Columbia Business School

Decision, Risk & Operations
Working Papers Series

Optimal Order Routing in a Fragmented Market

C. Maglaras, C.C. Moallemi, and H. Zheng

May 2012
DRO-2012-02

Optimal Order Routing in a Fragmented Market

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: May 21, 2012

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 the "price-time" priority rule. Taking into account the effect of investors' order routing decisions across exchanges, we find that the equilibrium of this decentralized market exhibits a certain state space collapse property, whereby (a) the states at different exchanges are coupled in a fairly intuitive manner, (b) the behavior of the market is captured through a one-dimensional process that can be viewed as a weighted aggregate depth of the market at the best bid and offer across all exchanges, and (c) the behavior of the various exchanges is inferred through a set of simple mappings from that of the aggregated market depth process. 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 strong support for the predicted state space collapse property.

1. Introduction

Modern equity markets are highly fragmented. In the United States alone there are over a dozen exchanges and about forty alternative trading systems (ATS) where investors may choose to trade. Market participants, including institutional investors, market makers, and opportunistic investors, interact within today's high-frequency, fragmented marketplace with the use of electronic algorithms. 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. These algorithms seek to optimize their own best execution objectives while taking into account short term differences or opportunities across the various exchanges. Characterizing the interaction effects between market fragmentation and high-frequency, self-interested order routing is an issue of central importance in understanding trade execution, market design, and policy questions that have come to the fore in recent years. This paper studies this problem, and, specifically, highlights an important structural property that emerges through these otherwise complex interaction effects.

Exchanges typically function as electronic limit order books, operating under a "price-time" priority rule. Here, resting limit orders are prioritized for trade first based on their associated prices, and then, at a given price level, according to their time of arrival. In this way, the high-frequency dynamics of an individual exchange can be understood as that of a multi-class queueing system. Each queue consists of a collection of limit orders posted at a single price level. Job arrivals correspond to new limit orders that are posted. Market orders trigger executions which, in queueing systems parlance, correspond to service completions of a single server. The service discipline considers queues first by price, and then each queue is operated in a first-in-first-out (FIFO) fashion. The entire market, consisting of multiple exchanges, can then be viewed as a stochastic network, which evolves as a collection of parallel single-server, multi-class queueing systems (cf. Figure 1).

Exchanges publish real-time information for each security that allow investors to know or compute the quantities available for trade at each price level (i.e., the queue lengths). At any point in time, the conditions at the exchanges may differ with respect to the best bid and offer[1] price levels, the market depth at various prices, recent trade activity, etc. These, in turn, imply differences in a number of metrics that capture the quality of execution over time and across various exchanges, such as the probability that the order will be filled, the expected delay until such a fill, or the adverse selection[2] 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 make-take fees range in magnitude, and are typically between $-\$0.0010$ and $\$0.0030$ per share traded. 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.

Market participants employ so-called "smart order routers" that take into account real-time market data 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 tradeoff metrics such as price, rebates, delays, and adverse selection effects. Because of this, when viewed as a queueing network, the fragmented marketplace is quite different than the so-called parallel server systems that have been studied extensively in the

[1]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, and, of course, 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).

[2]Roughly speaking, adverse selection measures the anticipated adverse price movement over a short time interval, conditional on the execution of a limit order. For example, conditional on executing a limit order to buy at the bid, the anticipated drift of the price in the near term is typically negative, indicating that one could have possibly purchase stock at a lower price at a later time. Given that exchanges differ in the fees and rebates, market orders will preference lower fee exchanges over higher fee ones, and as a result, expected adverse selection will also differ across these exchanges.

literature for three reasons: (a) the parallel sub-systems (the exchanges in our case) are not centrally controlled, (b) arrivals and service completions into sub-systems (the order flow) are optimized by self-interested agents, and (c) the agents are heterogeneous with respect to their preferences.

This paper formulates a mathematical model of the fragmented marketplace that captures many of the aforementioned phenomena. Specifically, we focus on the bid side of the book and study how the optimized limit order placement and market order routing decisions made by self-interested and heterogeneous investors affect the behavior at the various exchanges, characterize equilibrium market outcomes, and order flow equilibria. Investors that arrive and consider placing a limit order to buy stock have a choice between the various exchanges and face a tradeoff that balances monetary considerations (e.g., rebates, adverse selection) with temporal considerations (e.g., expected delay). We allow investors to be heterogeneous with respect to the way they tradeoff between these two. We capture the routing decisions of contra-side market sell orders via a reduced form model that we motivate and estimate empirically. The decision of where to route a marketable order, i.e., one that will execute immediately, is somewhat simpler as it does not involve tradeoffs over time. The reduced form model posits a notion of "attraction" of each exchange at a given point in time as function of its depth and economics, and computes routing decisions given the relative attraction of the various exchanges.

At a high level, this paper makes three key contributions:

- *A novel model for order routing in fragmented markets.* First, from a modeling perspective, our work is among the first papers to capture the order routing decision in today's fragmented market structure in a detailed manner. Specifically, we incorporate the idiosyncratic economic and temporal considerations of investors so as to accurately model the microstructure of order routing decisions in a state dependent manner. This appears novel viz. the existing literature in market microstructure and that in limit order book dynamics. Simultaneously, it offers concrete motivation for a decentralized version of the parallel server system that exhibits novel features and is of independent interest from a stochastic modeling viewpoint.

- *A theoretical analysis establishing state space collapse.* The optimization of order routing decisions couples the dynamics of the different exchanges. Our second, methodological, contribution is to 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 process, known as the *workload*. This phenomenon is known as *state space collapse* (SSC). The workload process can be interpreted as a single 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.

- *Empirical verification of state space collapse.* Third, our model and its main theoretical results lead to several testable hypotheses, most notably regarding the effective dimensionality of

the market dynamics, the relation between the expected delays across exchanges, and their ordering according to the exchange economics. We report on an econometric study of 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.

We test the implications of our model in several ways. First, we perform a principal component analysis 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 principle component explains 83% of the variability of the expected delays across exchanges, and that the first two principle components explain 90%. The second testable hypothesis derived from the model is that expected delays across different exchanges should be linearly related, with a specific, predicted constant of proportionality. We test this hypothesis in two different ways. First, for each security, we regress the expected delay of each exchange against the expected delay of a reference exchange. We find that the average R^2 across all of these models is 88%. Moreover, the regression coefficients are generally consistent with the constant of proportionality predicted by our model. To further study this relationship, we analyze the residuals between the expected delay estimates based on our model prediction using the one-dimensional workload process and the expected delay estimates computed using all available information, i.e., the depth and trading volume statistics at each exchange. We find that the model predictions explain 87% the variability in the realized expected delays across all of the exchanges. Overall, the empirical analysis provides strong statistical support for the theoretical SSC prediction of our model. Separately, the reduced form model that we postulated and used to explain the routing decisions of contra-side market orders is both statistically significant and insightful in the sense that the routing weights are generally consistent with the respective take fees. The findings are robust across securities and insensitive to the stochastic and non-stationary nature of the markets over long periods of time that is assumed away in the theoretical analysis.

