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

Costis Maglaras
Graduate School of Business
Columbia University
email: c.maglaras@gsb.columbia.edu

Ciamac C. Moallemi*
Graduate School of Business
Columbia University
email: ciamac@gsb.columbia.edu

Hua Zheng
Graduate School of Business
Columbia University
email: hzheng14@gsb.columbia.edu

Current Version: February 25, 2014

Abstract

In modern equity markets, participants have a choice of many exchanges at which to trade. Exchanges typically operate as electronic limit order books operating under a "price-time" priority rule and, in turn, can be modeled as multi-class FIFO 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 delays until their order may fill against financial considerations. These limit orders can be thought as jobs waiting for service. Simultaneously, traders that submit market orders select the exchange, i.e., the queue, to direct their order. These market orders trigger instananeous 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. This predicted dimension reduction is the result of high-frequency order routing decisions that essentially couple the dynamics across exchanges. We derive a characterization of the market equilibrium and the associated aggregated depth process. Analyzing a TAQ dataset for a sample of stocks over a one month period, we find empirical support for the predicted state space collapse.

1. Introduction

Motivation. Modern equity markets are highly fragmented. In the United States alone there are over a dozen exchanges and about forty alternative trading systems where investors may choose to trade. Market participants, including institutional investors, market makers, and opportunistic

*The second author acknowledges the support of NSF Grant CMMI–1235023.

investors, interact within today's high-frequency, fragmented marketplace with the use of electronic algorithms. These algorithms 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 taking into account 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 high-frequency dynamics of one exchange can be understood as that of a multi-class 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 which, in queueing system parlance, correspond to service completions. That is, exchanges do not possess their own "server," as is typical in the study of service and production systems, but the service process is the result of service completions arising from market orders routed to the exchanges by traders over time.

The entire market, consisting of multiple exchanges, can be viewed as a stochastic network that evolves as a collection of parallel, multi-class 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 is an important issue in understanding trade execution, market design, and policy questions that have come to the fore in recent years. This paper focuses on this problem, and, specifically, establishes and then empirically tests a structural property that emerges through these otherwise complex interaction effects.[1]

At a point in time, the 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, etc. 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 metrics that capture the quality of execution over time and 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. In addition, exchanges 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, the exchanges charge a fee to "takers" of liquidity that initiate trades using marketable orders that transact against posted limit orders. These fees range in magnitude, and are typically between $-\$0.0010$ and $\$0.0030$ per share traded.

[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).

Since 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 employ 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 execute their order.

From a stochastic modeling viewpoint, the aforementioned system consists of parallel multi-class 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 while optimizing heterogeneous trade-offs between economics and operational metrics related to queueing effects.

Summary of results. At a high level, this paper makes three contributions. First, it offers a novel model for order routing in fragmented markets. It proposes a double-sided queueing system that takes into effect the atomistic limit order placement and market order (service completions) routing decisions. This paper appears to be one of the first in financial engineering or market microstructure to study the exchange queueing dynamics and their effect on order routing decisions. In parallel, it is one of the first stochastic modeling papers to focus in this application domain, and apart from introducing related queueing questions, it concretely motivates a decentralized variant of the well-studied parallel queueing system. Some modeling features motivated by the financial application, specifically the self-interested routing of the service completions, may be more broadly applicable and of independent methodological interest; e.g., one possible application might be in modeling personnel that work in retailing that may strategize over which customer to help next.

Second, from a methodological viewpoint, we study a deterministic and continuous fluid model associated with the above 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.[3] The market equilibrium and SSC are not the result of the price protection mechanism[4] imposed in the U.S. equities market. Rather, they arise out of order routing decisions

[3]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 in this paper.

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

among exchanges that offer the same price level yet that differ with respect to other factors such as their exchange fees. We characterize the nature of this coupling effect, and highlight a strikingly simplifying property whereby the behavior of the multi-dimensional 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: queue lengths can be inferred from it, as can the routing behavior of investors. Moreover, 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. 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 the coupling depends on both the economics and anticipated delays of each exchange, as well as trader heterogeneity. For the case with $N = 2$ 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 1-dimensional workload system seems to offer a tractable model for downstream analysis of interesting questions that pertain to exchange competition (e.g., how to set take fees or associated volume tiers), policy questions that may affect the structure of the routing decision problem or impose exogenous transaction costs (e.g., a tax on the value of a transaction), and market design questions (e.g., whether the co-existence of competing exchanges that offer differentially priced execution platforms is beneficial from a welfare perspective).

Third, we empirically verify the state space collapse property for a sample of TAQ data for the month of 9/2011 for the 30 securities that comprise the Dow Jones Index. While all being 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.

We test the implications of our model in several ways. First, we perform a principal component analysis (PCA) to characterize the effective dimension of the joint vector of expected delays across exchanges. Our theoretical prediction of state space collapse suggests that this vector should be contained in a one-dimensional set. In support of this prediction, we find that the first principal component explains around 80% of the variability of the expected delays across exchanges, and that the first two principal components explain 90%. Second, our analysis suggests that the expected delays across exchanges are linearly related, in fact, proportional to each other. We test this prediction by running a set of linear regressions over different pairs of exchanges and find statistically significant support for a linear relationship in all cases. The R^2 varies between 76% and 87%. The

regression coefficients are also very close to the ones predicted by our analysis, despite the simple structure of the mathematical model and the inherent noise in the extensive market data sample. Similarly good fits are obtained if one were to carry through the analysis separately for each security. The SSC result suggests that the expected delays in each exchange can be inferred through the market-wide workload. This prediction can be tested again through a set of linear regressions between the the workload delay estimate and the delay estimate that uses information about the state of the exchange (queue length and trading rate). All these regressions are again 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, i.e., the errors between the workload and exchange-specific delay estimates, and find that the workload estimate captures on the average 80% of the variation in exchange delays. Overall, the empirical analysis provides statistical support for the theoretical SSC prediction of our model, despite the fact that many of the assumptions of our model may be not satisfied in practice. To our knowledge, this seems to be one of the first empirical verifications of SSC in a real and complex stochastic processing system. Most of the related literature, which we briefly review below, is prescriptive in that it formulates and studies models that are meant to offer insights for how systems should operate.

The remainder of this paper is organized as follows. This section concludes with a very brief literature survey. Section 2 sets up the one-sided, top-of-book model of the limit order book markets and then describes two order routing models: one for limit orders and the other for market orders. Our main results on market equilibrium and state space collapse are given in Section 3. In Section 4, we show empirical evidence of state space collapse.

Literature Survey. There are two strands of literature that we will briefly review. The first is on market microstructure and financial engineering, and focuses on the structure and behavior of limit order books. The second 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.

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, from which we mention the work of Bouchaud et al. (2004), Griffiths et al. (2000), and Hollifield et al. (2004). Parlour (2008) offers a good review of markets operating as limit order books. Related to the above work, there is a body of literature that studies the question of adverse selection, which is important for the placement of limit orders. Good references that span both the empirical and theoretical angles of this topic include the work of Keim and Madhavan (1998), Dufour and Engle (2000), Holthausen et al. (1990), Huberman and Stanzl (2004), Gatheral (2010), and Sofianos (1995).

Second, there is a growing body of work that develops models of limit order book dynamics and studies optimal execution problems. This includes the work of Obizhaeva and Wang (2006), Cont et al. (2010), Rosu (2009), Alfonsi et al. (2010), Foucault et al. (2005), Parlour (1998), Stoikov et al. (2011), Maglaras and Moallemi (2011), Cont and Larrard (2013), and Guo et al. (2013). Most of

that work treats the market as one limit order book and examine its discrete queueing dynamics, or uses an aggregated model of market impact and abstracts away the discrete dynamics of the exchanges. More recently, work as Lakner et al. (2014) describe the whole limit order book by measure valued processes and study its asymptotics under high frequency regimes.

Third, 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 those by 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, while in the markets we consider, the tick-size is uniform. Also related to our work are the papers that study the effect of make-take pricing on market outcomes and liquidity cycles. 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.

Of the market fragmentation literature, closest to our paper is work that examines the impact of smart order routing by market participants. Foucault and Menkveld (2008) studies the effect of smart order routing decisions across multiple (two) exchanges. Their focus, however, is on smart order routing to optimize the execution price (i.e., in a setting without a price protection mechanism like Reg NMS that applies to the U.S. equities market). On the other hand, we focus on the case the execution price is the same, but other, more nuanced factors motivate the order routing decision. van Kervel (2012) considers the impact of order routing in a setting where market makers place limit orders on multiple exchanges simultaneously so as to increase execution probabilities. This analysis, however, ignores economic differences between venues such as rebates and fees, or execution delays. In a similar vein, 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, while Cont and Kukanov (2013) formulates a version of the smart order routing problem.

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, see, e.g., Cont et al. (2010), Maglaras and Moallemi (2011), Cont and Larrard (2013), Lakner et al. (2013), Blanchet and Chen (2013). Some ideas from stochastic network theory are relevant in our work. To start with, the 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 and its consequences has been pioneered by the work of Harrison (1988), Harrison and Van Mieghem (1996), and Harrison (2000). Workload fluid models were first introduced by Harrison (1995), while the use of workload relaxations in fluid model control problems were proposed by Meyn (2001). The condition that guarantees that parallel server systems exhibit SSC and reduce 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.

(2010). Our paper models market order routing decisions via a reduced form state dependent service rate process. There is a broad literature on queues with state dependent rates, including Mandelbaum and Pats (1995, 1998), that derive fluid and diffusion approximations for such systems.

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 Armony and Haviv (2003). Relaxations of the early papers include the work of Loch (1991), who studied symmetric $M/GI/1$ providers, and Lederer and Li (1997) that allowed for an arbitrary number of service providers. 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 the above papers look at static rules, where consumers make decisions based on steady-state expected delays. Chen et al. (2010) 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. In our model, coupling also results through the optimized service completion process. Moreover, 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.

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; 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., the 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.

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.

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, the highest bid price available across all exchanges – the "top-of-book." A profit-maximizing seller would only choose to trade at the top

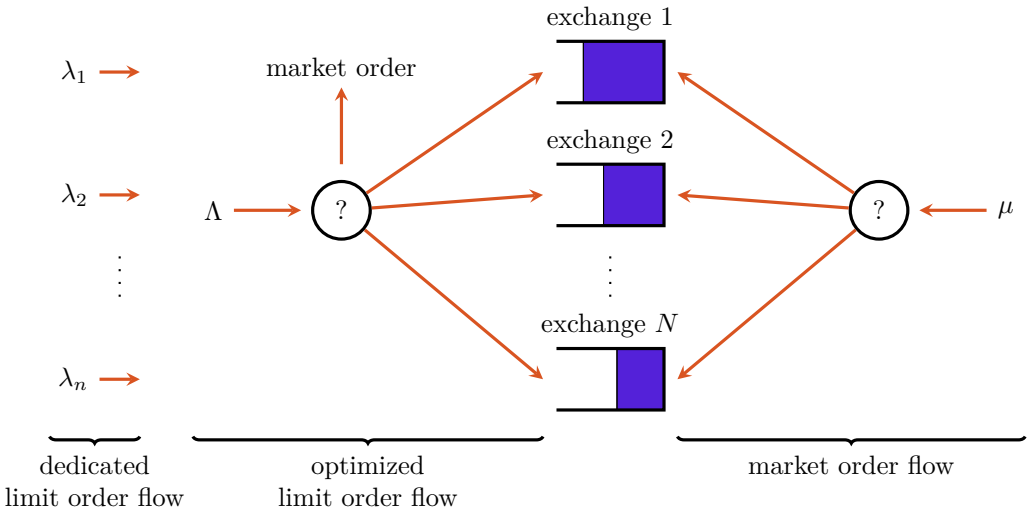


Figure 1: A one-sided, top-of-book model of multiple limit order books. Limit orders (i.e., jobs) arrive to each exchange (modeled by the respective queues) in a) dedicated streams and b) optimized limit order placement decisions. Liquidity is removed through the arrival of decentralized, self-interested market orders, acting as service completions.

of book, and, in fact, in the United States, this is enforced *de jure* by 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, and the time scale of such transients is similar to that of the queueing delays, i.e., the time it takes for queue lengths to 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, while 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 $Q(t) \triangleq (Q_1(t), Q_2(t), \ldots, Q_N(t)) \in \mathbb{R}_+^N$. Sections 2–3 impose the above assumptions. The empirical analysis of § 4, which is based on a large sample of real market data, will test our findings in a setting that is discrete and stochastic, and where system parameters, such as arrival rates, are non-stationary.

2.1. Limit Order Routing

We imagine a continuous and deterministic flow of investors arriving 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 that are common. 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; etc. This assumption simplifies the order routing decision and leads to a tractable analysis.

Investors are heterogeneous in the way they trade off delays and rebates.

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 due to 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 1000 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 $\mathsf{ED}_0 \triangleq 0$.

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 when compared to 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

above. 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; etc. In order to maintain tractability, we will focus on the direct trade-off 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 $\mathcal{E}(t) \triangleq \{(\tilde{r}_i, \mathsf{ED}_i(t)) \ : \ 0 \leq i \leq N\}$. Investors are heterogeneous with respect to their way of trading off rebate against delay. Each investor is characterized by its type, denoted by $\gamma \geq 0$, that is assumed to be an independent identically distributed (i.i.d.) draw from a cumulative distribution function $F(\cdot)$, that is differentiable and has a continuous density function, and selects a routing decision $i^*(\gamma)$ so as to maximize his "utility" according to the rule [5]

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

In other words, γ is a trade-off 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 trade-off coefficient should roughly be in the range of 1 to 10^4 seconds per \$.01.

