# Microstructure Analysis of Dynamic Markets: Limit Order Books and Dynamic Matching Markets

**Hua Zheng**

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

## ABSTRACT

**Microstructure Analysis of Dynamic Markets:**
**Limit Order Books and Dynamic Matching Markets**

**Hua Zheng**

This thesis is concerned with addressing operational issues in two types of dynamic markets where queueing plays an important role: limit order books (financial industry), and dynamic matching markets (residential real estate).

We first study the smart order routing decisions of investors in fragmented limit order book markets and the implications on the market dynamics. In modern equity markets, participants have a choice of many exchanges at which to trade. Exchanges typically operate as electronic limit order books operating under a "price-time" priority rule and, in turn, can be modeled as multi-class FIFO queueing systems. A market with multiple exchanges can be thought as a decentralized, parallel queueing system. Heterogeneous traders that submit limit orders select the exchange to place their orders by trading off delays until their order may fill against financial considerations. Simultaneously, traders that submit market orders select the exchange to direct their orders. These market orders trigger instantaneous service completions of queued limit orders. Taking into account the effect of investors' order routing decisions, we find that the equilibrium of this decentralized market exhibits a state space collapse property. The predicted dimension reduction is the result of high-frequency order routing decisions that essentially couple the dynamics across exchanges. Analyzing a TAQ dataset for a sample of stocks over a one month period, we find empirical support for the predicted state space collapse.

In the second part of this thesis, we model an electronic limit order book as a multi-class

queueing system under fluid dynamics, and formulate and solve a problem of limit and market order placement to optimally buy a block of shares over a short, predetermined time horizon. Using the structure of the optimal execution policy, we identify microstructure variables that affect trading costs over short time horizons and propose a resulting microstructure-based model of market impact costs. We use a proprietary data set to estimate this cost model, and highlight its insightful structure and increased accuracy over conventional (macroscopic) market impact models that estimate the cost of a trade based on its normalized size but disregarding measurements of limit order book variables.

In the third part of this thesis, we study the residential real estate markets as dynamic matching systems with an emphasis on their microstructure. We propose a stylized microstructure model and analyze the market dynamics and its equilibrium under the simplifying approximation where buyers and sellers use linear bidding strategies. We motivate and characterize this near closed-form approximation of the market equilibrium, and show that it is asymptotically accurate. We also provide numerical evidence in support of this approximation. Then with the gained tractability, we characterize steady-state properties such as market depth, price dispersion, and anticipated delays in selling or buying a unit. We characterize congestion and matching patterns for sellers and buyers, taking into account market dynamics, heterogeneity, and supply and demand imbalance manifested in the competition among buyers and sellers. Furthermore, we show the effects of market primitives with comparative statics results.

# Contents

## ACKNOWLEDGEMENT

I greatly acknowledge my thesis advisors, Professor Costis Maglaras and Professor Ciamac C. Moallemi. They have recreated my horizons throughout the years as mentors of illuminating guidance and unparalleled support. They introduced, distilled their profound knowledge of, and inspired me to progress in the research that is the subject matter of not only this thesis but also the career path that I am about to undertake at the moment. I have been incredibly fortunate to observe them work and profit much from their very generous investment of time in tutorials, discussions, and conversations. I feel a deep sense of gratitude toward both of them.

I would like to thank Professor Gabriel Weintraub for being an invaluable source of advice and help during my graduate study. Gabriel provided suggestions that goes into parts of this thesis and I have enjoyed all our conversations. I am indebted to Professor Garrett van Ryzin as being an exceptional role model. He is always insightful and provides for me constant support as the chair of the department. I am very grateful to Professor Paul Glasserman and Professor Garud Iyengar for serving on my thesis committee.

I have also benefited from my fellow graduate students. I want to thank John Yao, Yonatan Gur, Yunru Han, Damla Gunes, Cathy Yang, Yina Lu, Jia Liu, and Xinyun Chen. It has been a great pleasure to share the interesting debates and this whole rewarding experience with you.