State space collapse implies that, in equilibrium, all delays across exchanges are proportional to the aggregate depth or workload process. For example, if one exchange has long queues and is experiencing long delays, then the other exchanges will also be experiencing proportionally long delays. Put differently, if 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: people join the shortest queue and as a result quickly erase any queue length differences across booths or checkout counters. In our model the choice behavior is more intricate, and the coupling depends on both the economics and anticipated delays of each exchange.

The SSC property is a direct consequence of order routing optimization. The specific structure

of the one dimensional manifold depends on the form of the order routing problem we use, but the SSC property itself seems general and remains valid if the modeling assumptions are changed or relaxed. Specifically, SSC is not the result of the price protection mechanism[3] imposed in U.S. equities market, but rather that of routing decisions among exchanges at the same price level that differ, however, with respect to their economics and delays.

State space collapse results have played an important role in performance analysis and control of large scale stochastic networks, i.e., networks that are characterized by large volumes of inflow and large processing capacities. These results are typically derived in settings where processing resources are operating at near full utilization levels, and are based on asymptotic analysis that gives rise to appropriate fluid and diffusion models that exhibit SSC. It is also typical to assume that a centralized planner optimally controls the processing resources and capabilities of the system, or that a system is operating under a predefined control policy. In common to that body of work, our model is also characterized by large volumes of transactions, processing resources (the exchanges) that operate at full utilization (service completions, i.e., trades, are exactly matched with order arrivals), and our analysis leverages the tools just mentioned. The decentralized nature of the order routing decisions, both for order arrivals (limit orders) and service completions (market orders) is novel and of independent interest.

Our empirical analysis points that the fragmented market is in fact strongly coupled and acts in a way that is consistent with a lower dimensional system. This coupling is evident when one observes the exchange behavior over moderate time horizons, of the order of minutes or longer, that suppresses the short term differences that result from arrivals or departures of individual orders. The lower dimensional system is tractable and seems to offer a valuable framework for downstream analysis of many interesting questions that pertain to exchange competition (e.g., how to set make-take fees and rebates 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).

To our knowledge, this paper seems to be the first within the stochastic networks literature to offer empirical verification of SSC in a real and complex stochastic processing system during the past 25 years or so that the of SSC has been explored. Most of the literature, which we briefly review in §1.1, is prescriptive in that it formulates and studies models that are meant to offer insights for how systems should operate. In a similar vein, our theoretical analysis suggested that SSC should emerge, and despite the fact that many of the assumptions of our model may be not satisfied in practice, our empirical analysis found strong evidence that the complex and multi-dimensional U.S. equity market satisfies such a SSC property.

The remainder of this paper is organized as follows. This section concludes with a very brief literature survey. §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.

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

Our main results on market equilibrium and state space collapse are given in §3. In §4 we show empirical evidence of state space collapse.

1.1. Literature Review

Apart from some of 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 in order routing decisions. 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), and Cont and Larrard (2010). 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.

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 fees on market outcomes and liquidity cycles. Foucault et al. (2005) describe a theoretical model to understand make-take fees when monitoring the market is costly. Malinova and Park (2010) empirically study the introduction of make-take 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. One such paper is that of Foucault and Menkveld (2008); they study 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 US 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 make-take 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.

Finally, there is a broad literature on the study of mathematical models of stochastic networks, and multi-class queuing networks, in particular, that are useful in analyzing the high frequency behavior of limit order books. The latter connection has been explored in Cont et al. (2010) and Maglaras and Moallemi (2011). We offer a brief list of references that introduce and explore some of the related ideas we use in our work. To start with, the so called equivalent workload formulations and the associated idea of state space collapse arise in stochastic network control problems that are considered in the context of their approximate Brownian model formulations. This idea and its consequences in policy design has been pioneered by the work of Harrison (1988); Harrison and Van Mieghem (1996); Harrison (2000). Workload fluid models were first introduced in Harrison (1995), while the use of workload relaxations of fluid model control problems involving sequencing, routing and admission control were proposed in Meyn (2001). The justification of the fluid model that we suggest relies on the work by Mandelbaum and Pats (1995) on queues with state dependent parameters. The condition that guarantees that parallel server networks reduce under SSC to one-dimensional systems was first introduced in 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); the latter involves pricing and routing decisions that tradeoff delay against cost, similar to the types of decisions encountered in exchanges. Plambeck and Ward (2006) studies an assemble-to-order system, that involves a two-sided market fed by product requests on one side and raw materials necessary for product fulfillment on the other. In contrast to the above body of work, the model we study involves optimized, self-interested routing of the matching flow that in the stochastic network parlance represents the service completions.

2. Model

We develop a stylized model of a fragmented market consisting of N distinct electronic limit order books simultaneously trading a single underlying asset. Our focus here is to understand the interaction between multiple limit order books. Since this interaction is governed by investors who face a choice of where to route their orders, it is most natural to consider the dynamics of the market over the timescale that is relevant in order routing decisions, which is that of queueing delays, i.e., the time required for a limit order to execute, across the various exchanges. To focus on this, we make a number of simplifying assumptions that aid the tractability of our model.

One-sided market. We model only one side of the market in isolation. Without loss of generality, we choose this to be the bid side, i.e., we analyze the resting limit orders to buy the stock.

Top-of-book only. Limit orders on the bid side are further distinguished by an associated limit price. At any instant in time, we only consider limit orders at the national best bid price, the highest

bid price available across all exchanges, i.e., the "top-of-book". This is because a profit-maximizing seller who wishes to trade would not choose to trade at a lower price without exhausting all higher priced bids. Moreover, in the United States, price priority across market venues is enforced *de jure* by SEC Regulation NMS, a price protection mechanism that prevents trading at a price worse than is available at another exchange.

Fluid model. In an electronic limit order book, orders arrive at distinct and seemingly stochastic instances of time, and each order is associated with a discrete order size. We consider a deterministic fluid model, or "mean field" model, where discrete and stochastic 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 where the rates of events grows large using the functional strong law of large numbers. As an approximation, it is appropriate in settings where the rate of order arrivals grows large and the size of each individual order is small relative to the overall order flow over any interval of time, and is well suited for characterizing transient dynamics in such systems. The time scale of the latter is similar to that of the queueing delays, i.e., the time it takes for queue length to drain or move from one configuration to another, which is the relevant time scale in order routing decisions that tradeoff trading delays against rebates.

To put this in perspective, the dynamics of the underlying events on each exchange (i.e., arrival of limit or market orders) takes place on a time-scale measured in milliseconds to seconds for liquid securities. On the other hand, the time scale of queueing delays is of the order of seconds to minutes.