An equivalent formulation, which is 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. Yet 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. Overall, while (2) is a simplified criterion, it captures the fundamental trade-off 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 is 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. Since a market order executes without any delay, it is

[5] The criterion (2) is "static." In practice, order routing decisions are "dynamic," i.e., done and updated over the lifetime of the order in the market.

natural to route it to exchange i^* so as to minimize the fee paid:

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

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, due to 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 very little available liquidity, "clearing" the queue of resting limit orders is likely to result in greater price impact. (d) There maybe 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 of 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 $f_i(0) = 0$, i.e., market orders will not route to an exchange i with no liquidity. The discussion above 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,

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

Here, $\mu_i(Q(\cdot))$ is the arrival rate of market orders to exchange i, defined by (4)–(5). 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, i.e., 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 $j \notin \{0, i\}$, and $\gamma\tilde{r}_i - \mathsf{ED}_i(t) \geq \gamma\tilde{r}_0$, given the expected delays $\mathsf{ED}_j(t) = Q_j(t)/\mu_j(Q(t))$, for $j = 1, \ldots, N$, implied[6] 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 towards this exchange, increasing delays and erasing its relative advantage viz 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 trade-off taking into account the investors' heterogeneous preferences.

Expressing the fluid equations (6) in differential form, we have that

$$\dot{Q}_i(t) = \lambda_i + \Lambda\chi_i(Q(t)) - \mu_i(Q(t)), \quad i = 1, \ldots, N.$$

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 trade-off coefficient $\gamma \geq 0$, $\pi_i(\gamma)$ denotes the fraction 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 $\pi(\gamma) \triangleq (\pi_0(\gamma), \pi_1(\gamma), \ldots, \pi_N(\gamma))$ satisfy

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

[6]Here, we employ 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.

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*

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

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

(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, i.e.,

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

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[7] with the informal system of equations (8) since $\chi_i(Q^*) = \int_0^\infty \pi_i^*(\gamma)\, dF(\gamma)$.

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 than orders at unattractive exchanges (low β_i, typically corresponding to high $\tilde{r}_i$), since 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

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

[7]Strictly speaking, the informal definition (8) may not deal properly with situations where agents are indifferent between multiple routing decisions, while 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.

That is, the 1-dimensional workload is sufficient to determine delays at every exchange. Theorem 1 below 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

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

(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^*) satisfy (i)–(ii).

Proof. Suppose that (π^*, W^*) satisfy (i)–(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). Further, 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 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,

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

Combined with the fact that (π^*, Q^*) satisfies individual rationality condition (10), this implies that (π^*, W^*) satisfy (i). Further, if we sum up all N equations in the flow balance condition (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 trade-off 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 1-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 trade-off coefficients γ is non-atomic with a continuous and strictly positive density on the non-negative 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 (ii) ensures that the queueing system is stable ($\sum_{i=1}^N \lambda_i < \mu$) and leads to a non-trivial equilibrium where queue lengths are non-zero ($\mu < \Lambda + \sum_{i=1}^N \lambda_i$). 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 non-zero 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, i.e., those of type $\gamma \leq \gamma_0$, for some γ_0, chosen to satisfy (14), i.e.,

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

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

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

In order 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. *Suppose that (π^*, W^*) is an equilibrium and define γ_0 by* (18)*. Then,*

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

Further, 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 the Lemma 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 $\bar{i} \in \operatorname{argmax}_{j\neq 0}\kappa_j$, according to Lemma 1,

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

and therefore the equilibrium workload is $W^* = \gamma_0\mu\kappa_{\bar{i}}$. Theorem 2, whose proof can be found in the Online Supplement, summarizes the discussion above and characterizes the equilibrium.

Theorem 2 (Equilibrium Characterization)**.** *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

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

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

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

The above 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 multiple differentiated exchanges. One implication of Theorem 2 is that the equilibrium workload is unique, and that equilibrium routing policies are unique up to ties. Under an additional mild technical assumption, the following theorem (whose proof can be found in the Online Supplement) establishes that Q^* is unique:

Theorem 3 (Uniqueness of Equilibria). *Assume that the effective rebates* $\{\tilde{r}_i,\ i \neq 0\}$ *are distinct. Then, there is a unique equilibrium queue length vector* Q^*.

3.4. Convergence of Fluid Model to Equilibrium Configuration

Next we establish that the fluid model queue length vector $Q(t)$ converges to the unique equilibrium vector Q^* as $t \to \infty$. In addition to Assumption 1, we will assume the following:

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, i.e.,* $\tilde{r}_0 < \tilde{r}_1 < \cdots < \tilde{r}_N$.

Under Assumptions 1 and 2, Theorem 3 guarantees that Q^* is a unique equilibrium.

As in Section 2.3, define $\mathcal{G}_i(W(t)) \subset \mathbb{R}_+$ to be the set of optimizing limit order investor types γ that would prefer exchange i given a workload level[8] of $W(t)$, i.e., the set of all $\gamma \geq 0$ with

$$\gamma\tilde{r}_i - \frac{W(t)}{\mu\beta_i} \geq \gamma\tilde{r}_j - \frac{W(t)}{\mu\beta_j}, \quad \text{for all } j \notin \{0, i\}; \qquad \gamma\tilde{r}_i - \frac{W(t)}{\mu\beta_i} \geq \gamma\tilde{r}_0,$$

and the instantaneous fraction of arriving limit orders that are placed into exchange i as

$$\chi_i(W(t)) \triangleq \int_{\mathcal{G}_i(W(t))} dF(\gamma). \tag{24}$$

[8]Note that in Section 2.3, $\mathcal{G}_i(\cdot)$ and $\chi_i(\cdot)$ were defined to be functions of the vector of all queue lengths. However, since they depend on the queue length of each exchange only through the expected delay and therefore the workload, we will abuse notation and define these as functions of workload here.

Under Assumptions 1 and 2, (24) can be rewritten as

$$\chi_i(W) = \begin{cases} F\left(\frac{W\Gamma_i^+}{\mu}\right) - F\left(\frac{W\Gamma_i^-}{\mu}\right) & \text{if } \Gamma_i^+ \geq \Gamma_i^-, \\ 0 & \text{otherwise,} \end{cases} \tag{25}$$

where the constants Γ_i^+, Γ_i^- are defined by

$$\Gamma_i^+ \triangleq \begin{cases} \min\limits_{j>i} \ \frac{\beta_j^{-1} - \beta_i^{-1}}{\tilde{r}_j - \tilde{r}_i} & \text{if } i < N, \\ \infty & \text{if } i = N, \end{cases} \qquad \Gamma_i^- \triangleq \max\left\{ \frac{\beta_i^{-1}}{\tilde{r}_i - \tilde{r}_0}, \max\limits_{0<j<i} \ \frac{\beta_i^{-1} - \beta_j^{-1}}{\tilde{r}_i - \tilde{r}_j} \right\}.$$

Assumption 3. *Suppose that, for all $W > 0$,*

$$\sum_{i=1}^{N} \beta_i \frac{d\chi_i(W)}{dW} < 0. \tag{26}$$

Assumption 3 is essentially a local stability drift condition[9] that is easy to verify, and takes the form of a tail condition on $F(\cdot)$. Specifically, using (25), we have that:

$$\sum_{i=1}^{N} \beta_i \frac{d\chi_i(W)}{dW} = \sum_{i=1}^{N} \left(\frac{\Gamma_i^+}{\mu} f\left(\frac{W\Gamma_i^+}{\mu}\right) - \frac{\Gamma_i^-}{\mu} f\left(\frac{W\Gamma_i^-}{\mu}\right) \right) \mathbb{I}_{\{\Gamma_i^+ \geq \Gamma_i^-\}}, \tag{27}$$

where f is the density associated with F. A sufficient condition for Assumption 3 is that

$$t\Gamma_i^+ f(t\Gamma_i^+) < t\Gamma_i^- f(t\Gamma_i^-), \tag{28}$$

for all $t > 0$ and $1 \leq i \leq N$ such that $\Gamma_i^+ > \Gamma_i^-$. This expression can be easily verified in a particular problem instance, and it is satisfied for sufficiently broad class of distributions.

Definition 2 (Elastic Distribution)**.** *The cumulative distribution function F is elastic if $\gamma f(\gamma)$ is a strictly decreasing function over $\gamma \geq 0$.*

Examining (28), it is clear that elastic distributions will always satisfy Assumption 3. As an example, note that decreasing generalized failure rate distributions; see, e.g., Lariviere (2006), are included in the class of elastic distributions.

[9]The workload process evolves according to the differential equation $\dot{W}(t) = \sum_{i=1}^{N} \beta_i \dot{Q}_i(t) = \sum_{i=1}^{N} \beta_i \lambda_i + \Lambda \sum_{i=1}^{N} \beta_i \chi_i\big(W(t)\big) - \sum_{i=1}^{N} \beta_i \mu_i\big(Q(t)\big)$, which is itself a function of $W(t)$. In equilibrium, where $W(t) = W^*$, we have $\dot{W}(t) = 0$, i.e., $0 = \sum_{i=1}^{N} \beta_i \lambda_i + \Lambda \sum_{i=1}^{N} \beta_i \chi_i\big(W^*\big) - \sum_{i=1}^{N} \beta_i \mu_i\big(Q^*\big)$. Now, consider a small deviation from equilibrium of the form $Q(t) = (1+\epsilon)Q^*$ where ϵ is a small constant. Using the fact that $\mu_i\big((1+\epsilon)Q^*\big) = \mu_i(Q^*)$, the expression for $\dot{W}(t)$, and a Taylor approximation for small ϵ we get that $\dot{W}(t) = \sum_{i=1}^{N} \beta_i \lambda_i + \Lambda \sum_{i=1}^{N} \beta_i \chi_i\big((1+\epsilon)W^*\big) - \sum_{i=1}^{N} \beta_i \mu_i\big(Q^*\big) \approx \Lambda \epsilon W^* \sum_{i=1}^{N} \beta_i \frac{d\chi_i\left(W^*\right)}{dW}$. Assumption 3 guarantees that $\dot{W}(t) < 0$ when $\epsilon > 0$ and that $\dot{W}(t) > 0$ when $\epsilon < 0$. That is, it is necessary condition for local stability around W^*. Assumption 3 extends that condition to the entire state space.

In general, even under Assumptions 1 and 2, the queue lengths $Q(t)$ need not converge to the unique equilibrium Q^* — it is easy to construct numerical counterexamples. However, the following theorem illustrates that the additional condition of Assumption 3 is sufficient to guarantee convergence to equilibrium when there are $N = 2$ exchanges:

Theorem 4. *Suppose that there are $N = 2$ exchanges. Under Assumptions 1–3, given arbitrary initial conditions $Q(0) \in \mathbb{R}^N_+$, the queue lengths converge to the unique equilibrium Q^*.*

The proof of Theorem 4 can be found in the Online Supplement. For the $N > 2$ case, as discussed above, condition (26) is a necessary condition for local stability of the equilibrium Q^*. We conjecture that, as for $N = 2$, Assumption 3 is, in fact, also a sufficient condition when $N > 2$.

3.5. 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 marker order is increasing in its queue length. Both of these monotonicity conditions seem plausible even under different models of order routing optimization logic for arriving orders on both sides of the market, and would typically 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 order arrivals but relatively 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, while 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)–(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. 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 these predictions is, of course, predicated on the structure of (2) and (4)–(5) and the deterministic and stationary nature of the model. The sample of market data analyzed below captures more

complex and diverse trading behaviors, and is both stochastic and non-stationary. The statistical tests do not rely on the simplifying modeling assumptions, and the study over time will examine whether SSC holds in a pathwise sense; cf., footnote 3 in the introduction.

4.1. Overview of the Data Set

We use trade and quote (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 is from the nationally consolidated data feeds (i.e., the CTS, CQS, UTDF, and UQDF data feeds). Here, we identify the fluid volume at the bid (or, respectively, the ask) on a particular exchange as in the model of Section 2 with the total number of shares available on the bid (resp., ask) on that exchange. That is, we assume that the quantity at the bid or ask is made up of individual infinitesimal orders ignore the fact that the quantity actually arises from a collection of discrete, non-infinitesimal orders.

We consider the 30 component stocks of the Dow Jones Industrial Average 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.

We restrict attention to the $N = 6$ most liquid U.S. equity exchanges: NASDAQ, NYSE,[10] ARCA, EDGX, BATS, and EDGA. Smaller, regional exchanges were excluded as 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 9:45am and 3:45pm (i.e., excluding the opening 15 minutes and the closing 15 minutes).[11] 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, denote this 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, i.e., 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

[10] Note that 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.

[11] The time intervals should be sufficiently long so as to get reliable estimates of the event rates, and also long when compared to the event inter-arrival times, so that one could expect that the transient dynamics of the market due to changes in these rates settle down during these time intervals.