This journey has been brightened up in the last three years by the companionship and constant encouragement of Yu Gu, who himself is an academic role model of mine. I have enjoyed his loving support, patience, and candid advices. He has unfailingly helped me to keep clear perspectives, stay focused, and more importantly, know when to rest.

I am forever indebted to my parents, Shun and Zhehao. I appreciate them for their

constant demonstration of love. They have sacrificed a lot of quality time for my absence in the past few years but indulged my pursuit with all their support. The thesis is dedicated to them.

*To my parents - Shun and Zhehao*

# Chapter 1

# Introduction

Many marketplaces of current interest are ones in which participants may (choose to) seek for transactions over time, resulting in queue/inventory of buyers and sellers. This thesis is concerned with operational issues that arise in dynamic markets where the queueing phenomenon plays an important role in the functioning of the market and the trading decisions therein. In the chapters to come, we explore two specific applications that are of great economic significance within this framework:

(1) In financial markets, exchanges operate as electronic limit order books. The orders posted by investors are prioritized for trade first based on price, and then, at a given price, queue based on their arrival time. The choice of price and exchange at which to post an order affect both the probability that it will trade and its anticipated delay.

(2) Another example is provided by residential real estate markets, where buyers and sellers of houses arrive sequentially over time, and can observe only a fraction of the entire market at any given time and might need to conduct several rounds of bidding or bargaining before finding the right match. This search friction creates a tradeoff between

the possibility of finding a better match in the future and the cost of time, e.g., explicit carrying costs of sellers, and/or usage needs of buyers.

In these applications, the delay induced by the competitive structure of the market, and the resulting queue/inventory of sellers and buyers, is a key ingredient in understanding the functioning of the market as well as the optimization of trading decisions therein. This thesis focuses on building and studying queueing models of such markets with specific bearing on operational questions, e.g., what are the transaction costs of buying or selling a large quantity of stock on electronic exchanges within a certain time horizon? How much faster will a house sell if you lower the price by 5%? How can one interpret market conditions and quantify risks/rewards from observations such as historical housing transaction data?

In the sequel, we provide some detail about our modeling approach and results that we obtain in each chapter.

## 1.1. Queueing Dynamics in Limit Order Book Markets

### 1.1.1. Fragmented Market and State Space Collapse

In Chapter 2 of this thesis, we study the smart order routing decisions of investors in fragmented limit order book markets and the implications on the market dynamics. In many financial markets, participants have a choice of many exchanges at which to trade one security, e.g., NASDAQ, NYSE, BATS in the context of U.S. equities. A market with multiple exchanges can be thought as a decentralized, parallel queueing system. Traders that submit limit orders specify limits on acceptable price of their order and wait to trade in the book according to a price-time priority rule. These limit orders can be thought as jobs waiting for service. Traders that submit market orders are willing to accept the best available price at the time and trigger instantaneous service completions of queued limit orders. In this way,

both the arrival and the server in the queueing system are the aggregation of self-interested, atomistic traders that tradeoff the delay until their order may fill against financial considerations when selecting the exchange, i.e., the queue, to route their orders under heterogeneous time-money preferences.

We find that the equilibrium of this decentralized market exhibits a state space collapse property, whereby (a) the queue lengths at different exchanges are coupled in an intuitive manner; (b) the behavior of the market is captured through a one-dimensional process that can be viewed as a weighted aggregate queue length across all exchanges; and (c) the behavior at each exchange can be inferred via a mapping of the aggregated market depth process that takes into account the heterogeneous trader characteristics. This predicted dimension reduction is the result of high-frequency order routing decisions that essentially couple the dynamics across exchanges.

By analyzing a TAQ dataset for a sample of stocks over a one month period, we find empirical support for the predicted state space collapse. This seems to be one of the first examples of a complex stochastic network model where state space collapse has been empirically verified.