Constant arrival rates. Over the course of a trading day, the order arrival rates follow some relatively predictable profile that exhibits strong time-of-day effects, and further exhibit significant and unpredictable fluctuations around these forecasts. These underlying profiles exhibit significant variation over longer time scales (e.g., minutes to hours) than what we consider here, and for that reason we assume that arrival rates are constant for purposes of our analysis. We will also assume that the event rates do not depend on the state, i.e., the queue lengths and prices, at the various exchanges, which is also an idealization.

In what follows, we will describe a system model based on the above assumptions, which we will use to derive structural results about the equilibrium behavior across many exchanges. The empirical analysis of §4 ,which is based on a large sample of real market data, will nevertheless test our findings in a setting that is discrete and stochastic, and where system parameters such as arrival rates are non-stationary.

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. We denote by $Q(t) \triangleq (Q_1(t), Q_2(t), \ldots, Q_N(t)) \in \mathbb{R}_+^N$ the vector of queue lengths (or, quantities of resting limit orders) at each exchange at time t.

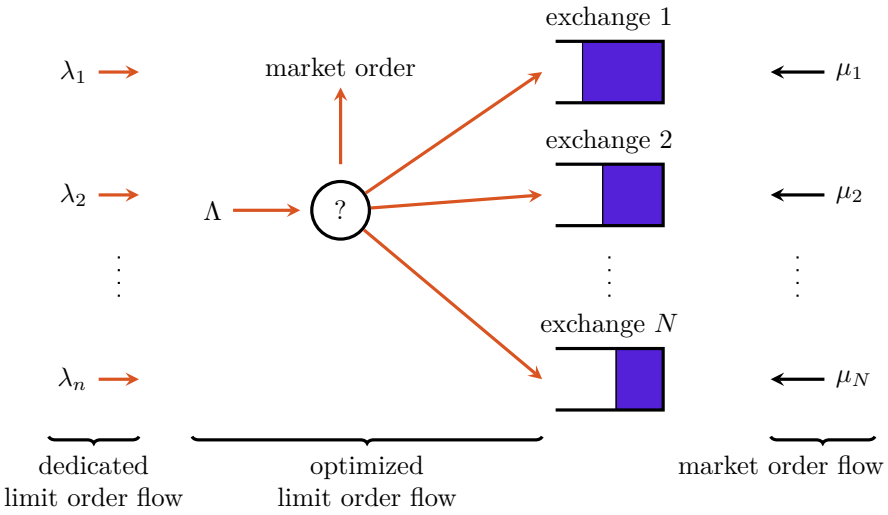


Figure 1: A one-sided, top-of-book model of multiple limit order books. Here, the quantity of liquidity (i.e., resting limits orders) at the national best bid price is depicted at each exchange. Additional liquidity accumulates through the arrival of optimized limit order flow, which seems to optimally route limit orders across exchanges, or though specific flows that are dedicated to each exchange. Liquidity is removed through the arrival of market order flow.

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. This flow arrives at rate $\lambda_i \geq 0$ and is destined for a particular exchange i, independent of the conditions at that exchange or at other exchanges. This flow could represent, for example, investors that may not have the ability to route orders to all exchanges, may not have the ability to make real-time order routing decisions, or even have special economic incentives for liquidity provision at particular exchanges.

Optimized limit order flow. A second type of limit orders arrive at a rate $\Lambda > 0$ and represents investors that attempt to make an "optimal" routing decision. We assume that each infinitesimal optimized limit order investor arrives to the market, observes the current state of the market, as summarized by the queue length vector $Q(t)$, and tactically selects to which venue $i \in \{1, \ldots, N\}$ to route the order, based on conditions at the time of arrival. Such an investor can also decide, if conditions are unfavorable, not to leave a limit order and to trade instead with a market order at the offered side of the market. We denote this possibility by the index $i = 0$.

In our model, once a limit order arrives at a particular exchange, it remains queued at that exchange until the order 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 that the instantaneous risk of adverse selection is high; when prices move away from the posted limit order; etc. This assumption simplifies the order routing decision and leads to a more tractable

analysis whose findings are supported by the empirical study of §4 that incorporates the effect of cancellations that are present in market data.

Investors may have different concerns or objectives when submitting a limit order that will influence their routing decisions. Some key factors affecting limit order placement are the following:

Expected delay. All things being equal, an investor would prefer to trade earlier rather than later. Trading later increases the risk of an adverse price movement, for example. Further, in many instances (e.g., algorithmic trading applications), the limit order in question may be a "child order" or part of the execution trajectory of a large "parent order". The investor may have a limited time horizon for the execution of the entire parent order as well as constraints on its execution trajectory defined by its "strategy," and delays in the execution of the child order adversely impact that goal. Hence, given a choice of exchanges that differ in their expected delays until a newly placed limit order would execute, the investor would prefer a route with a shorter expected delay. As will be seen in §4, the expected delay of a typical limit order execution is in the range of 1 to 1000 seconds.

If exchange i has a current queue length $Q_i(t)$ and market orders arrive removing liquidity from the exchange at rate $\mu_i > 0$, then the expected delay in our fluid model is defined

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

For purposes of analysis, the rates μ_i are assumed to be known to the investors. In practice, the μ_i's can be approximated by observing recent real-time trading activity at each exchange. In the case where the investor decides to take liquidity ($i = 0$), the resulting market order is immediately executed, however, so we set $\mathsf{ED}_0 \triangleq 0$.

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, for example, to the bid-ask spread of a typical liquid stock of $\$0.01$ per share, and represent an important part of the overall trading cost, and, as such, influence the order routing decisions. All things being equal, investors providing liquidity prefer higher rebates to lower 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. We capture this by setting $r_0 < 0$ corresponding to the sum of these payments.

Adverse selection. Suppose that an investor with no information on future price movements (i.e., a zero expectation) places a limit order to buy. Adverse or negative selection refers to the fact that, in general, conditional on the limit order getting executed, the expectation of the future price change of the stock is negative. This is because the resting limit order is exposed to the arrival of informed investors with private information on the future movement of the stock price. When placing a market order, on the other hand, the investor would suffer no adverse selection since

execution is certain. Adverse selection can thus be viewed as a cost associated with limit orders. The cost of adverse selection is of a similar order of magnitude at that of the liquidity rebates.