	Symbol	Listing Exchange	Price		Average Bid-Ask Spread	Volatility	Average Daily Volume
			Low ($)	High ($)	($)	(daily)	(shares, $\times 10^6$)
Alcoa	AA	NYSE	9.56	12.88	0.010	2.2%	27.8
American Express	AXP	NYSE	44.87	50.53	0.014	1.9%	8.6
Boeing	BA	NYSE	57.53	67.73	0.017	1.8%	5.9
Bank of America	BAC	NYSE	6.00	8.18	0.010	3.0%	258.8
Caterpillar	CAT	NYSE	72.60	92.83	0.029	2.3%	11.0
Cisco	CSCO	NASDAQ	14.96	16.84	0.010	1.7%	64.5
Chevron	CVX	NYSE	88.56	100.58	0.018	1.7%	11.1
DuPont	DD	NYSE	39.94	48.86	0.011	1.7%	10.2
Disney	DIS	NYSE	29.05	34.33	0.010	1.6%	13.3
General Electric	GE	NYSE	14.72	16.45	0.010	1.9%	84.6
Home Depot	HD	NYSE	31.08	35.33	0.010	1.6%	13.4
Hewlett-Packard	HPQ	NYSE	21.50	26.46	0.010	2.2%	32.5
IBM	IBM	NYSE	158.76	180.91	0.060	1.5%	6.6
Intel	INTC	NASDAQ	19.16	22.98	0.010	1.5%	63.6
Johnson & Johnson	JNJ	NYSE	61.00	66.14	0.011	1.2%	12.6
JPMorgan	JPM	NYSE	28.53	37.82	0.010	2.2%	49.1
Kraft	KFT	NYSE	32.70	35.52	0.010	1.1%	10.9
Coca-Cola	KO	NYSE	66.62	71.77	0.011	1.1%	12.3
McDonalds	MCD	NYSE	83.65	91.09	0.014	1.2%	7.9
3M	MMM	NYSE	71.71	83.95	0.018	1.6%	5.5
Merck	MRK	NYSE	30.71	33.49	0.010	1.3%	17.6
Microsoft	MSFT	NASDAQ	24.60	27.50	0.010	1.5%	61.0
Pfizer	PFE	NYSE	17.30	19.15	0.010	1.5%	47.7
Procter & Gamble	PG	NYSE	60.30	64.70	0.011	1.0%	11.2
AT&T	T	NYSE	27.29	29.18	0.010	1.2%	37.6
Travelers	TRV	NYSE	46.64	51.54	0.013	1.6%	4.8
United Tech	UTX	NYSE	67.32	77.58	0.018	1.7%	6.2
Verizon	VZ	NYSE	34.65	37.39	0.010	1.2%	18.4
Wal-Mart	WMT	NYSE	49.94	53.55	0.010	1.1%	13.1
Exxon Mobil	XOM	NYSE	67.93	74.98	0.011	1.6%	26.2

Table 1: Descriptive statistics for the 30 stocks over the 21 trading days of September 2011. The average bid-ask spread is a time average computed from our TAQ data set. The volatility is an average of daily volatilities over Sept 2011. All the other statistics were retrieved from Yahoo Finance.

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

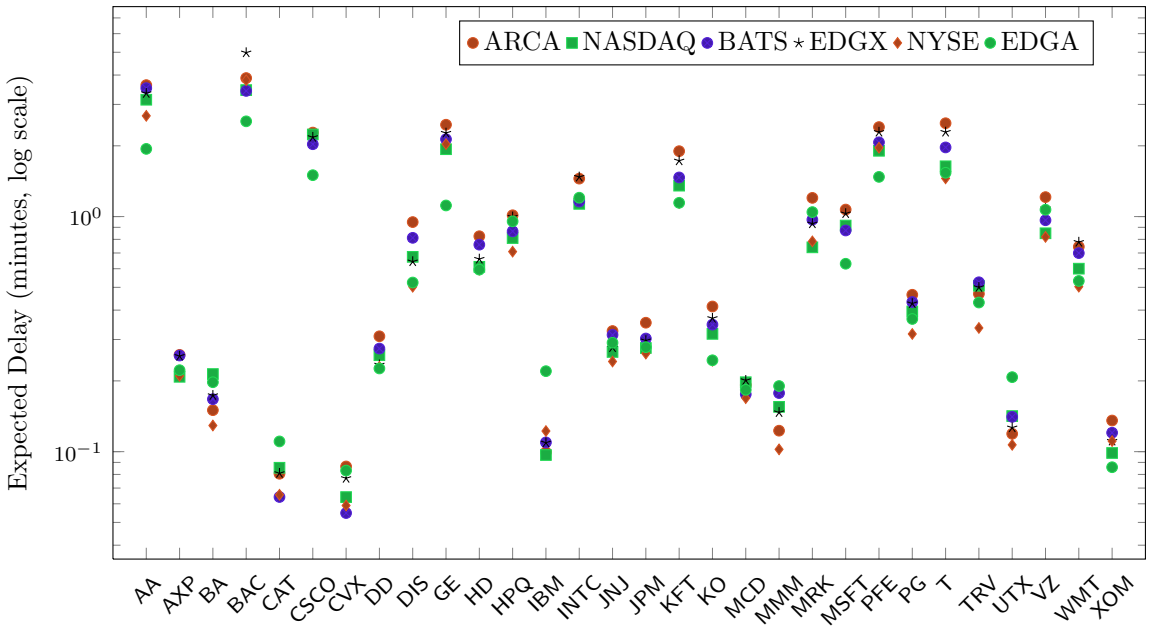
Table 2: Rebates and fees of the 6 major U.S. stock exchanges during the September 2011 period, per share traded. †Rebates on NASDAQ and ARCA are subject to "tiering": higher rebates than the ones quoted may be available to traders that contribute significant volume to the respective exchange.

$$\mathsf{ED}_i^{(s,j)}(t) \triangleq \frac{Q_i^{(s,j)}(t)}{\mu_i^{(s,j)}(t)}. \tag{29}$$

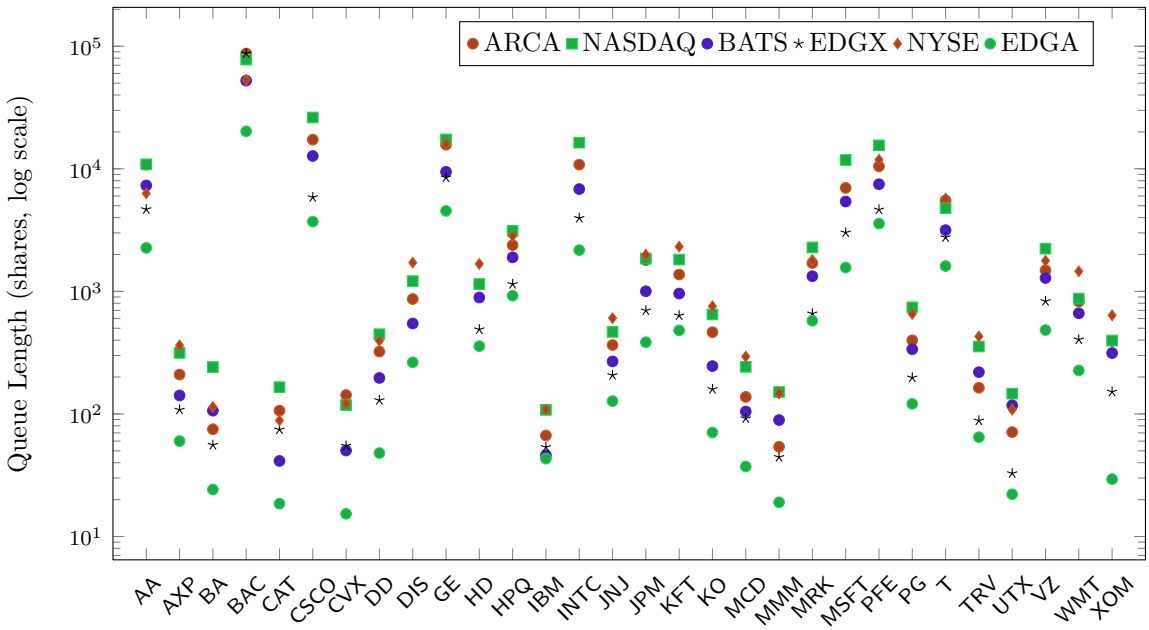
The above expression disregards the effect of order cancellations from the bid and ask queues, and serves as a practical proxy for expected delay that is commonly used in trading systems. For each stock and each exchange, Figure 2(a) shows the expected delay, averaged across time slots and the bid and ask sides of the market. Delays range from 5 seconds to about 5 minutes across the 30 stocks we studied, and we observe 2x to 3x 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 market. Queue lengths range from 10 to 100,000 shares across securities, and exhibit about a 10x variation in the queue sizes across exchanges for the same security. Deeper queues correspond to longer delays.

Principle component analysis (PCA). The state space collapse result of our model predicts that delays are coupled across exchanges and are restricted to a 1-dimensional subspace. Define the empirically observed expected delay vector trajectories $\left\{\mathsf{ED}^{(s,j)}(t) \ : \ t = 1, \ldots, T; \ s = \mathsf{BID}, \mathsf{ASK}\right\}$, where $\mathsf{ED}^{(s,j)}(t)$ was estimated in (29) 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 PCA by examining the number of principle components necessary to explain the 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 around 80% of the variability of the expected delays across exchanges, and that the first two principle components explain about 90%. This is consistent with the hypothesis of low effective dimension.

Intuitively, in the high flow environment of our observation universe, i.e., where Λ and μ are large, queue length deviations from the equilibrium configuration would be quickly erased as new limit or market order arrivals would make decisions taking into account relative opportunities. The end result is that the queue lengths at the various exchanges stay close to their equilibrium configurations, and that order routing optimization leads to coupling in the state of the various exchanges. The equilibrium state itself changes over time as the rates of events change, but the coupling across exchanges remains strong, and persists even if we shorten the time period over which market statistics are averaged from 1 hour down to 15 minutes. 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), while the first two principle components explains 82% of the variability.



(a) Average expected delay across stocks and exchanges.



(b) Average queue length (number of shares at the NBBO) across stocks and exchanges.

Figure 2: Averages of hourly estimates of the expected delays and queue lengths for the Dow 30 stocks on the 6 exchanges during September 2011. 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.

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

Table 3: Results of PCA: how much variance in the data can the first two principle components explain.

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 relationship

$$\mu_i^{(s,j)}(t) = \mu^{(s,j)}(t)\ \frac{\beta_i^{(j)} Q_i^{(s,j)}}{\sum_{i'=1}^{N} \beta_{i'}^{(j)} Q_{i'}^{(s,j)}}, \tag{30}$$

where $\beta_i^{(j)}$ is the attraction coefficient for security j on exchange i. Note that 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 on (30). The results are given in Table 4. Note that all attraction coefficient estimates are statistically significant.

4.3. Empirical Evidence of State Space Collapse

At its core, 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, which is not the consolidated depth at the bid or ask.

Verification of linear dependence of expected delays via regression analysis. Define $W^{(s,j)}(t)$

	Attraction Coefficient					
	ARCA	NASDAQ	BATS	EDGX	NYSE	EDGA
Alcoa	0.73	0.87	0.76	0.81	1.00	1.33
American Express	1.19	1.08	0.99	0.94	1.00	0.94
Boeing	0.95	0.67	0.81	0.74	1.00	0.73
Bank of America	0.94	1.04	1.01	0.77	1.00	1.43
Caterpillar	0.82	0.78	1.13	0.70	1.00	0.58
Cisco	0.95	1.00	1.06	0.98	-	1.45
Chevron	0.70	0.93	1.17	0.65	1.00	0.75
DuPont	0.90	0.98	0.98	1.03	1.00	1.00
Disney	0.69	0.88	0.78	0.88	1.00	1.04
General Electric	0.79	1.01	0.94	0.73	1.00	1.63
Home Depot	0.76	0.98	0.79	0.84	1.00	1.02
Hewlett-Packard	1.04	1.04	1.02	0.68	1.00	0.82
IBM	1.25	1.20	1.20	1.05	1.00	0.54
Intel	0.83	1.00	0.96	0.84	-	1.04
Johnson & Johnson	0.80	0.94	0.86	0.92	1.00	0.77
JPMorgan	0.78	0.99	0.93	0.84	1.00	0.91
Kraft	0.72	0.89	0.83	0.73	1.00	1.06
Coca-Cola	0.68	0.84	0.79	0.76	1.00	0.88
McDonalds	0.90	0.86	1.03	0.82	1.00	0.82
3M	0.89	0.67	0.62	0.66	1.00	0.57
Merck	0.68	1.01	0.83	0.90	1.00	0.81
Microsoft	0.83	1.00	1.02	0.95	-	1.41
Pfizer	0.84	1.01	0.96	0.87	1.00	1.29
Procter & Gamble	0.79	0.89	0.88	0.89	1.00	0.89
AT&T	0.62	0.94	0.75	0.59	1.00	1.00
Travelers	0.80	0.69	0.69	0.84	1.00	0.80
United Tech	1.18	0.89	0.79	0.87	1.00	0.53
Verizon	0.77	0.95	0.88	0.72	1.00	0.85
Wal-Mart	0.72	0.88	0.79	0.71	1.00	0.91
Exxon Mobil	0.89	1.13	0.97	0.89	1.00	1.35

Table 4: Estimates of the attraction coefficients β_i from nonlinear regression. Note that the attraction coefficient of the listing exchange is normalized to be 1.

to be the workload for side s of security j in time slot t, i.e.,

$$W^{(s,j)}(t) \triangleq \sum_{i=1}^{N} \beta_i^{(j)} Q_i^{(s,j)}(t), \tag{31}$$

and observe that the vector of expected delays can be written as

$$\mathsf{ED}^{(s,j)}(t) = \frac{W^{(s,j)}(t)}{\mu^{(s,j)}(t)} \left(\frac{1}{\beta_1^{(j)}}, \ldots, \frac{1}{\beta_N^{(j)}} \right). \tag{32}$$

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,

$$\mathsf{ED}_i^{(s,j)}(t) = \frac{\beta_{i'}^{(j)}}{\beta_i^{(j)}} \, \mathsf{ED}_{i'}^{(s,j)}(t), \tag{33}$$

for each time slot t.