### 1.1.2. Optimal Execution and Market Microstructure

In Chapter 3 of this thesis, we study the problem of optimally trading a block of shares over a short, predetermined horizon in a limit order book and propose an associated microstructure-based market impact model. Execution of a large stock order in an electronic limit order book can be achieved by a combination of limit and market orders that are posted at different prices and at different times. The overall price and duration of the order depends on how the trader allocates the shares between limit and market orders, across different price levels as limit orders, and over time.

We propose a microstructure-based model of market impact that directly relates transaction costs of stock orders to "micro-level" order book features such as queue lengths and arrival rates of orders. These features are directly estimatable and bring the market impact model closer to data. Using a proprietary dataset of three months of actual algorithmic trades, we find significant increase in out-of-sample prediction accuracy of our microstructure model over market impact models in existing literature which are based on "macro-level" features.

The microstructure-based market impact model is motivated by the structure of the optimal execution policy that we characterize. In particular, the dynamics of a limit order book can be thought as a double-sided multiclass queueing system with a specific priority rule. We formulate and solve a problem to optimally buy a block of shares over a short, predetermined time horizon with limit and market order placement as an optimal fluid control problem, and characterize the optimal execution policy.

## 1.2. Dynamic Matching Markets

### 1.2.1. Residential Real Estate

In Chapter 4 of this thesis, we study a dynamic microstructure model of the residential real estate market and propose a linear approximation method to tractably analyze its dynamics, market depth, and buyer/seller bidding strategies. In the residential real estate market, sellers arrive dynamically over time to put their units up for sale. These assets may differ in their attributes, including location, acreage, etc., while sellers themselves differ in their own financial constraints, such as carrying costs. Buyers arrive dynamically over time, differing in their preferences and delay sensitivity. This market evolves sequentially, and is subject to other frictions, such as the fact that sellers and buyers can consider only a fraction of the

entire market at any time. Both phenomena imply sellers and buyers face search friction and probably will experience delay, which results in inventory of sellers that incurred explicit carrying costs and inventory of buyers whose utilities are decreased as they spend more time searching. Their decisions, i.e., how to price, which bid to accept, and whether to wait for better outcomes in the future, depend on the available inventory, its characteristics, potential mismatch between buyers and sellers, and their beliefs for potential future arrivals of better units or less patient buyers, etc., and vice versa.

We formulate a microstructure model of this market, explicitly accounting for its dynamics and the heterogeneity of buyers, sellers, and inventory. We motivate and characterize an almost closed-form approximation of the dynamic matching equilibrium that is asymptotically accurate. We also provide numerical evidence in support of the approximation. We observe in the equilibrium that delay increases with sellers' costs and decreases with buyers' valuations as a power law, or exponentially in some cases. Besides insights on congestion, our results also characterize the matching pattern between buyers and sellers when dynamics play a role, and clarify the effect of attractiveness (meaning having a low cost as a seller or having a high valuation as a buyer) in several aspects that would be otherwise ambiguous. Finally, if the market is symmetric with respect to buyers' and sellers' primitive parameters - that is, buyers and sellers arrive at the same rate, incur the same search cost, and have equal bargaining power - we can establish the existence and uniqueness of equilibrium. Furthermore, we can provide a series of comparative statics results on how the equilibrium would react to different kinds of changes in market primitives, e.g., the meeting technology, the prevailing interest rate, or the participation cost.