In general, the adverse selection cost associated with an exchange depends on the exchange as well as the current market conditions. For example, if the current queue length in exchange i is given by $Q_i(t)$, one would expect that the adverse selection cost associated with placing an incremental limit order on exchange i would be an increasing function of $Q_i(t)$. This is because a larger number of orders with higher time priority would need to trade in order ahead of the execution of the incremental order. Thus, contingent on the execution of the incremental order, the arrival of significantly informed investors and a subsequent adverse price movement are more likely. In a similar manner, the probability of an adverse price movement conditional on getting executed at exchange i, may depend on the aggregate depth at other exchanges. As we will see in §2.2, exchanges also differ in the economics of market orders, hence, typically, low fee exchanges trade before high fee exchanges. This results in low fee exchanges receiving a larger proportion of uninformed market orders, and hence also experiencing less adverse selection. We allow that adverse selection depends on the identity of the exchange, and we denote the adverse selection associated with exchange i by AS_i. In the case where the investor chooses to take liquidity ($i = 0$), a market order executes with certainty and faces no adverse selection, so we set $\mathsf{AS}_0 \triangleq 0$.

Among the factors discussed above, both liquidity rebates and adverse selection costs are, respectively, direct financial benefits or costs that vary by exchange and are measured in dollars per share traded. As such, it is natural to combine these two effects, and define the *effective rebate* $\tilde{r}_i$ for exchange i by $\tilde{r}_i \triangleq r_i - \mathsf{AS}_i$. This represents the overall direct financial benefit per share of trading on exchange i — all else being equal, an investor prefers an exchange with higher effective rebate to one with lower effective rebate.

On the other hand, the expected delay is not a direct financial cost measured in dollars, but relates instead to the urgency to trade of the investor and is measured in units of time. While it is natural to assume that all investors prefer higher effective rebate and less delay, the precise nature of this tradeoff is idiosyncratic and depends on the investor. For example, an investor executing an algorithmic trading program seeking to execute a large parent order over a tight time horizon might be very concerned about delays in trade execution and less sensitive to rebates or adverse selection. On the other hand, a market maker faces no time horizon constraint, but would instead seek to directly maximize trading profit and thus would be very concerned about rebates and adverse selection.

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. Specifically, 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

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

In other words, γ is a tradeoff coefficient between price and delay, with units of time per dollar that characterizes the type of the heterogeneous, rational, utility maximizing investors. Given the approximate range of rebates, adverse selection costs, and expected delays, this tradeoff coefficient should roughly be in the range of 1 to 10^4 seconds per \$.01.

An alternative and equivalent formulation, which is commonly used in the economic analysis of queueing systems, is to instead convert the delay into a monetary cost by multiplying it with a delay sensitivity parameter. Finally, yet another alternative interpretation of the above criterion 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. If the anticipated delays are long relative to the delay tolerance, then investors would focus on optimizing their relative rebate; as the anticipated delays increase, however, investors would shift towards lower effective rebate exchanges that maximize their fill probability. Such a reformulation of (2) would still involve a fundamental tradeoff between monetary rebate and adverse selection, weighted against measures of delay and risk until the order gets traded. This behavior is captured in our model through the parameter γ that is inversely related to investor patience: lower values of γ correspond to more delay sensitive investors that tend to prefer exchanges with lower expected delays, while higher values of γ correspond to more patient investors that tend to prefer exchanges with higher effective rebates. In addition, an important dimension in order routing decisions is the fact that they are "dynamic," i.e., done and updated over the lifetime of the order in the market, as opposed to being "static" as portrayed in (2).

Overall, while (2) is a simplified order routing criterion, it captures the fundamental tradeoff between time and money. In spite of its simplicity, we shall see that our simplified model implies a number of structural results that is strongly supported by 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.

If such an infinitesimal investor arrives to the market at time t when the queue lengths are given by the vector $Q(t)$, the investor faces a routing decision amongst the set of exchanges $\{i : Q_i(t) > 0\}$, i.e., the exchanges with liquidity available. One factor influencing this decision is that each exchange charges a fee for taking liquidity, and these fees vary across exchanges. The fees are closely linked with the rebates discussed earlier, and typically the fee is slightly higher than the rebate, and the exchange pockets the difference as a profit. Fee and rebate data is given in §4. For the purposes of this discussion, we assume that the fee for taking liquidity on exchange i is exactly equal to the rebate r_i. Since a market order does not face any execution delay or adverse selection,

it is natural to route the order to exchange i^* so as to minimize the fee paid according to

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

In reality, routing decisions may differ from those predicted by fee minimization for a number of reasons: (a) Real order sizes are not infinitesimal. In order to trade a significant quantity, an investor may need to split an order across many exchanges. This biases orders to exchanges with larger available liquidity, i.e., greater queue lengths. (b) Even 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 effect is accentuated if there are only a few limit orders posted at an exchange. Hence, this also creates 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 incur 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 economics and marketing to capture consumer choice behavior. In particular, given queue lengths $Q(t)$, 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 $f_i(Q_i(t))$. We assume that $f_i(0) = 0$ — market orders will not route to exchange i when there is no liquidity present. 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, i.e., available liquidity attracts market orders. Given a fixed quantity q of liquidity, however, we expect that the attraction values $f_1(q), \ldots, f_N(q)$ be approximately ordered according to the fees of the corresponding exchanges. In other words, in our model market orders are biased towards exchanges with low fees and large liquidity.

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

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

Here, the β_i is a coefficient that captures the attraction of exchange i, per unit of available liquidity. We posit (but our model does not require) that $\beta_1, \ldots, \beta_N$ be ordered inversely to the fees of the corresponding exchanges. We will revisit this empirically in §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

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

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[4] by the current market state $Q(t)$.

3. Equilibrium Analysis

Suppose that at some point in time a high rebate exchange has a very short delay relative to other exchanges. Then, the routing logic in (2) will divert many limit orders towards this exchange, increasing the delays and erasing its relative advantage viz the other exchanges. This type of argument suggests that the queue lengths evolve over time and eventually converge into some equilibrium configuration where no exchange seems to have a relative advantage over others with respect to its rebate/delay tradeoff taking into account the investors' heterogeneous preferences. We wish to characterize and understand such equilibrium configurations.

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

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

Denote the equilibrium queue length vector by Q^*. In equilibrium, we have the intuitive conditions that the total rate of arrival of orders into any exchange matches the rate of departure. This leads to the system of equations

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

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. The focus of this paper

[4]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.

will be on characterizing this equilibrium and its structural consequences, and subsequently testing their validity econometrically.

The remainder of this section will offer a formal definition of equilibrium that highlights the interplay between flow balance and investor heterogeneity in §3.1, and, in §3.2, demonstrate that the equilibrium queue lengths are tightly coupled across exchanges, establishing a form of *state space collapse*. Finally, in §3.3, we exploit this state space collapse property to explicitly characterize the structure of the market equilibrium.

3.1. Equilibrium Definition

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

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

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 paramaters (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.,

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

When the queue lengths are given by Q^*, the expected delay on each exchange i is given by $Q_i^*/\mu_i(Q^*)$. Hence, the individual rationality condition ensures that all limit order investors of type γ are routed to destinations optimal for their price-delay trade-off, consistent with the optimization criteria (2), assuming fixed queue lengths Q^*.