To test this prediction, for each security, we perform a linear regression of the left side of (33), which is the expected delay of that security on a particular exchange, as a function of the right side of (33), which is the expected delay on a benchmark exchange (ARCA) rescaled by the ratio of the attraction coefficients of the two exchanges. Scaled in this way, the regression slope predicted by (33) is 1. The regression is performed using the expected delay measurements outlined in (29), i.e., 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.

The results of these regressions are summarized in Table 5.

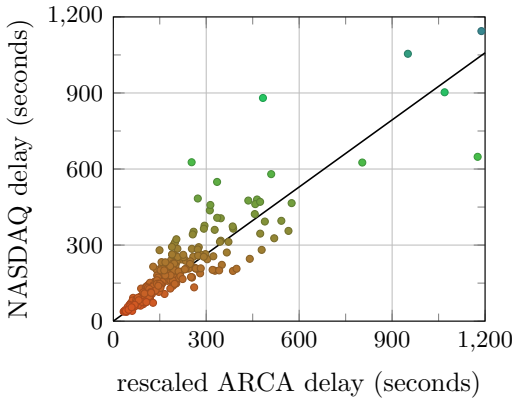
	Dependent Variable: $\mathsf{ED}_{\text{exchange}}$				
	NASDAQ OMX	BATS	DirectEdge X	NYSE	DirectEdge A
Intercept	$6.96 \times 10^{-4***}$	$1.27 \times 10^{-3***}$	-1.02×10^{-4}	$-4.60 \times 10^{-4***}$	$9.42 \times 10^{-4***}$
	(1.14×10^{-4})	(1.09×10^{-4})	(2.02×10^{-4})	(1.60×10^{-4})	(1.05×10^{-4})
Rescaled $\mathsf{ED}_{\text{ARCA}}$	0.92^{***}	0.89^{***}	0.97^{***}	0.98^{***}	0.87^{***}
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
R^2	85%	87%	76%	77%	79%
Note:					$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 5: Linear regressions of the expected delay on a particular exchange, versus that of the benchmark exchange (ARCA) rescaled by the ratio of the attraction coefficients of the two exchanges.

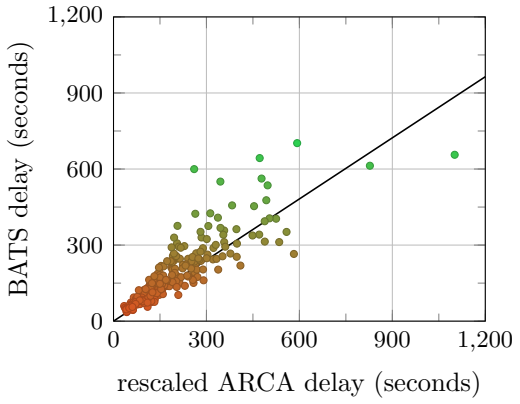
The R^2 varies between 76% and 87% across the five exchanges. Moreover, all of the regressions are statistically significant and we are able to reject the null hypothesis that the delay on a particular exchange has a zero regression coefficient relative to the rescaled delay on ARCA. These results statistically verify the linear dependence of delays across different exchanges suggested by (33). Note that (33) further predicts that the regression should have a zero intercept and the slope of the rescaled $\mathsf{ED}_{\text{ARCA}}$ 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.

While 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.

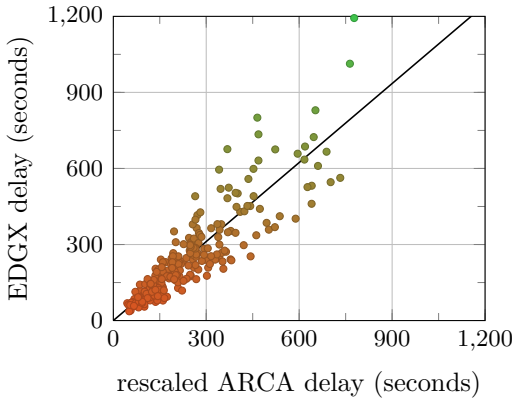
Residual analysis and accuracy of delay estimates based on the aggregate workload. The SSC result culminated in a specific relationship (32) that makes expected delay predictions in each exchange based on the 1-dimensional aggregated workload process. Specifically, given the market model coefficients $\beta_i^{(j)}$ and a measurement of the queue sizes at the various exchanges, $Q_i^{(s,j)}(t)$, one can compute the workload via (31), and then construct estimates for the expected delays at the various exchanges via (32). We denote the resulting delay estimates by $\hat{\mathsf{ED}}^{(s,j)}(t)$, where the $\hat{}$



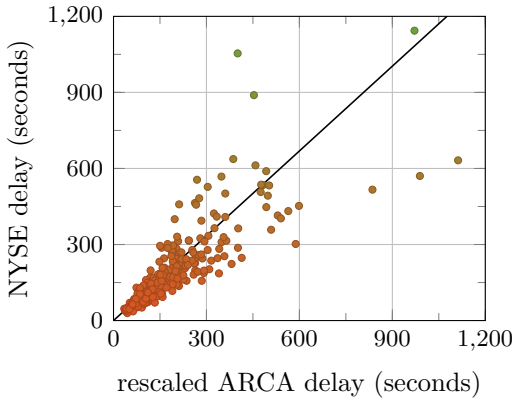
(a) slope $= 0.88$, intercept $= 6 \times 10^{-3}$, $R^2 = 84\%$



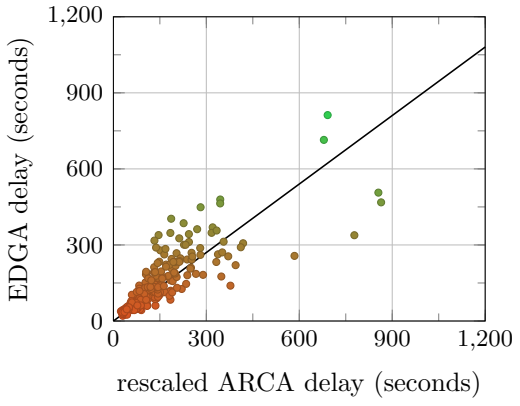
(b) slope $= 0.80$, intercept $= 9 \times 10^{-3}$, $R^2 = 79\%$



(c) slope $= 1.04$, intercept $= 9 \times 10^{-4}$, $R^2 = 71\%$



(d) slope $= 1.11$, intercept $= -4 \times 10^{-3}$, $R^2 = 63\%$



(e) slope $= 0.90$, intercept $= 4 \times 10^{-3}$, $R^2 = 73\%$

Figure 3: Scatter plots of the expected delay for Bank of America (BAC) on each exchange, versus the delay on ARCA rescaled by the ratio of the attraction coefficients of the two exchanges. The black lines correspond to linear regressions with intercept.

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

Table 6: The measure of performance R^2_*, which given the reduction of variability in expected delays explained by the workload relationship (32).

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 (29). This prediction can be tested again through a set of linear regressions between the workload delay estimate and the delay estimate that uses information about the state of the exchange (queue length and trading rate). All these regressions are again 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, i.e., 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 (32), 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 (32). Numerical results for R^2_* across securities are given in Table 6. Typical values for R^2_* are around 80%, highlighting the predictive power of the one-dimensional workload model as a means of capturing the state of the decentralized fragmented market.

Our analysis showed that optimized order routing couples the exchange dynamics in terms of their delay estimates as opposed to their queue depths. As mentioned earlier in discussing Figure 2(b), queue lengths across exchanges exhibit significantly more variation than their corresponding delays. One could repeat the above analysis, for example, starting from trying to see whether the queue length processes live on a lower dimensional manifold, similarly to what we observed in studying the respective delay estimates. Not surprisingly, and as suggested through our analysis and the above comments, the PCA of the queue length trajectories yields weaker results, and similarly all of the subsequent tests lead to noticeably lower quality of fit. Our model suggests two explanations: (a) the limit order routing logic seems to rely on delay estimates as opposed to queue lengths; and, (b) the model capturing the routing of market orders is itself nonlinear. Both

(a) and (b) hinge on some of our modeling assumptions that build on insight from practical smart order router optimization logic, where indeed limit order placement decisions depend crucially on delays or fill probabilities, and market order routing follow variants of fee minimization arguments that depend nonlinearly on the displayed quantities.

Finally, it is worth remarking that this seems to be one of the first examples of a complex stochastic network model, where state space collapse has been empirically verified.

References

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

G. Allon and A. Federgruen. Competition in service industries. *Operations Research*, 55(1):37–55, 2007.

M. Armony and M. Haviv. Price and delay competition between two service providers. *European Journal of Operational Research*, 147:32–50, 2003.

M. J. Barclay, T. Hendershott, and T. D. McCormick. Competition among trading venues: Information and trading on electronic communications networks. *Journal of Finance*, 58:2637–2666, 2003.

A. Bassamboo, J.M. Harrison, and A. Zeevi. Dynamic routing in large call centers: Asymptotic analysis of an LP-based method. *Operations Research*, 10:1074–1099, 2004.

H. Bessembinder. Quote-based competition and trade execution costs in NYSE listed stocks. *Journal of Financial Economics*, 70:385–422, 2003.

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

J. Blanchet and X. Chen. Continuous-time modeling of bid-ask spread and price dynamics in limit order books. Working paper, 2013.

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

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

S. Buti, B. Rindi, Y. Wen, and I.M. Werner. Tick size regulation, intermarket competition and sub&penny trading. Working paper, 2011.

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

Y-J. Chen, C. Maglaras, and G. Vulcano. Design of an aggregated marketplace under congestion effects: Asymptotic analysis and equilibrium characterization. Working paper, 2010.

R. Cont and A. Kukanov. Optimal order placement in limit order markets. Working paper, 2013.

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

R. Cont, S. Stoikov, and R. Talreja. A stochastic model for order book dynamics. *Operations Research*, 58: 549–563, 2010.

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

A. Dufour and R. F. Engle. Time and the price impact of a trade. *Journal of Finance*, 55:2467–2498, 2000.

T. Foucault and A. J. Menkveld. Competition for order flow and smart order routing systems. *Journal of Finance*, 63:119–158, 2008.

T. Foucault, O. Kadan, and E. Kandel. Limit order book as a market for liquidity. *Review of Financial Studies*, 18:1171–1217, 2005.

J. Gatheral. No-dynamic-arbitrage and market impact. *Quantitative Finance*, 10:749–759, 2010.

L. Glosten. Is the electronic order book inevitable? *Journal of Finance*, 49:1127–1161, 1994.

L. Glosten. Competition, design of exchanges and welfare. Working paper, 1998.

L. R. Glosten. Components of the bid/ask spread and the statistical properties of transaction prices. *Journal of Finance*, 42:1293–1307, 1987.

L. R. Glosten and P. R. Milgrom. Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14:71–100, 1985.

M. D. Griffiths, B. F. Smith, D. A. S. Turnbull, and R. W. White. The costs and the determinants of order aggressiveness. *Journal of Financial Economics*, 56:65–88, 2000.

X. Guo, A. De Larrard, and Z. Ruan. Optimal placement in a limit order book. Working paper, 2013.

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

J. M. Harrison. Brownian models of queueing networks with heterogeneous customer populations. In W. Fleming and P. L. Lions, editors, *Stochastic Differential Systems, Stochastic Control Theory and Applications*, volume 10 of *Proceedings of the IMA*, pages 147–186. Springer-Verlag, New York, 1988.

J. M. Harrison. Balanced fluid models of multiclass queueing networks: a heavy traffic conjecture. In F. Kelly and R. Williams, editors, *Stochastic Networks*, volume 71, pages 1–20. Proceedings of the IMA, 1995.

J. M. Harrison. Brownian models of open processing networks: Canonical representation of workload. *Ann. Appl. Prob.*, 10:75–103, 2000.

J. M. Harrison and M. J. Lopez. Heavy traffic resource pooling in parallel- server systems. *Queueing Systems*, 33:339–368, 1999.

J. M. Harrison and J. A. Van Mieghem. Dynamic control of brownian networks: State space collapse and equivalent workload formulations. *Ann. Appl. Prob.*, 7:747–771, 1996.

R. Hassin and M. Haviv. *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston, MA, 2003.

B. Hollifield, R. A. Millerz, and P. Sandas. Empirical analysis of limit order markets. *Review of Economic Studies*, 71:1027–1063, 2004.

R. W. Holthausen, R. W. Leftwich, and D. Mayers. Large-block transactions, the speed of response, and temporary and permanent stock-price effects. *Journal of Financial Economics*, 26:71–95, 1990.

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

B. Jovanovic and A. J. Menkveld. Middlemen in limit-order markets. Working paper, 2011.

D. B. Keim and A. Madhavan. The cost of instituional equity trades. *Financial Analysts Journal*, 54:50–59, 1998.

A. S. Kyle. Continuous auctions and insider trading. *Econometrica*, 53:1315–1335, 1985.

P. Lakner, J. Reed, and S. Stoikov. High frequency asymptotics for the limit order book. Working paper, 2013.

P. Lakner, J. Reed, and F. Simatos. Scaling limit of a limit order book model via the regenerative characterization of lévy trees. Working paper, 2014.