# Chapter 2

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

## 2.1. Introduction

**Motivation.** Modern equity markets are highly fragmented. In the United States alone there are over a dozen exchanges and about forty alternative trading systems where investors may choose to trade. Market participants, including institutional investors, market makers, and opportunistic investors, interact within today's high-frequency, fragmented marketplace with the use of electronic algorithms that differ across participants and types of trading strategies. At a high level, they dynamically optimize where, how often, and at what price to trade, seeking to achieve their own best execution objectives while taking into account short term differences or opportunities across the various exchanges. Exchanges function as electronic limit order books, typically operating under a "price-time" priority rule: resting orders are

prioritized for trade first based on their respective prices, and then, at a given price, according to their time of arrival, i.e., in first-in-first-out (FIFO) order. The dynamics of an exchange can be understood as that of a multi-class system of queues, where each queue is associated with a price level. Job arrivals into these queues correspond to new limit orders posted at the respective prices. Market orders trigger executions which, in queueing system parlance, correspond to service completions.

The market, consisting of multiple exchanges, can be viewed as a stochastic network that evolves as a collection of parallel, multi-class queueing systems. Figure 2.1 depicts one side of the market at one price level. Heterogeneous, self-interested traders optimize where to route their limit and market orders, coupling the dynamics of these parallel queues. Studying the interaction effects between market fragmentation and high-frequency, optimized order routing decisions is an important issue in understanding market behavior and trade execution, and is the main focus of this chapter. [1]

At a point in time, conditions at the exchanges may differ with respect to the best bid and offer[2] price levels, the market depth at various prices, recent trade activity, etc. Exchanges publish real-time information for each security that allow investors to know or compute these quantities. These, in turn, imply differences in a number of execution metrics across exchanges, such as the probability that an order will be filled, the expected delay until such a fill, or the adverse selection associated with a fill. Exchanges also differ with respect to their underlying economics. Under the "make-take" pricing that is common, exchanges typically

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

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

offer a rebate to liquidity providers, i.e., investors that submit limit orders that "make" markets when their orders get filled; simultaneously, exchanges charge a fee to "takers" of liquidity that initiate trades using marketable orders that transact against posted limit orders. Fees range in magnitude, and are typically between $-\$0.0010$ 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, and material in optimizing over routing decisions. Most retail investors do not have access to this information, but essentially all institutional investors and market makers — that, taken together, account for almost all trading activity — have access and do make use of this information. They employ so-called "smart order routers" that take into account real-time state information and formulate an order routing problem that considers various execution metrics in order to decide whether to place a limit order or trade immediately with a market order, and accordingly to which venue(s) to direct their order. Investors are heterogeneous; specifically they differ with respect to the way that they trade off metrics such as price, rebates, and delays, primarily driven by their intrinsic patience until they fill their order.

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

**Summary of results.** This chapter makes three contributions. First, it offers a novel model for order routing in fragmented markets. It proposes a queueing system that takes into effect the atomistic limit order placement and market order (service completions) routing

decisions. This research appears to be one of the first in financial engineering or market microstructure to study the exchange queueing dynamics and their effect on order routing decisions. It is also one of the first stochastic modeling studies to focus in this application domain, and introduce some related queueing considerations. Finally, the self-interested routing of the service completion process, may be of independent interest; e.g., one possible application might be in modeling personnel that work in retailing that may strategize over which customer to help next.

Second, from a methodological viewpoint, we study a deterministic and continuous fluid model associated with the above system, that takes into account the routing decisions of atomistic limit order placements and market orders (service completions). The key result is to characterize the structural form of the equilibrium state of this fluid model and derive a form of state space collapse (SSC) property.[3] The market equilibrium and SSC are not the result of the price protection mechanism[4] imposed in the U.S. equities market. Rather, they arise out of order routing decisions among exchanges that offer the same price level at different (rebate, delay) combinations. We characterize this coupling effect that yields a strikingly simplifying property whereby the behavior of the multi-dimensional market reduces to that of a one-dimensional system expressed in terms of what we refer to as *workload*, which is an aggregate measure of the total available liquidity. In equilibrium, the workload is a sufficient statistic that summarizes the state of the market: queue lengths can be inferred from it, as can the

[3]SSC results tend to be pathwise properties, established via an asymptotic analysis after an appropriate rescaling of time. In our system, arrival rates of limit and market orders vary stochastically over time on a slower time scale than that of the transient fluid model dynamics. An asymptotic analysis on the slower time scale of the event rate variations, in the spirit of the so called Pointwise-Stationary-Fluid-Models (PSFM), would establish such a pathwise SSC property by exploiting the transient fluid model results of this chapter. Standard machinery for establishing such results either exploit the work by Bramson (1998) or Bassamboo et al. (2006). Our model seems to satisfy the key requirements that one would need to derive the PSFM and as a result the sample path version of the SSC property, but we will not pursue this in this chapter.