The flow balance condition (ii), ensures that under π^*, the queue length vector Q^* remains stationary, since the inflow of limit orders adding liquidity to each exchange, consisting of both dedicated limit order flow and optimized limit order flow, is exactly balanced by the outflow of market orders removing liquidity from each exchange. Note that Definition 1 is consistent[5] with the informal equilibrium system of equations (8) given earlier, given the observation that

$$\chi_i(Q^*) = \int_0^\infty \pi_i^*(\gamma)\, dF(\gamma).$$

3.2. State Space Collapse

The equilibria of Definition 1 consist of a set of routing decisions π^* and an N-dimensional vector of queue lengths Q^*. In general, at any instant of time, the queue lengths $Q(t)$ encapsulate the "state" of the network that determines the routing behavior of both optimized limit orders through (2) and of market orders through (4). At the *equilibrium* (π^*, Q^*), however, the state of the network can be describe with a more parsimonious sufficient statistic.

In particular, 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. It provides a measure of the total amount of "work" already in the system, as viewed from the perspective of an arriving limit order investor contemplating a routing decision. 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 and more greatly impact the delays experienced by the limit order investor at all exchanges. In fact, from the delay equation (1) and the market order routing equation (4), the expected delay on exchange i is given by

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

In order words, the 1-dimensional workload is sufficient to determine delays at every exchange. This suggests that the workload provides a single aggregate measure of the liquidity available across exchanges.

In fact, Theorem 1 below establishes something stronger: in equilibrium, the entire configuration of resting limit orders Q^* (i.e., the state vector of the system) is determined by the equilibrium workload W^*. This is a notion of *state space collapse*: in equilibrium, the state variables Q^* are not arbitrary points in $\mathbb{R}_+^N$, but, rather, are contained in a one dimensional manifold parameterized by the workload W^*.

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

[5]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.

(i) π^* is an optimal solution for

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

(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 Q^* is defined by

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

for each exchange $i \neq 0$.

Alternatively, suppose that (π^*, Q^*) is an equilibrium, and 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). ■

Theorem 1 provides a characterization of equilibria in terms of (π^*, W^*). Condition (i) simply states that all limit order investors in the system are individually rational, when faced with delays implied by the workload W^*, cf. (10) and (12). Condition (ii), is a market-wide flow balance equation: it equates the total number of limit orders arriving to all exchanges to the total number of market orders arriving to all exchanges.

Given a pair (π^*, W^*) satisfying conditions (i) and (ii), the queue lengths Q^* are determined as a function of workload W^* through the *lifting map* (15). The lifting map distributes the workload across exchanges in a way that balances effective rebates with delays so that the arrivals equate departures at every exchange taking into account investor heterogeneity with respect to their price-delay tradeoff coefficient γ. The intuition behind the lifting map is that the right side of (15) is the product of the total arrival rate of limit orders (both dedicated and optimized) to exchange i and the equilibrium expected delay at exchange i, cf. (12). This is the total quantity of new limit orders that arrive over the average time that a limit order spends in the queue, which depends on $F(\cdot)$. By Little's law, this must be the equilibrium number of limit orders in the queue.

3.3. Equilibrium Characterization

The state space collapse result of Theorem 1 allows us to study the behavior of N independent limit order books as a one dimensional system through the sufficient statistic of workload. In this section, we use this structure to explicitly construct equilibria.

To begin, we make the following assumption, which we will see shortly is sufficient for the existence of an equilibrium:

Assumption 1. *Assume that*

(i) The cumulative distribution function $F(\cdot)$ over the price-delay tradeoff coefficients γ for optimizing limit order investors is non-atomic with a continuous and strictly positive density on the non-negative reals.

(ii) The arrival rates (λ, Λ, μ) satisfy

$$\sum_{i=1}^{N} \lambda_i < \mu < \Lambda + \sum_{i=1}^{N} \lambda_i.$$

In other words, the arrival rate of market orders is bounded by the minimum and maximum possible arrival rate of limit orders.

(iii) Each exchange $i \in \{1, \ldots, N\}$ satisfies $\tilde{r}_i > \tilde{r}_0$. In other words, considering only price and assuming immediate execution, limit orders on any exchange are preferred to a contra-side market order.

Now, observe that given arrival rates (λ, Λ, μ) for dedicated limit order flow, optimized limit order flow, and market order flow, respectively, condition (ii) of Theorem 1 requires that, in equilibrium, the total quantity of limit orders arriving to the market balance with the total quantity of market orders arriving to the market. Since the arrival rates for dedicated limit order flow and for market order flow are exogenously fixed, the only degree of freedom available to satisfy this condition is to induce a subset of the optimized limit order flow to forgo limit orders. Intuitively, it is reasonable to expect that this be the most impatient investors, i.e., those of type $\gamma \leq \gamma_0$, for some threshold γ_0. The threshold γ_0 is 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 $\gamma \leq \gamma_0$ not to submit limit orders, the routing criteria (2) requires that

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

for all $\gamma \leq \gamma_0$. 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. The following lemma, whose proof is provided in the Online Supplement, formally establishes that this holds in every equilibrium.

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 the exchange has maximal κ_i, i.e., if $i \in \operatorname{argmax}_{j \neq 0} \ \kappa_j$.*

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 investors. These are the marginal investors in the sense of being 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,

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

Then, the equilibrium workload is determined via $W^* = \gamma_0\mu\kappa_{\bar{i}}$. Theorem 2, whose proof can be found in the Online Supplement, summarizes the discussion above and gives a complete equilibrium characterization.

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

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

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., 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* (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 several market structure and market design issues, as well as exchange 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 the equilibrium queue lengths are 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^*.*

To recapitulate, smart order routing decisions in the placement of limit orders, via (2), and the routing of market orders, via (4)–(5), couple the dynamics across exchanges in two ways: the expected delays $\{\mathsf{ED}_i\}$ across exchanges affect the former, while the market depths $\{Q_i\}$ across exchanges affect the latter.

4. Empirical Results

The preceding results lead to several testable hypotheses that we will explore in this section. First, motivated by our state space collapse result 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. The specific prediction of our model is that the expected delays across exchanges $\{\mathsf{ED}_i\}_{i=1,\ldots,N}$ live in a one dimensional subspace. Second, (12) crisply articulates that the coupling across exchanges should be in term of their anticipated delays, which itself can be tested. The precise form of the equilibrium state and coupling relation across exchanges derived above 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.

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 of the best bid and offer on each exchange) and trades (price and size of all market transactions), 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). We restrict our attention to 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,[6] 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. In our subsequent analysis we will also assume that the structural and economic characteristics of the market as captured by 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, and exhibit both predictable time-of-day and day-of-month variations as well as unpredictable variations due to news, market events, etc. In our subsequent analysis we will estimate these rates for each stock by averaging the event activity over one hour time intervals[7] Specifically, we divided each trading day into 1 hour slots between 9:45am and 3:45pm (i.e., excluding the opening 15 minutes and the closing 15 minutes). This

[6]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.