M. A Lariviere. A note on probability distributions with increasing generalized failure rates. *Operations Research*, 54(3):602–604, 2006.

P. Lederer and L. Li. Pricing, production, scheduling and delivery -time competition. *Operations Research*, 45:407–420, 1997.

D. Levhari and I. Luski. Duopoly prcing and waiting lines. *European Economic Review*, 11:17–35, 1978.

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

C. Loch. *Pricing in markets sensitive to delay.* Ph.D. dissertation, Stanford University, Stanford, CA, 1991.

I. Luski. On partial equilibrium in a queueing system with two servers. *The Review of Economic Studies*, 43:519–525, 1976.

C. Maglaras and C. Moallemi. A multiclass queueing model of limit order book dynamics. Working paper, 2011.

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

A. Mandelbaum and G. Pats. State-dependent queues: approximations and applications. In F. Kelly and R. Williams, editors, *Stochastic Networks*, volume 71, pages 239–282. Proceedings of the IMA, 1995.

A. Mandelbaum and G. Pats. State-dependent stochastic networks. Part I: Approximations and applications with continuous diffusion limits. *Ann. Appl. Prob.*, 8(2):569–646, 1998.

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

S. P. Meyn. Sequencing and routing in multiclass queueing networks: Part I: feedback regulation. *SIAM J. on Control and Optimization*, 40(3):741–776, 2001.

A. Obizhaeva and J. Wang. Optimal trading strategy and supply/demand dynamics. Working paper, 2006.

M. O'Hara and M. Ye. Is market fragmentation harming market quality? *Journal of Financial Economics*, 100(3):459–474, June 2011.

C. Parlour. Limit order markets: A survey. *Handbook of Financial Intermediation & Banking A.W.A. Boot and A. V. Thakor eds.*, 2008.

C. A. Parlour. Price dynamics in limit order markets. *Review of Financial Studies*, 11:789–816, 1998.

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

I. Rosu. A dynamic model of the limit order book. *Review of Financial Studies*, 22:4601–4641, 2009.

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

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

G. Sofianos, J. Xiang, and A. Yousefi. Smart order routing: All-in shortfall and optimal order placement. *Goldman Sachs, Equity Executions Strats, Street Smart*, 42, 2011.

S. Stoikov, M. Avellaneda, and J. Reed. Forecasting prices from level-i quotes in the presence of hidden liquidity. *Algorithmic Finance, Forthcoming*, 2011.

A. L. Stolyar. Optimal routing in output-queued flexible server systems. *Probability in the Engineering and Informational Sciences*, 19:141 – 189, 2005.

V. van Kervel. Liquidity: What you see is what you get? Working paper, 2012.

S. H. Zak. *Systems and Control*. Oxford University Press, 2003.

A. Proofs: Equilibrium Characterization

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

$$\max_{i\neq 0}\ \gamma_0(\tilde{r}_i - \tilde{r}_0) - \frac{W^*}{\mu\beta_i} = 0. \tag{A.1}$$

Further, suppose that for a given W^*, (A.1) *holds, and for each exchange* i*, define*

$$\kappa_i \triangleq \beta_i(\tilde{r}_i - \tilde{r}_0). \tag{A.2}$$

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

Proof. For $\gamma \geq 0$, define

$$\mathcal{L}(\gamma) \triangleq \max_{i\neq 0}\ \gamma(\tilde{r}_i - \tilde{r}_0) - \frac{W^*}{\mu\beta_i}.$$

Clearly $\mathcal{L}$ is a continuous function, and under Assumption 1(iii), it is also increasing. We wish to show that $\mathcal{L}(\gamma_0) = 0$.

Suppose that $\mathcal{L}(\gamma_0) < 0$. Then, there exists $\bar{\gamma} > \gamma_0$ with $\mathcal{L}(\gamma) < 0$ for all $\gamma \in [0, \bar{\gamma}]$. Thus, in equilibrium, investors with types $\gamma \in [0, \bar{\gamma}]$ strictly prefer placing market orders, i.e., $\pi_i^*(\gamma) = 0$ for $i \neq 0$. Then,

$$\begin{aligned}\sum_{i=1}^{N}\left(\lambda_i + \Lambda\int_0^\infty \pi_i^*(\gamma)\,dF(\gamma)\right) &= \sum_{i=1}^{N}\lambda_i + \Lambda\int_{\bar{\gamma}_0}^\infty\left(\sum_{i=1}^{N}\pi_i^*(\gamma)\right)\,dF(\gamma)\\ &\leq \sum_{i=1}^{N}\lambda_i + \Lambda\big(1 - F(\bar{\gamma})\big) < \mu,\end{aligned}$$

where the last inequality follows from (17) and Assumption 1(i). This contradicts the flow balance equation (14).

Alternatively, suppose that $\mathcal{L}(\gamma_0) > 0$. Then, there exists $\bar{\gamma} < \gamma_0$ with $\mathcal{L}(\gamma) > 0$ for all $\gamma \in [\bar{\gamma}, \infty)$. Thus, in equilibrium, investors with types $\gamma \in [\bar{\gamma}, \infty)$ strictly prefer *not* placing market orders, i.e., $\pi_0^*(\gamma) = 0$. Then,

$$\begin{aligned}\sum_{i=1}^{N}\left(\lambda_i + \Lambda\int_0^\infty \pi_i^*(\gamma)\,dF(\gamma)\right) &= \sum_{i=1}^{N}\lambda_i + \Lambda\int_0^\infty\left(1 - \pi_0^*(\gamma)\right)\,dF(\gamma)\\ &\geq \sum_{i=1}^{N}\lambda_i + \Lambda\big(1 - F(\bar{\gamma})\big) > \mu,\end{aligned}$$

where the last inequality follows from (17) and Assumption 1(i). This contradicts the flow balance equation (14). Thus, we must have $\mathcal{L}(\gamma_0) = 0$ and (A.1) holds.

Now, suppose exchange i achieves the maximum in (A.1). Then, from the right side of (A.1),

it follows that

$$\kappa_i = \beta_i(\tilde{r}_i - \tilde{r}_0) = \frac{W^*}{\mu\gamma_0}.$$

Further, for any exchange j, (A.1) implies that

$$\kappa_j = \beta_j(\tilde{r}_j - \tilde{r}_0) \leq \frac{W^*}{\mu\gamma_0} = \kappa_i.$$

For the converse, if

$$\kappa_i = \max_{j\neq 0} \ \kappa_j, \tag{A.3}$$

and there exists an exchange j satisfying

$$0 = \gamma_0(\tilde{r}_j - \tilde{r}_0) - \frac{W^*}{\mu\beta_j} > \gamma_0(\tilde{r}_i - \tilde{r}_0) - \frac{W^*}{\mu\beta_i},$$

then

$$\kappa_j = \beta_j(\tilde{r}_j - \tilde{r}_0) = \frac{W^*}{\mu\gamma_0} > \beta_i(\tilde{r}_i - \tilde{r}_0) = \kappa_i,$$

which contradicts with (A.3). ■

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

$$W^* \triangleq \gamma_0\mu \max_{i\neq 0} \ \kappa_i, \tag{A.4}$$

and

$$\begin{aligned} \pi_0^*(\gamma) &= 1, \qquad \text{for all } \gamma < \gamma_0, \\ \pi_i^*(\gamma) &= 0, \qquad \text{for all } i \notin \mathcal{A}^*(\gamma_0) \cup \{0\}, \\ \pi_i^*(\gamma) &= 0, \qquad \text{for all } \gamma > \gamma_0,\ i \notin \mathcal{A}^*(\gamma), \end{aligned} \tag{A.5}$$

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

Conversely, suppose that $(\pi^*, W^*) \in \mathcal{P} \times \mathbb{R}_+$ *is an equilibrium, i.e., it satisfies* (13)-(14). *Then,* W^* *must satisfy* (A.4) *and* π^* *must satisfy* (A.5)*, except possibly for* γ *in a set of* F*-measure zero.*

Proof. Suppose (π^*, W^*) satisfies (A.4)–(A.5). We want to show that (π^*, W^*) is an equilibrium, i.e., it must satisfy (13)–(14).

We first establish (13). In particular, we will establish that for any $\pi \in \mathcal{P}$ and all γ,

$$\pi_0(\gamma)\gamma\tilde{r}_0 + \sum_{i=1}^{N} \pi_i(\gamma)\left(\gamma\tilde{r}_i - \frac{W^*}{\mu\beta_i}\right) \leq \pi_0^*(\gamma)\gamma\tilde{r}_0 + \sum_{i=1}^{N} \pi_i^*(\gamma)\left(\gamma\tilde{r}_i - \frac{W^*}{\mu\beta_i}\right).$$

Equivalently,

$$\sum_{i=1}^{N} \pi_i(\gamma) \left(\gamma \left(\tilde{r}_i - \tilde{r}_0 \right) - \frac{W^*}{\mu \beta_i} \right) \leq \sum_{i=1}^{N} \pi_i^*(\gamma) \left(\gamma \left(\tilde{r}_i - \tilde{r}_0 \right) - \frac{W^*}{\mu \beta_i} \right). \tag{A.6}$$

If $\gamma \leq \gamma_0$ and $i \neq 0$, using (A.4) and Assumption 1(iii), we have that

$$\gamma \left(\tilde{r}_i - \tilde{r}_0 \right) - \frac{W^*}{\mu \beta_i} = \frac{\gamma \beta_i \kappa_i - \gamma_0 \max_{j \neq 0} \kappa_j}{\beta_i} \leq \frac{\gamma_0 \beta_i \kappa_i - \gamma_0 \max_{j \neq 0} \kappa_j}{\beta_i} \leq 0 \tag{A.7}$$

Since, by (A.5), $\pi_i^*(\gamma) = 0$ for $i \neq 0$, we have that (A.6) holds for all $\gamma < \gamma_0$. For $\gamma = \gamma_0$, note that equality holds in (A.7) iff $\kappa_i = \max_{j \neq 0} \kappa_j$, i.e., $i \in \mathcal{A}^*(\gamma_0)$. Thus, (A.6) also holds for $\gamma = \gamma_0$. Finally, if $\gamma > \gamma_0$ and $i \neq 0$,

$$\gamma \left(\tilde{r}_i - \tilde{r}_0 \right) - \frac{W^*}{\mu \beta_i} = \frac{\gamma \kappa_i - \gamma_0 \max_{j \neq 0} \kappa_j}{\beta_i} \geq \frac{\gamma \kappa_i - \gamma \max_{j \neq 0} \kappa_j}{\beta_i} \geq 0. \tag{A.8}$$

Thus, (A.6) continues to hold.

Next, we establish (14). By (A.5), $1 - \pi_0^*(\gamma) = 0$ when $\gamma < \gamma_0$ and $1 - \pi_0^*(\gamma) = 1$ when $\gamma > \gamma_0$. Thus,

$$\int_0^\infty \left(1 - \pi_0^*(\gamma)\right) dF(\gamma) = \int_{\gamma_0}^\infty dF(\gamma) = 1 - F(\gamma_0).$$

Using this and (17),

$$\begin{aligned}
\mu &= \sum_{i=1}^{N} \lambda_i + \Lambda \int_0^\infty \left(1 - \pi_0^*(\gamma)\right) dF(\gamma) \\
&= \sum_{i=1}^{N} \lambda_i + \Lambda \int_0^\infty \left(\sum_{i=1}^{N} \pi_i^*(\gamma) \right) dF(\gamma) \\
&= \sum_{i=1}^{N} \left(\lambda_i + \Lambda \int_0^\infty \pi_i^*(\gamma) \, dF(\gamma) \right).
\end{aligned}$$

Thus, (π^*, W^*) satisfies (14) as well and is an equilibrium.

Now suppose (π^*, W^*) is an equilibrium. We would like to show that (π^*, W^*) must satisfy (A.4)–(A.5), except possibly for γ in a set of F-measure zero.

First, by Lemma 1, we have that

$$\gamma_0 \tilde{r}_0 = \max_{i \neq 0} \ \gamma_0 \tilde{r}_i - \frac{W^*}{\mu \beta_i} = \gamma_0 \tilde{r}_{\bar{i}} - \frac{W^*}{\mu \beta_{\bar{i}}},$$

where $\bar{i} \in \operatorname{argmax}_{j \neq 0} \ \kappa_j$. By solving for W^*, (A.4) follows immediately.

