2011) Is market fragmentation harming market quality? *J. Financial Econom.* 100(3):459–474.

Obizhaeva A, Wang J (2006) Optimal trading strategy and supply/demand dynamics. Working paper, MIT, Cambridge, MA.

Parlour CA, Seppi DJ (2008) Limit order markets: A survey. Boot AWA, Thakor AV, eds. *Handbook of Financial Intermediation & Banking 5*, 63–95.

Parlour CA (1998) Price dynamics in limit order markets. *Rev. Financial Stud.* 11:789–816.

Pearson P (2015) Takeways from the NASDAQ pilot program. Technical report, Investment Technology Group, New York, NY.

Plambeck EL, Ward AR (2006) Optimal control of a high-volume assemble-to-order system. *Math. Oper. Res.* 31(3):453–477.

Rosu I (2009) A dynamic model of the limit order book. *Rev. Financial Stud.* 22:4601–4641.

So K (2000) Price and time competition for service delivery. *Manufacturing Service Oper. Management* 2(4):392–409.

Sofianos G (1995) Specialist gross trading revenues at the New York Stock Exchange. Working paper, New York Stock Exchange, New York, NY.

Sofianos G, Xiang J, Yousefi A (2011) Smart order routing: All-in shortfall and optimal order placement. Technical report, Goldman Sachs, New York, NY.

Stolyar AL (2005) Optimal routing in output-queued flexible server systems. *Probab. Engrg. Inform. Sci.* 19:141–189.

van Kervel V (2012) Liquidity: What you see is what you get? Working paper, Tilburg University, Tilburg, Netherlands.

Zak SH (2003) *Systems and Control* (Oxford University Press).

Costis Maglaras is dean and the David and Lyn Silfen Professor of Business at Columbia Business School. His current research centers on stochastic modeling and data science, with an emphasis on stochastic networks, financial engineering, and quantitative pricing and revenue management.

Ciamac C. Moallemi is an associate professor of business in the decision, risk, and operations division of the Graduate School of Business at Columbia University, where he has been since 2007. His research interests are in the area of the optimization and control of large-scale stochastic systems and decision-making under uncertainty, with an emphasis on applications in financial engineering.

Hua Zheng is a quantitative researcher at JP Morgan Chase, where she leads the equity delta one systematic trading team in North America. Since joining JP Morgan in 2015, she has been developing algorithmic and data-driven trading strategies on electronic market making, optimal execution, and risk management across cash equities central risk book, exchange traded fund (ETF) market making, equities delta one, and foreign exchange (FX) and rates electronic trading.