[7]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. At a high level, for the liquid stocks of our observation set, the results we obtain do not seem sensitive to the choice of time discretization. For example, we can obtain qualitatively similar results using 15 minute time slots.

	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 the period in question. 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". That is, the quoted rebate is only available to traders at the highest tiers of volume on the exchange; other traders may receive significantly lower rebates.

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 $\mu_j^{(s,j)}(t)$ 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

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

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 also 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. These 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. These 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.

4.2. Estimation of the Market Order Routing Model

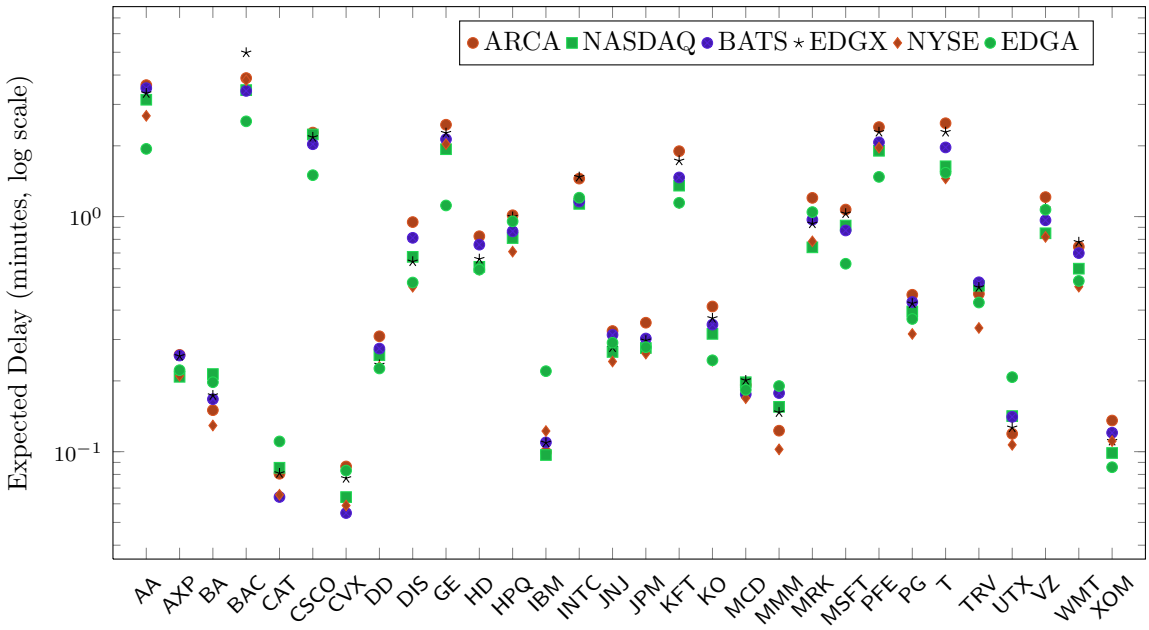
Define $\mu^{(s,j)}(t)$ to be the total arrival rate of market orders for security j and side $s \in \{\mathsf{BID}, \mathsf{ASK}\}$ in time slot t, i.e.,

$$\mu^{(s,j)}(t) \triangleq \sum_{i=1}^{N} \mu_i^{(s,j)}(t).$$

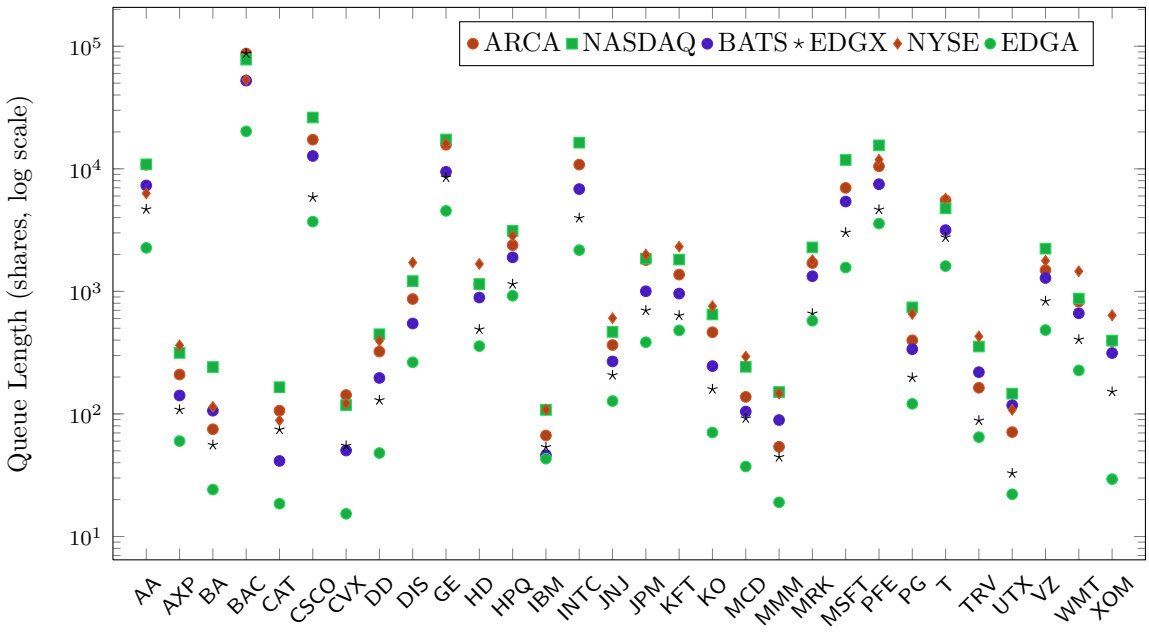
The attraction model of §2.2 for market orders suggests the relationship

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

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. Then, given that $\{\mu_i^{(s,j)}(t)\}$, $\{\mu^{(s,j)}(t)\}$, and $\{Q_i^{(s,j)}(t)\}$ are all observable, we estimated the attraction coefficients for each stock on each exchange using a nonlinear regression on (25). The results are given in Table 3. Note that all attraction coefficient estimates are statistically significant.



(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.

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

4.3. Empirical Evidence of State Space Collapse

The main theoretical predictions of our model concern the state space collapse property and its consequences with respect to the role of the workload process and the coupling across exchanges. 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 coupled, and restricted to a one-dimensional subspace. Moreover, it specifies that the states across exchanges are linearly related, and that this with respect to their estimated delays, rather than their quote depths, and finally 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.