Next, we verify (A.5). Define $\mathcal{M}$ to be the set of $\gamma \geq 0$ such that $\pi^*(\gamma)$ does not satisfy (A.5). Define $\bar{\pi} \in \mathcal{P}$ to be a set of routing decisions such that $(\bar{\pi}, W^*)$ satisfies (A.5), such a $\bar{p}i$ can easily

be constructed by solving the optimization problem for $\mathcal{A}^*(\gamma)$ for each $\gamma \geq 0$. Define

$$\begin{aligned}\Delta(\gamma) &\triangleq \pi_0^*(\gamma)\tilde{r}_0 + \sum_{i=1}^N \pi_i^*(\gamma)\left(\gamma\tilde{r}_i - \frac{W^*}{\mu\beta_i}\right) - \bar{\pi}_0(\gamma)\tilde{r}_0 - \sum_{i=1}^N \bar{\pi}_i(\gamma)\left(\gamma\tilde{r}_i - \frac{W^*}{\mu\beta_i}\right)\\ &= \sum_{i=1}^N \pi_i^*(\gamma)\left(\gamma\left(\tilde{r}_i - \tilde{r}_0\right) - \frac{W^*}{\mu\beta_i}\right) - \sum_{i=1}^N \bar{\pi}_i(\gamma)\left(\gamma\left(\tilde{r}_i - \tilde{r}_0\right) - \frac{W^*}{\mu\beta_i}\right),\end{aligned}$$

for $\gamma \geq 0$. Following the same arguments as in (A.7)–(A.8), it is easy to see that

$$\begin{aligned}&\Delta(\gamma) = 0 \qquad \text{if } \gamma \notin \mathcal{M},\\ &\Delta(\gamma) < 0 \qquad \text{if } \gamma \in \mathcal{M} \text{ and } \gamma \neq \gamma_0.\end{aligned} \tag{A.9}$$

On the other hand, Since π^* is optimal for the program (13), we have that

$$0 \leq \int_0^\infty \Delta(\gamma)\, dF(\gamma) = \int_{\mathcal{M}} \Delta(\gamma)\, dF(\gamma) = \int_{\mathcal{M}\cap[0,\gamma_0)} \Delta(\gamma)\, dF(\gamma) + \int_{\mathcal{M}\cap(\gamma_0,\infty)} \Delta(\gamma)\, dF(\gamma), \tag{A.10}$$

where, for the final equality, we use the fact that the point $\{\gamma_0\}$ has F-measure zero under Assumption 1(i). Together, (A.9)–(A.10) imply that $\mathcal{M}$ has F-measure 0. ■

Theorem 3 (Uniqueness of Equilibria). *Assume that the effective rebates $\{\tilde{r}_i,\ i \neq 0\}$ are distinct. Then, there is a unique equilibrium queue length vector Q^*.*

Proof. Suppose $(\pi^{(1)}, Q^{(1)})$ and $(\pi^{(2)}, Q^{(2)})$ are both equilibria. Define $W^{(\ell)} \triangleq \beta^\top Q^{(\ell)}$, for $\ell \in \{1, 2\}$. By Theorem 1, both $(\pi^{(1)}, W^{(1)})$ and $(\pi^{(2)}, W^{(2)})$ satisfy (13)-(14). By Theorem 2, we have that

$$W^{(1)} = W^{(2)} = W^* \triangleq \gamma_0\mu \max_{i\neq 0}\ \kappa_i. \tag{A.11}$$

Now, suppose that $\gamma < \gamma_0$. Theorem 2 states that $\pi_i^{(1)}(\gamma) = \pi_i^{(2)}(\gamma) = 0$ for $i \neq 0$, except possibly on a set of γ of F-measure zero. On the other hand, if $\gamma > \gamma_0$, by Theorem 2, $\pi^{(1)}(\gamma)$ and $\pi^{(2)}(\gamma)$ can only differ when $\mathcal{A}^*(\gamma)$ contains at least two exchanges (ignoring a set of γ of at most F-measure zero). Suppose $\{i, j\}$ are two exchanges such that $\{i, j\} \subset \mathcal{A}^*(\gamma)$, i.e., a type-$\gamma$ investor is indifferent between exchanges i and j. Then,

$$\gamma(\tilde{r}_i - \tilde{r}_j) = \frac{W^*}{\mu\beta_i} - \frac{W^*}{\mu\beta_j}. \tag{A.12}$$

The right hand side of (A.12) is independent of γ, and $\tilde{r}_i - \tilde{r}_j \neq 0$, by the assumption that the effective rebates are distinct. Then, $\{i, j\} \subset \mathcal{A}^*(\gamma)$ for at most a single value of γ. As there are only finitely many pairs of exchanges, we have that $|\mathcal{A}^*(\gamma)| = 1$ except for possibly finitely many $\gamma > \gamma_0$. Then, under Assumption 1(i), $\pi^{(1)}(\gamma)$ and $\pi^{(2)}(\gamma)$ differ on a set of γ of at most F-measure zero.

Combining these facts with the flow balance condition (11), we have that

$$\begin{aligned} Q_i^{(1)} &= Q_i^{(1)} \times \frac{\mu\beta_i}{\mu\beta_i} \times \frac{W^*}{\beta^\top Q^{(1)}} = \mu_i(Q^{(1)})\frac{W^*}{\mu\beta_i} \\ &= \left(\lambda_i + \Lambda \int_0^\infty \pi_i^{(1)}(\gamma)\, dF(\gamma)\right)\frac{W^*}{\mu\beta_i} \\ &= \left(\lambda_i + \Lambda \int_0^\infty \pi_i^{(2)}(\gamma)\, dF(\gamma)\right)\frac{W^*}{\mu\beta_i} \\ &= Q_i^{(2)}, \end{aligned}$$

for $i = 1, \ldots, N$, i.e., the equilibrium queue lengths are unique. ■

B. Proofs: Equilibrium Convergence

In this appendix, we prove the convergence of the queue length process $Q(t)$ to the unique equilibrium vector Q^* at $t \to \infty$, in the two-dimensional case.

As in Section 2.3, define $\chi_i(W(t))$ to be the instantaneous fraction of arriving limit orders that are placed into exchange i. The evolution of the queue length process $Q(t)$ is characterized by the following system of ordinary differential equations,

$$\dot{Q}_i(t) = \lambda_i + \Lambda\chi_i(W(t)) - \mu_i(Q(t)), \quad i = 1, \cdots, N. \tag{B.1}$$

In the remainder of this appendix, we focus on the two dimensional cases, i.e., $N = 2$. Also, without loss of generality, we assume $\lambda_i = 0$, for $i = 1, 2$.[12]

The fact that the equilibrium queue length vector exhibits state space collapse and leads us to consider a new coordinate system in which workload $W \triangleq \beta^\top Q$ is one of the new coordinates. In the two dimensional case, the workload W together with the sum of queue lengths $S \triangleq \mathbf{1}^\top Q$ characterize the individual queue lengths and vice versa. Thus, the convergence of $(W(t), S(t))$ to (W^*, S^*) where $W^* \triangleq \beta^\top Q^*$ and $S^* \triangleq \mathbf{1}^\top Q^*$, is equivalent to the convergence of the queue length process $Q(t)$ to the unique equilibrium vector Q^*. We perform the change of coordinates and rewrite the original ordinary differential equations in terms of W and S as follows:

$$\begin{cases} \dot{W}(t) = \Lambda\beta^\top\chi(W(t)) - \mu(\beta_1 + \beta_2)\cdot\mathbb{I}_{\{W(t)\neq 0\}} + \mu\dfrac{\beta_1\beta_2 S(t)}{W(t)}\cdot\mathbb{I}_{\{W(t)\neq 0\}}, \\ \dot{S}(t) = \Lambda\mathbf{1}^\top\chi(W(t)) - \mu\cdot\mathbb{I}_{\{S(t)\neq 0\}}. \end{cases} \tag{B.2}$$

We will restrict attention to this new (W, S) coordinate system for the remainder of this appendix.

[12] The proof that follows can be easily adapted to all other cases where $\lambda_1, \lambda_2 > 0$ and, $\lambda_1 + \lambda_2 < \mu$.

B.1. Overview of the Proof for $(W(t), S(t))$ Convergence

In the following we prove that under Assumptions 1–3, given arbitrary initial conditions $(W(0), S(0)) \in \mathbb{R}^2_+$, the process $(W(t), S(t))$ converges to the unique equilibrium (W^*, S^*) at $t \to \infty$.

Define the set $\mathcal{W}^+ \triangleq \{(W, S) : W = W^*, S > S^*\}$, i.e., the upper half of the vertical line $W = W^*$ in $\mathbb{R}^2$. We will show that $(W(t), S(t))$ either hits the set $\mathcal{W}^+$ or enters a local stability region within a finite time, starting from any initial point. This will imply that $(W(t), S(t))$ returns to set $\mathcal{W}^+$ with finite inter-arrival times, if it has not entered the local stability region. Each recurrence corresponds to a point on the upper half of the vertical line $W = W^*$ in $\mathbb{R}^2$, i.e., to a value of $S \geq S^*$. We then show that each recurrence has a smaller (closer to S^*) S value than the previous appearance in set $\mathcal{W}^+$. Moreover, the step size is bounded away from zero as long as the trajectory is outside the local stability region. This ensures there are finite iterations until $(W(t), S(t))$ enters the local stability region, and thus has to converge.

Accordingly, the proof will be organized around the following main steps, each of which corresponds to one of Lemmas 3-5 in the following subsection:

1. Lemma 3 (Local Stability). There exists $\varepsilon > 0$, such that if $(W(0), S(0))$ is in the set

$$\mathcal{W}_{local} \triangleq \{(W, S) : |W - W^*| < \varepsilon, |S - S^*| < \varepsilon\},$$

   then $(W(t), S(t))$ converges to (W^*, S^*).

2. Lemma 4 (Finite Inter-arrival Time). Starting from any initial point, a sample path either enters the local stability region $\mathcal{W}_{local}$ or hits the set $\mathcal{W}^+$ in finite time; in the latter case, starting from any point in $\mathcal{W}^+$ the sample path must, in finite time, either

   (i) reach the set $\mathcal{W}_{local}$,

   (ii) return to the set $\mathcal{W}^+$.

3. Lemma 5 (Guaranteed Decay). There exists $\varphi > 0$, such that if $\tau_1 < \tau_2$ are times where

$$(W(\tau_1), S(\tau_1)), (W(\tau_2), S(\tau_2)) \in \mathcal{W}^+ \text{ and } (W(t), S(t)) \notin \mathcal{W}_{local} \text{ for } t \in [\tau_1, \tau_2],$$

   then $S(\tau_2) \leq S(\tau_1) - \varphi$.

This method of proving $(W(t), S(t))$ convergence shows that each sample path is a decaying spiral in $\mathbb{R}^2$ centered around the unique equilibrium point (W^*, S^*). Analyzing the spiral, we show that each rotation takes finite time, and has a guaranteed decay towards the equilibrium along the S coordinate at times when the set $\mathcal{W}^+$ is hit.

Therefore, the spiral enters the local stability region after finite iterations and within finite time, at which point it much converge to the unique equilibrium.

B.2. Proving $(W(t), S(t))$ Convergence

We begin with a lemma that provides a series of bounds on the trajectory. First, we postulate that $(W(t), S(t))$ should be within the first quadrant $\mathbb{R}^2_+$, since both components are positive weighted sum of the queue lengths. Second, the ratio $S(t)/W(t)$ is bounded by the largest and smallest of $\{1/\beta_i\}_{i=1,2}$. Recall that we assume attraction coefficients are distinct. Without loss of generality, assume that $\beta_1 > \beta_2$ and define $\mathcal{C} \triangleq \{(W, S) : S/W \in [1/\beta_1, 1/\beta_2]\}$. The trajectory should be confined within this cone. Third, we provide a lower bound $\underline{W}$ and an upper bound $\overline{W}$ on the workload $W(t)$ and argue that after finite time the workload will be restricted within that range. As a result, after finite time an inequality with respect to the vector of routing fractions $\chi(W(t))$ holds, which will be useful in proving convergence later on.

Lemma 2 (Bounded Trajectory). *There exists $\zeta \in (0, \mu\beta_2)$ and $\underline{W}, \overline{W} \in [0, +\infty)$ with $\underline{W} < \overline{W}$, such that given initial conditions $S(0) = \mathbf{1}^\top Q(0)$ and $W(0) = \beta^\top Q(0)$ where $Q(0) \in \mathbb{R}^2_+$, there exists finite time $T_b \in [0, +\infty)$ such that at any time $t > T_b$,*

(1) the trajectory is contained within $\mathbb{R}^2_+ \bigcap \mathcal{C} \bigcap \mathcal{B}$ where $\mathcal{B} \triangleq \{(W, S) : W \in [\underline{W}, \overline{W}]\}$;

(2) $\Lambda\beta^\top\chi(W(t)) - \mu(\beta_1 + \beta_2) \leq -\zeta$.

Proof. Since $Q(t) \in \mathbb{R}^2_+$, it is obvious that $(S(t), W(t)) \in \mathbb{R}^2_+$. Moreover, for all $Q = Q(t) \in \mathbb{R}^2_+$,

$$\begin{aligned} \mathbf{1}^\top Q &\leq \frac{\beta_1}{\beta_2} Q_1 + Q_2 = \frac{1}{\beta_2}\left(\beta^\top Q\right), \\ \mathbf{1}^\top Q &\geq Q_1 + \frac{\beta_2}{\beta_1} Q_2 = \frac{1}{\beta_1}\left(\beta^\top Q\right). \end{aligned} \tag{B.3}$$

Therefore $1/\beta_1 \leq S(t)/W(t) \leq 1/\beta_2$.

For the third bound, we will use the following definitions of $\underline{W}$ and $\overline{W}$: Pick an arbitrary $0 < \zeta < \mu\beta_2$. Because of the monotonicity assumption, and that $\Lambda\beta^\top\chi(0) - \mu\beta_2 \geq \Lambda\beta_2 - \mu\beta_2 > 0$, $\Lambda\beta^\top\chi(W) - \mu\beta_2 \to -\mu\beta_2$ as $W \to \infty$, there will be a unique workload value satisfying $\Lambda\beta^\top\chi(W) - \mu\beta_2 = -\zeta$, which we denote as $\overline{W}$. Also because of the monotonicity assumption, and the fact that $\Lambda\beta^\top\chi(W) - \mu(\beta_1 + \beta_2) \to -\mu(\beta_1 + \beta_2)$ as $W \to \infty$, if $\Lambda\beta^\top\chi(0) - \mu(\beta_1 + \beta_2) \geq -\zeta$, there will be a unique workload value satisfying $\Lambda\beta^\top\chi(W) - \mu(\beta_1 + \beta_2) = -\zeta$, which we denote as $\underline{W}$. Otherwise, we define $\underline{W} = 0$. In both cases, $\Lambda\beta^\top\chi(\underline{W}) - \mu(\beta_1 + \beta_2) \leq -\zeta$.