Principle component analysis (PCA). A direct statistical test of the above observation that does not rely on the structural assumptions of our underlying model, is to study the effective

	% 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 4: Results of PCA: how much variance in the data can the first two principle components explain

dimensionality of the set of empirically observed expected delay vector trajectories

$$\left\{\mathsf{ED}^{(s,j)}(t) \ : \ t=1,\ldots,T;\ s=\mathsf{BID},\mathsf{ASK}\right\} \tag{26}$$

where $\mathsf{ED}^{(s,j)}(t)$ was estimated in (24) and the trajectories consider all one hour time slots in the 21 days of our observation period. A natural way to do this is via PCA. In particular, we estimate the effective dimension of the set of expected delay vectors in (26) by examining the number of principle components necessary to explain the variability of the data. Numerical results across the stocks in our data set are given in Table 4. We find that 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 results in the observed coupling across exchnages. On the slower timescale where the rates of events change over time, possibly in a unpredictable manner, one observes that the system evolves from one configuration state vector to another, but in each case staying close to some target equilibrium state that depends on the prevailing vector λ, Λ, μ at that time. The strong dimension reduction observed in these results persists as 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.

Regression analysis. If we define $W^{(s,j)}(t)$ to be the workload for side s of security j in time slot t, i.e.,

$$W^{(s,j)}(t) \triangleq \sum_{i=1}^{N} \beta_i^{(j)} Q_i^{(s,j)}(t), \tag{27}$$

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

$$\mathsf{ED}^{(s,j)}(t) = \frac{W^{(s,j)}(t)}{\mu^{(s,j)}(t)} \left(\frac{1}{\beta_1^{(j)}}, \ldots, \frac{1}{\beta_N^{(j)}} \right). \tag{28}$$

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,

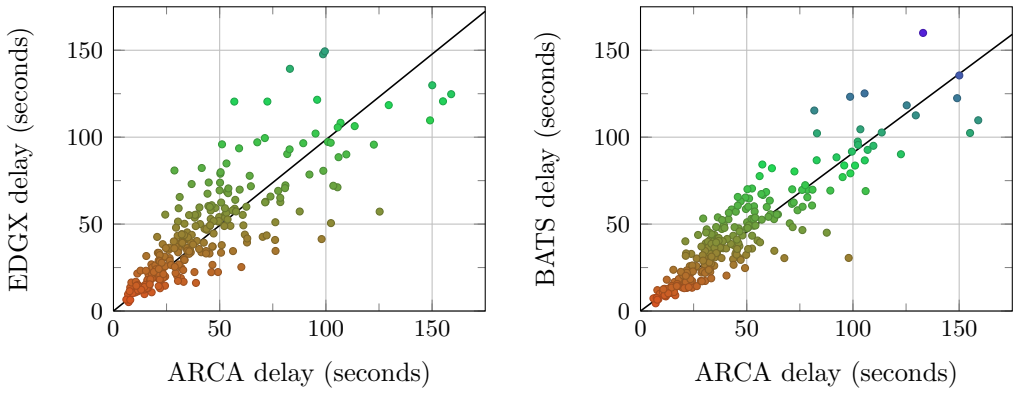
$$\mathsf{ED}_i^{(s,j)}(t) = \frac{\beta_{i'}^{(j)}}{\beta_i^{(j)}} \, \mathsf{ED}_{i'}^{(s,j)}(t), \tag{29}$$

for each time slot t.

To test this prediction, for each security, we perform a linear regression of the expected delay of that security on a particular exchange, as a function of the expected delay on a benchmark exchange (ARCA). This regression is without an intercept, and is performed using the expected delay measurements outlined in (24) for all time slots and both sides of the market, i.e., by dividing the average observed queue size in each exchange with its respective observed rate of trading. The results of these regressions are summarized in Table 5. The average R^2 across all regressions is 82%, and all of the regressions are statistically significant. This provides strong evidence of a linear relationship between delays across exchanges, and indirectly validates that the stationary attraction model estimated in §4.2 results in a reasonably good fit for market order routing decisions.

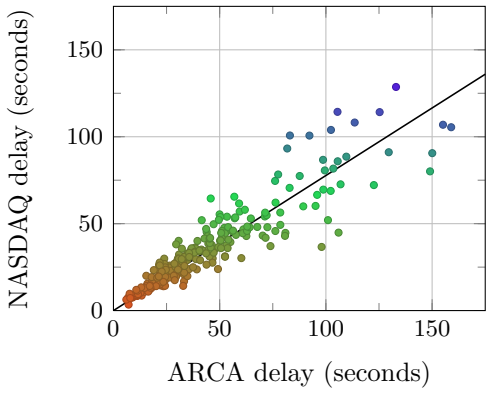
This linear relationship is illustrated in particular for the stock of Wal-Mart in Figure 3. Here, we see that the linear relationship holds across all exchanges over periods that vary significantly with respect to their prevailing market conditions, as is manifested in the roughly two orders of magnitude variation in estimated expected delays. While the regression slopes in Figure 3 differ from those predicted by the linear relationship (29), they have the same ordering. That is, the relative slopes of any two exchanges in Figure 3 are inversely ordered according to the corresponding attraction coefficients of the exchanges (cf. Table 3).

Residual analysis. The third prediction of the model pertains to the specific relationship (28) that gives a one-dimensional prediction of the state of the various exchanges. That is, 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 (27), and then construct estimates for the expected delays at the various exchanges via (28). We denote the resulting delay estimates by $\hat{\mathsf{ED}}^{(s,j)}(t)$, where the $\hat{}$ notation denotes in this context is the estimate obtained via the one-dimensional workload process, as opposed to measuring the actual expected delay $\mathsf{ED}^{(s,j)}(t)$ via (24). In the sequel

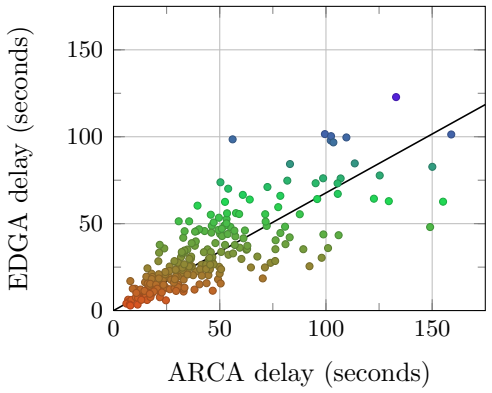


(a) slope = 0.99, $R^2 = 89\%$

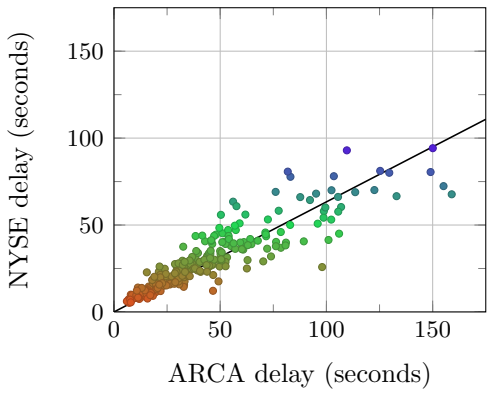
(b) slope = 0.91, $R^2 = 94\%$



(c) slope = 0.78, $R^2 = 95\%$



(d) slope = 0.68, $R^2 = 87\%$



(e) slope = 0.63, $R^2 = 94\%$

Figure 3: Scatter plots of the expected delay for Wal-Mart on each exchange, versus the delay on ARCA. The black lines correspond to linear regressions.

	NASDAQ		BATS		EDGX		NYSE		EDGA	
	Slope	R^2	Slope	R^2	Slope	R^2	Slope	R^2	Slope	R^2
Alcoa	0.85	0.83	0.95	0.93	0.90	0.76	0.72	0.88	0.50	0.91
American Express	0.53	0.66	0.69	0.68	0.68	0.60	0.53	0.64	0.56	0.62
Boeing	1.29	0.91	1.01	0.86	1.12	0.85	0.77	0.90	1.22	0.81
Bank of America	0.84	0.92	0.82	0.90	1.28	0.84	1.01	0.77	0.63	0.86
Caterpillar	0.97	0.91	0.77	0.89	0.94	0.75	0.76	0.91	1.19	0.80
Cisco	0.97	0.95	0.86	0.93	0.95	0.90	-	-	0.63	0.90
Chevron	0.72	0.92	0.61	0.92	0.83	0.84	0.65	0.92	0.87	0.78
DuPont	0.78	0.95	0.85	0.93	0.69	0.83	0.67	0.94	0.65	0.86
Disney	0.66	0.95	0.82	0.92	0.65	0.87	0.46	0.91	0.50	0.86
General Electric	0.77	0.96	0.83	0.94	0.90	0.82	0.81	0.94	0.43	0.94
Home Depot	0.71	0.96	0.88	0.95	0.77	0.90	0.70	0.95	0.70	0.92
Hewlett-Packard	0.75	0.93	0.79	0.93	0.94	0.86	0.64	0.91	0.89	0.88
IBM	0.92	0.92	1.07	0.91	1.05	0.78	1.18	0.92	2.05	0.90
Intel	0.72	0.92	0.73	0.93	1.01	0.85	-	-	0.83	0.89
Johnson & Johnson	0.73	0.92	0.88	0.87	0.76	0.86	0.65	0.91	0.74	0.86
JPMorgan	0.76	0.96	0.83	0.95	0.81	0.90	0.71	0.96	0.74	0.92
Kraft	0.58	0.85	0.65	0.85	0.81	0.80	0.49	0.87	0.44	0.73
Coca-Cola	0.74	0.97	0.83	0.95	0.88	0.87	0.54	0.94	0.53	0.83
McDonalds	0.94	0.93	0.89	0.94	0.99	0.78	0.81	0.90	0.87	0.86
3M	1.07	0.82	1.27	0.87	1.02	0.75	0.71	0.88	1.24	0.72
Merck	0.57	0.92	0.77	0.92	0.73	0.82	0.62	0.93	0.83	0.88
Microsoft	0.85	0.92	0.80	0.95	0.99	0.77	-	-	0.59	0.95
Pfizer	0.74	0.92	0.83	0.94	0.92	0.87	0.78	0.92	0.58	0.92
Procter & Gamble	0.83	0.88	0.93	0.93	0.91	0.80	0.63	0.94	0.73	0.90
AT&T	0.61	0.90	0.72	0.89	0.92	0.79	0.55	0.93	0.58	0.88
Travelers	0.97	0.90	1.03	0.91	1.03	0.79	0.62	0.90	0.84	0.87
United Tech	1.11	0.92	1.07	0.91	1.04	0.84	0.79	0.91	1.37	0.61
Verizon	0.64	0.94	0.75	0.93	0.82	0.85	0.63	0.92	0.85	0.85
Wal-Mart	0.78	0.95	0.91	0.94	0.99	0.89	0.63	0.94	0.68	0.87
Exxon Mobil	0.70	0.97	0.86	0.97	0.78	0.84	0.79	0.92	0.61	0.89

Table 5: Linear regressions of the expected delays of each security on a particular exchange, versus that of the benchmark exchange (ARCA).

we analyze the residuals $\mathsf{ED}^{(s,j)}(t) - \hat{\mathsf{ED}}^{(s,j)}(t)$. 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 (28), 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 (28). 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.

To recapitulate, the analysis of the previous section showed that order routing optimization cou-

	R^2_*		R^2_*		R^2_*
Alcoa	83%	Home Depot	95%	Merck	89%
American Express	73%	Hewlett-Packard	90%	Microsoft	86%
Boeing	88%	IBM	85%	Pfizer	89%
Bank of America	83%	Intel	88%	Procter & Gamble	89%
Caterpillar	80%	Johnson & Johnson	91%	AT&T	87%
Cisco	93%	JPMorgan	95%	Travelers	79%
Chevron	87%	Kraft	86%	United Tech	63%
DuPont	91%	Coca-Cola	95%	Verizon	89%
Disney	89%	McDonalds	86%	Wal-Mart	94%
General Electric	91%	3M	82%	Exxon Mobil	91%

Table 6: The measure of performance R^2_*, which given the reduction of variability in expected delays explained by the workload relationship (28).

ples the dynamics across exchanges, and indeed simplifies their behavior by effectively reducing the multi-dimensional market into a one-dimensional process. The exchange dynamics couple in terms of their delay estimates as opposed to their queue depths, and the relationship across exchanges is linear. Finally, the one-dimensional "workload" process offers a simple characterization of the state at the various exchanges, and leads to delay estimates for each exchange. The three separate tests performed above focused sequentially on these three predictions. First, the statistical PCA analysis verified that the fragmented market dynamics can be effectively reduced to those of a lower dimensional process, and specifically that the first principle component has striking explanatory power. Second, the regression analysis between delay trajectories across exchanges verified the simple linear relation across these variables, and, third, the delay estimate based on the one-dimensional workload process and the actual delays measured in each exchange was very accurate.

As mentioned earlier in discussing Figure 2(b), queue lengths across exchanges exhibit significantly more variation that 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 complex stochastic network models, where state space collapse has been empirically verified. Stochastic network theory is rich in examples where network dynamics under certain control policies are theoretically proved to exhibit the type of dimension reduction suggested in the analysis of the previous section, and empirically observed above.

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.

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.

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.-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.

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

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. De Larrard. Price dynamics in a Markovian limit order market. Woking Paper, 2010.

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.

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.

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.

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.

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.

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.

Online Supplement to "Optimal Order Routing in a Fragmented Market"

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: May 21, 2012

A. Proofs

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,

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

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,

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

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

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

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 $\left(\pi^{(1)}, Q^{(1)}\right)$ and $\left(\pi^{(2)}, Q^{(2)}\right)$ 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. ■