For $W \geq \overline{W}$,

$$\begin{aligned} \dot{W} &= \Lambda\beta^\top\chi(W) - \mu(\beta_1 + \beta_2) + \mu\beta_1\beta_2\frac{S}{W} \\ &\leq -\zeta - \mu\beta_1 + \mu\beta_1\beta_2\frac{1}{\beta_2} = -\zeta. \end{aligned} \tag{B.4}$$

So if the trajectory starts with $W(0) > \overline{W}$, it decreases and goes under $\overline{W}$ within finite time $(W(0) - \overline{W})/\zeta$. And since at $W = \overline{W}$, $\dot{W} \leq -\zeta$, as soon as the trajectory goes below $\overline{W}$, it will stay below $\overline{W}$.

If $\underline{W} = 0$, then we always have $W \geq \underline{W}$. If $\underline{W} > 0$, for $W \leq \underline{W}$,

$$\begin{aligned} \dot{W} &= \Lambda\beta^\top \chi(W) - \mu(\beta_1 + \beta_2) + \mu\frac{\beta_1\beta_2 S}{W} \\ &\geq \zeta + \mu\beta_1\beta_2\frac{1}{\beta_1} = \zeta + \mu\beta_2. \end{aligned} \tag{B.5}$$

So if the trajectory starts with $W(0) < \underline{W}$, it increases and goes above $\underline{W}$ within finite time $(\underline{W} - W(0))/(\zeta + \mu\beta_2)$. And since at $W = \underline{W}$, $\dot{W} \geq \zeta + \mu\beta_2$, as soon as the trajectory goes above $\underline{W}$, it will stay above $\underline{W}$. Therefore, $(W(t), S(t)) \in \mathcal{B}$ after finite time $T_b = \max\{0, (W(0) - \overline{W})/\zeta, (\underline{W} - W(0))/(\zeta + \mu\beta_2)\}$.

When $(W(t), S(t)) \in \mathcal{B}$, $W(t) \geq \underline{W}$. Because of the monotonicity assumption,

$$\Lambda\beta^\top \chi(W(t)) - \mu(\beta_1 + \beta_2) \leq \Lambda\beta^\top \chi(\underline{W}) - \mu(\beta_1 + \beta_2) \leq -\zeta. \tag{B.6}$$

■

The bounded region of $\mathbb{R}^2_+ \bigcap \mathcal{C} \bigcap \mathcal{B}$ is divided into four quadrants according to the signs of $\dot{W}$ and $\dot{S}$ as follows:

First, the vertical line $W = W^*$ divides the space into two half-spaces in which S is monotonically changing. This is because

$$\begin{aligned} \mathbf{1}^\top \chi(W) &= \chi_1(W) + \chi_2(W) \\ &= \mathsf{P}\left(\max_{i=1,2}\ \gamma\tilde{r}_i - \frac{W}{\mu\beta_i} > \gamma\tilde{r}_0\right), \end{aligned} \tag{B.7}$$

is strictly decreasing in W, and, at the equilibrium,

$$\dot{S} = \dot{Q}_1 + \dot{Q}_2 = \Lambda\mathbf{1}^\top \chi(W^*) - \mu = 0. \tag{B.8}$$

Thus,

$$\begin{aligned} \dot{S} &> 0, \qquad \text{when } W < W^*, \\ \dot{S} &= 0, \qquad \text{when } W = W^*, \\ \dot{S} &< 0, \qquad \text{when } W > W^*. \end{aligned}$$

Second, denote by $\bar{S}(W)$ for which $\dot{W} = 0$ at a given workload W, in other words,

$$\bar{S}(W) \triangleq \frac{W}{\mu\beta_1\beta_2}\left(\mu(\beta_1 + \beta_2) - \Lambda\beta^\top \chi(W)\right). \tag{B.9}$$

Because of the monotonicity assumption, $\bar{S}(W) > 0$ for all $W > \underline{W}$. We can rewrite $\dot{W}$ in terms of

$\bar{S}(W)$ as

$$\begin{aligned}\dot{W} &= \Lambda\beta^\top\chi(W) - \mu(\beta_1+\beta_2) + \mu\frac{\beta_1\beta_2\bar{S}(W)}{W} + \mu\frac{\beta_1\beta_2(S-\bar{S}(W))}{W} \\ &= \mu\frac{\beta_1\beta_2(S-\bar{S}(W))}{W}.\end{aligned} \tag{B.10}$$

Thus,

$$\begin{aligned}\dot{W} &> 0, \qquad \text{when } S > \bar{S}(W), \\ \dot{W} &= 0, \qquad \text{when } S = \bar{S}(W), \\ \dot{W} &< 0, \qquad \text{when } S < \bar{S}(W).\end{aligned}$$

For later reference, we clockwise index the four quadrants by even numbers and the bordering regions in between by odd numbers, as listed below and illustrated in the following figure. These nine regions are mutually exclusive and collectively exhaustive:

- **Region 1:** $W = W^*, S > \bar{S}(W)$,
- **Region 2:** $W > W^*, S > \bar{S}(W)$,
- **Region 3:** $W > W^*, S = \bar{S}(W)$,
- **Region 4:** $W > W^*, S < \bar{S}(W)$,
- **Region 5:** $W = W^*, S < \bar{S}(W)$,
- **Region 6:** $W < W^*, S < \bar{S}(W)$,
- **Region 7:** $W < W^*, S = \bar{S}(W)$,
- **Region 8:** $W < W^*, S > \bar{S}(W)$,
- **Region 9:** $W = W^*, S = \bar{S}(W)$.

As indicated by the following lemma, the system of ordinary differential equations in (B.2) is locally asymptotically stable. So there exists a local stability region around the equilibrium such that all points inside the region converge.

Lemma 3 (Local Stability). *There exists $\varepsilon > 0$, such that if $(W(0), S(0))$ is in the set*

$$\mathcal{W}_{local} \triangleq \{(W,S) : |W - W^*| < \varepsilon, |S - S^*| < \varepsilon\},$$

then $(W(t), S(t))$ converges to (W^*, S^*).

Proof. For $W > 0, S > 0$, the Jacobian matrix corresponding to the system of ordinary differential equations in (B.2) is

$$J(W,S) = \begin{bmatrix} \Lambda\beta^\top\frac{\partial\chi(W)}{\partial W} - \mu\beta_1\beta_2\frac{S}{W^2} & \mu\beta_1\beta_2\frac{1}{W} \\ \Lambda\mathbf{1}^\top\frac{\partial\chi(W)}{\partial W} & 0 \end{bmatrix}$$

Denote λ_1, λ_2 as its two eigenvalues, then

$$\lambda_1 + \lambda_2 = \mathrm{tr}(J(W,S)) = \Lambda\beta^\top \frac{\partial \chi(W)}{\partial W} - \mu\beta_1\beta_2 \frac{S}{W^2} < 0. \tag{B.11}$$

$$\lambda_1 \cdot \lambda_2 = \det(J(W,S)) = -\Lambda\mu\beta_1\beta_2 \mathbf{1}^\top \frac{\partial \chi(W)}{\partial W} > 0. \tag{B.12}$$

So both of the eigenvalues have negative real parts and the system is locally asymptotically stable (Zak, 2003). ■

Now we are ready to set out the argument for convergence. As laid out in the overview of Section B.1, the next step is to show that the trajectory returns to the set $\mathcal{W}^+$ and thus to region 1 with finite inter-arrival time as long as it does not enter the local stability region.

Lemma 4 (Finite Interarrival Time). *There exists finite time $T_r \in (0, +\infty)$, such that for any $(W(0), S(0)) \in \mathbb{R}^2_+ \bigcap \mathcal{C} \bigcap \mathcal{B}$, there exists time $0 < t < T_r$ where $(W(t), S(t)) \in \mathcal{W}^+$ or $(W(t), S(t)) \in \mathcal{W}_{local}$.*

Proof. To prove that the trajectory starting from time t and with any initial point will reach region 1 after finite time, we will show that, starting from any point in any of the nine regions, unless the trajectory enters the local stability region, it will reach the next numbered region within finite time, and thus the trajectory has to return to region 1 within finite time. Therefore the trajectory will keep returning to region 1 with a finite interval of time (unless it enters the local stability region). In the following we discuss the cases region by region:

- **Region 1:** $W = W^*$, $S > \bar{S}(W)$, then $\dot{W} > 0, \dot{S} = 0$. So the trajectory instantly exits region 1 and reaches region 2.
- **Region 2:** $W > W^*$, $S > \bar{S}(W)$, then $\dot{W} > 0, \dot{S} < 0$. So the trajectory can only reach region 3 if the trajectory leaves region 2.

  Since $\bar{S}(W)$ is continuous and $\bar{S}(W^*) = S^*$, for $\varepsilon^- \in (0, \varepsilon)$, there exists a small $\delta_{\varepsilon^-} > 0$ such that for $W \in (W^*, W^* + \delta_{\varepsilon^-})$, $|\bar{S}(W) - S^*| < \varepsilon^-$. For $W \in (W^*, W^* + \min\{\delta_{\varepsilon^-}, \varepsilon\})$, if $S - S^* < \varepsilon$, then the trajectory converges because of local stability. Otherwise, without entering the local stability region, for these W values, $\dot{W} = \mu\beta_1\beta_2 \frac{S - \bar{S}(W)}{W} > \mu\beta_1\beta_2 \frac{\varepsilon - \varepsilon^-}{W^* + \varepsilon}$. So the trajectory will exceed $W^* + \min\{\delta_{\varepsilon^-}, \varepsilon\}$ within finite time.

  For $W > W^* + \min\{\delta_{\varepsilon^-}, \varepsilon\}$, denote

  $$S_\delta(S, W) \triangleq S - \bar{S}(W). \tag{B.13}$$

  Since $W > \underline{W}$,

  $$\bar{S}'(W) = \frac{1}{\mu\beta_1\beta_2} \Big(\mu(\beta_1 + \beta_2) - \Lambda\beta^\top \chi(W)\Big) - \frac{W}{\mu\beta_1\beta_2} \Lambda\beta^\top \frac{\partial \chi(W)}{\partial W} > 0. \tag{B.14}$$

Then,

$$\dot{S}_\delta = \dot{S} - \bar{S}'(W)\cdot\dot{W} < \Lambda\mathbf{1}^\top\chi(W) - \Lambda\mathbf{1}^\top\chi(W^*), \quad W > W^* + \min\{\delta_{\varepsilon^-}, \varepsilon\}. \tag{B.15}$$

We know $\mathbf{1}^\top\chi(W)$ is strictly decreasing and W is bounded away from W^*, therefore S_δ will decrease to 0, i.e., the trajectory will reach region 3, within finite time.

- **Region 3:** $W > W^*$, $S = \bar{S}(W)$, then $\dot{W} = 0, \dot{S} < 0$. So the trajectory instantly exits region 3 and reaches region 4.

- **Region 4:** $W > W^*$, $S < \bar{S}(W)$, then $\dot{W} < 0, \dot{S} < 0$. So the trajectory can can only reach regions 3, 5, or 9 if the it leaves region 4.

  For $W \in (W^*, W^* + \min\{\delta_{\varepsilon^-}, \varepsilon\})$, if $-\varepsilon < S - S^* < \varepsilon^-$, then the trajectory converges because of local stability. Otherwise for these W values, $\dot{W} = \mu\beta_1\beta_2\frac{S-\bar{S}(W)}{W} < -\mu\beta_1\beta_2\frac{\varepsilon-\varepsilon^-}{W^*+\varepsilon}$. The trajectory will go to $W = W^*$ and reach region 5 within finite time.

  For $W > W^* + \min\{\delta_{\varepsilon^-}, \varepsilon\}$,

$$\dot{S} = \Lambda\mathbf{1}^\top\chi(W) - \Lambda\mathbf{1}^\top\chi(W^*), \quad W > W^* + \min\{\delta_{\varepsilon^-}, \varepsilon\}. \tag{B.16}$$

  Since $\mathbf{1}^\top\chi(W)$ is strictly decreasing, W is bounded away from W^*, and S is bounded below by the line $S = \frac{1}{\beta_1}W^*$, the trajectory will leave this region within finite time.

- **Region 5:** $W = W^*$, $S < \bar{S}(W)$, then $\dot{W} < 0, \dot{S} = 0$. So the trajectory instantly exists region 5 and reaches region 6.

- **Region 6:** $W < W^*$, $S < \bar{S}(W)$, then $\dot{W} < 0, \dot{S} > 0$. Thus, the trajectory can only reach region 7 if it leaves region 6.

  Since $\bar{S}(W)$ is continuous and $\bar{S}(W^*) = S^*$, for $\varepsilon^- \in (0, \varepsilon)$, there exists a small $\delta'_{\varepsilon^-} > 0$ such that for $W \in (W^* - \delta'_{\varepsilon^-}, W^*)$, $|\bar{S}(W) - S^*| < \varepsilon^-$. For $W \in (W^* - \min\{\delta'_{\varepsilon^-}, \varepsilon\}, W^*)$, if $0 > S - S^* > -\varepsilon$, then the trajectory converges because of local stability. Otherwise, without entering the local stability region, for these W values, $\dot{W} = \mu\beta_1\beta_2\frac{S-\bar{S}(W)}{W} < -\mu\beta_1\beta_2\frac{\varepsilon-\varepsilon^-}{W^*}$. So the trajectory will go below $W^* - \min\{\delta'_{\varepsilon^-}, \varepsilon\}$ within finite time.

  For $W < W^* - \min\{\delta'_{\varepsilon^-}, \varepsilon\}$,

$$\dot{S}_\delta = \dot{S} - \bar{S}'(W)\cdot\dot{W} > \Lambda\mathbf{1}^\top\chi(W) - \Lambda\mathbf{1}^\top\chi(W^*), \quad W < W^* - \min\{\delta'_{\varepsilon^-}, \varepsilon\}. \tag{B.17}$$

  Since $\mathbf{1}^\top\chi(W)$ is strictly decreasing and W is bounded away from W^*, S_δ will increase to 0, i.e., the trajectory will reach region 7, within finite time.

- **Region 7:** $W < W^*$, $S = \bar{S}(W)$, then $\dot{W} = 0, \dot{S} > 0$. So the trajectory instantly exists region 7 and reaches region 8.

- **Region 8:** $W < W^*$, $S > \bar{S}(W)$, then $\dot{W} > 0, \dot{S} > 0$. The trajectory can only reach regions 1, 7, or 9 if it leaves region 8.

  For $W \in (W^* - \min\{\delta'_{\varepsilon^-}, \varepsilon\})$, if $-\varepsilon^- < S - S^* < \varepsilon$, then the trajectory converges because of local stability. Otherwise for these W values, $\dot{W} = \mu\beta_1\beta_2 \frac{S-\bar{S}(W)}{W} > \mu\beta_1\beta_2 \frac{\varepsilon}{W^*}$. The trajectory will exceed W^* within finite time.

  For $W < W^* - \min\{\delta'_{\varepsilon^-}, \varepsilon\}$,

  $$\dot{S} = \Lambda \mathbf{1}^\top \chi(W) - \Lambda \mathbf{1}^\top \chi(W^*), \quad W < W^* - \min\{\delta'_{\varepsilon^-}, \varepsilon\}. \tag{B.18}$$

  Since $\mathbf{1}^\top \chi(W)$ is strictly decreasing, W is bounded away from W^*, and S is bounded above by the line $S = \frac{1}{\beta_2} W^*$, the trajectory will leave this region within finite time.

- **Region 9:** If $(W(t), S(t))$ is in region 9, then the trajectory has already converged.

■

The final step is to show that between two successive times when the trajectory returns to region 1, the S coordinate gets closer to the equilibrium value S^*. Furthermore, the step size is bounded away from zero as long as the trajectory does not enter the local stability region.

Lemma 5 (Guaranteed Decay). *There exists $\varphi > 0$, such that for any $(W(0), S(0)) \in \mathcal{W}^+ \bigcap \mathbb{R}^2_+ \bigcap \mathcal{C} \bigcap \mathcal{B}$, If $t_1 > 0$ is a time with $(W(t_1), S(t_1)) \in \mathcal{W}^+ / \mathcal{W}_{local}$, then $S(0) - S(t_1) > \varphi$.*

Proof. If $(W(t_1), S(t_1)) \in \mathcal{W}^+ / \mathcal{W}_{local}$, the trajectory has cycled back to region 1 without entering the local stability region. Along its path, trajectory will first reaches the lower half of the vertical line $W = W^*$, i.e., region 5, and then return to region 1. We denote the time that the trajectory hits region 5 as $t_5 \triangleq \inf\{s > t : W = W^*, S < \bar{S}(W)\}$.

The idea of the proof is to first show that the trajectory gets closer to the equilibrium when it reaches region 5, i.e., $(S(0) - S^*) - (S^* - S(t_5)) > \varphi_r$ for some $\varphi_r > 0$; and then make an analogous claim about the other half of the journey; and thus prove that the trajectory, when keeping returning to region 1, always moves closer to the equilibrium with a positive step size.

Denote $t_3 \triangleq \inf\{s > t : S = \bar{S}(W)\}$, i.e., the time that the trajectory reaches region 3. For any $W \in [0, W(t_3)]$, since W is first strictly increasing in region 2 and then strictly decreasing in region 4, it should be passed by the trajectory twice, once in region 2 and once in region 4. We denote

$$t_4(W(\tau)) \triangleq \inf\{s > t_3 : W(s) = W(t)\}, \quad \tau \in [0, t_3]. \tag{B.19}$$

Since $W(\tau) = W(t_4(W(\tau)))$,

$$\dot{W}(\tau) = \dot{W}(t_4(W(\tau))) \cdot t'_4(W(\tau)) \cdot \dot{W}(\tau), \quad \tau \in [0, t_3]. \tag{B.20}$$

$$t_4'(W(\tau)) = \frac{1}{\dot{W}_{t_4(W(\tau))}}, \quad \tau \in [0, t_3). \tag{B.21}$$

We define the following function,

$$F(\tau) \triangleq S(\tau) + S(t_4(W(\tau))) - 2\bar{S}(W)(\tau), \quad \tau \in [0, t_3]. \tag{B.22}$$

We are about to show for any $\tau \in (0, t_3)$, there exists a time $\nu \in [\tau, t_3)$ such that $F(\nu) > 0$, i.e., there exists arbitrarily close point to $(W(t_3), S(t_3))$ such that the trajectory is closer to line $S = \bar{S}(W)$ in region 4 than in region 2. By contradiction, for any $\tau \in (0, t_3)$, if $F(\nu) \leq 0, \forall \nu \in [\tau, t_3)$, then,

$$\begin{aligned}
\dot{F}(\nu) &= \dot{S}(\nu) + \dot{S}(t_4(W_\nu)) \cdot t_4'(W(\nu)) \cdot \dot{W}(\nu) - 2\bar{S}'(W(\nu)) \cdot \dot{W}(\nu) \\
&= \dot{S}(W(\nu)) + \dot{S}(W(\nu)) \cdot \frac{1}{\dot{W}(t_4(W_\nu))} \cdot \dot{W}(\nu) - 2\bar{S}'(W(\nu)) \cdot \dot{W}(\nu) \\
&= \dot{S}(W_\tau) \cdot \dot{W}_\tau \cdot \left(\frac{W_\tau}{\mu\beta_1\beta_2(S_\tau - \bar{S}(W_\tau))} + \frac{W_\tau}{\mu\beta_1\beta_2(S_{t_4(W_\tau)} - \bar{S}(W_\tau))} \right) - \bar{S}'(W_\tau) \cdot \dot{W}_\tau \\
&= \frac{\dot{S}(W(\nu)) \cdot \dot{W}(\nu) \cdot W(\nu)}{\mu\beta_1\beta_2} \cdot \frac{F(\nu)}{(S(\nu) - \bar{S}(W(\nu))) \cdot (S(t_4(W(\nu))) - \bar{S}(W(\nu)))} \\
&\quad - \bar{S}'(W(\nu)) \cdot \dot{W}(\nu) < 0.
\end{aligned} \tag{B.23}$$

Then,

$$F(t_3) - F(\tau) = \int_\tau^{t_3} \dot{F}(\nu) d\nu < 0, \tag{B.24}$$

which contradicts with the fact that $F(t_3) = 0$ and $F(\tau) \leq 0$.

Now we are about to show $(S(0) - S^*) - (S^* - S(t_5)) > 0$, i.e., $F(0) > 0$. By contradiction, if $F(0) \leq 0$, and choose τ as a point that is close to $(W(t_3), S(t_3))$ with $F(\tau) > 0$. By continuity of $F(\cdot)$, it has to be zero at some points between region 3 and region 5. Denote $t_{equal} \triangleq \sup\{0 < s < \tau : W(s) = W(t_4(W(s))\}$ as the closest to $(W(\tau), S(\tau))$ among such points. At t_{equal}, $F(t_{equal}) = 0$ and

$$\dot{F}(t_{equal}) = -\bar{S}'(W(t_{equal})) \cdot \dot{W}(t_{equal}) < 0 \tag{B.25}$$

so there has to be some time between t_{equal} and τ such that the two distances equate, which contradicts with the fact that t_{equal} is the closest to $(W(\tau), S(\tau))$ among such points. In fact, with such argument we can make a stronger claim: $F(\tau) > 0$ for all $\tau \in [0, t_3)$.

We still need to show that not only $F(0) > 0$, but also there exists a $\varphi_r > 0$ such that $F(0) > \varphi_r$ as long as the trajectory does not enter the local stability region, i.e., either $|W(t) - W^*| > \varepsilon$ or $|S(t) - \bar{S}(W(t))| > \varepsilon$ for any $(W(t), S(t))$.

Recall from equation (B.23) that

$$\dot{F}(\tau) = \frac{\dot{S}(\tau)\cdot\dot{W}(\tau)\cdot W(\tau)}{\mu\beta_1\beta_2}\cdot\frac{F(\tau)}{(S(\tau)-\bar{S}(W(\tau)))\cdot(S(t_4(W(\tau)))-\bar{S}(W(\tau)))} - \bar{S}'(W(\tau))\cdot\dot{W}(\tau). \tag{B.26}$$

Define

$$G(\tau) \triangleq \frac{\dot{S}(\tau)\cdot\dot{W}(\tau)\cdot W(\tau)}{\mu\beta_1\beta_2(S(\tau)-\bar{S}(W(\tau)))\cdot(S(t_4(W(\tau)))-\bar{S}(W(\tau)))}, \quad \tau\in[0,t_3]. \tag{B.27}$$

Then,

$$\dot{F}(\tau) = G(\tau)\cdot F(\tau) - \bar{S}'(W(\tau))\cdot\dot{W}(\tau). \tag{B.28}$$

Note that $G(0)=0$, and because of continuity of $G(\cdot)$, for a small ε_G such that

$$0<\varepsilon_G<\frac{\beta_2\zeta(\varepsilon-\varepsilon^-)}{W^*(W^*+\varepsilon)}-\frac{\beta_2\Delta}{W^*} \tag{B.29}$$

where $0<\Delta<\frac{\zeta(\varepsilon-\varepsilon^-)}{W^*+\varepsilon}$, there exists $t_G>0$ such that for $t\in[0,t_G)$, $|G(t)|<\varepsilon_G$.

Recall that for $W\in[W^*,W^*+\min\{\delta_\varepsilon,\varepsilon\})$, $\dot{W}>\frac{\mu\beta_1\beta_2(\varepsilon-\varepsilon^-)}{W^*+\varepsilon}$. At the same time, the starting position in region 1 satisfies $S(0)\le W^*/\beta_2$. $S(\tau)\le S(0)\le W^*/\beta_2$ for all $\tau\in[0,t_1]$, because the trajectory first decreases until it reaches region 5 and then increases to return to region 1 at a lower level. Therefore, we also have $\dot{W}<\frac{\mu\beta_1\beta_2(W^*/\beta_2+\varepsilon^-)}{W^*}$. Then for time $\tau\in[0,\frac{(W^*+\min\{\delta,\varepsilon\})W^*}{\mu\beta_1\beta_2(W^*/\beta_2+\varepsilon^-)})$, $W\in[W^*,W^*+\min\{\delta,\varepsilon\})$.

$F(\tau)<S(\tau)\le W^*/\beta_2$ is bounded. So is $\bar{S}'(W(\tau))>\frac{1}{\mu\beta_1\beta_2}\left(\mu(\beta_1+\beta_2)-\Lambda\beta^\top\chi(W)\right)\ge\zeta/\mu\beta_1\beta_2$.

For $\tau\in[0,\min\{t_G,\frac{(W^*+\min\{\delta,\varepsilon\})W^*}{\mu\beta_1\beta_2(W^*/\beta_2+\varepsilon^-)}\})$,

$$\begin{aligned}\dot{F}(\tau) &< \varepsilon_G F(\tau) - \frac{\zeta*}{\mu\beta_1\beta_2}\frac{\mu\beta_1\beta_2(\varepsilon-\varepsilon^-)}{W^*+\varepsilon}\\ &< \left(\frac{\beta_2\zeta(\varepsilon-\varepsilon^-)}{W^*(W^*+\varepsilon)}-\frac{\beta_2\Delta}{W^*}\right)\cdot\frac{W^*}{\beta_2}-\frac{\zeta(\varepsilon-\varepsilon^-)}{W^*+\varepsilon}\\ &= -\Delta.\end{aligned} \tag{B.30}$$

Therefore,

$$F(0)>F(0)-F\left(\min\left\{t_G,\frac{(W^*+\min\{\delta,\varepsilon\})W^*}{\mu\beta_1\beta_2\varepsilon}\right\}\right)>\Delta\cdot\min\left\{t_G,\frac{(W^*+\min\{\delta,\varepsilon\})W^*}{\mu\beta_1\beta_2\varepsilon}\right\}. \tag{B.31}$$

We can make analogous claims on $(S^*-S(t_5))-(S(t_1)-S^*)$, i.e., on the other half of the trajectory from region 5 back to region 1, and thus prove that in each cycle the trajectory gets

closer to the equilibrium with a positive step size as long as it does not enter the local stability region and therefore has to converge. ■