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

routing behavior of investors. The expected delay at each exchange is proportional to the workload, where the proportionality constant depends on exchange specific parameters. In equilibrium, if one exchange is experiencing long delays, then the other exchanges will also be experiencing proportionally long delays. Conversely, if (out of equilibrium) one exchange has temporarily an atypically small associated delay relative to its cost structure, the new order flow will quickly take advantage of that delay/cost opportunity and erase that difference. A simpler version of this effect is the familiar picture we encounter in highway toll booths or supermarket checkout lines, where people join the shortest queue; in our model choice behavior is more intricate, and depends on economics, anticipated delays, as well as trader heterogeneity. For $N = 2$ exchanges, we use a geometric argument to prove that the fluid model transient starting from an arbitrary initial condition converges to the equilibrium state in finite time. We conjecture that a similar argument carries through when there are $N > 2$ exchanges.

The 1-dimensional workload system seems to offer a tractable model for downstream analysis of interesting questions that pertain to exchange competition (e.g., how to set fees or associated volume tiers), policy questions that may affect the routing decision problem or impose exogenous transaction costs (e.g., a transaction tax), and market design questions (e.g., whether the co-existence of competing, differentially priced exchanges is beneficial from a welfare perspective).

Third, we empirically verify the state space collapse property for a sample of TAQ data for the month of 9/2011 for the 30 securities that comprise the Dow Jones Index. While all being liquid stocks, these securities differ in their trading volumes, price, volatility, and spread. Our methodological results suggest certain testable hypotheses, most notably regarding the effective dimensionality of the market dynamics, the linear relation between the expected delays across exchanges, and the relation between expected delays and market-wide workload.

We test the implications of our model in several ways. First, we perform a principal component analysis (PCA) to characterize the effective dimension of the joint vector of expected delays across exchanges. In support of our theoretical prediction that this vector should be contained in a one-dimensional set, we find that the first principal component explains around 80% of the variability of the expected delays across exchanges, and that the first two principal components explain 90%. Second, our analysis suggests that the expected delays across exchanges are linearly related, in fact, proportional to each other. We test this prediction by running a set of linear cross-sectional regressions over different pairs of exchanges and find statistically significant support for a linear relationship in all cases. The $R^2$ varies between 76% and 87%. The regression coefficients are also very close to the ones predicted by our analysis. Similarly good fits are obtained if one were to carry through the analysis separately for each security. The SSC result suggests that the expected delays in each exchange can be inferred through the market-wide workload. This prediction can be tested again through a set of linear regressions between the the workload delay estimate and the delay estimate that uses information about the state of the exchange (queue length and trading rate). All these regressions are again statistically significant and are accompanied with high $R^2$ values. We do not report on these results, instead we pursue a more detailed analysis of the residuals, i.e., the errors between the workload and exchange-specific delay estimates, and find that the workload estimate captures on the average 80% of the variation in exchange delays. Overall, the empirical analysis provides statistical support for the SSC prediction of our model, despite the fact that many of the assumptions of our model may be not satisfied in practice. To our knowledge, this seems to be one of the first empirical verifications of SSC in a real and complex stochastic processing system.

This section concludes with a literature survey. Section 4.2 sets up the one-sided, top-of-book model of the limit order book markets and then describes two order routing models:

one for limit orders and the other for market orders. Our main results on market equilibrium and state space collapse are given in Section 2.3. In Section 3.6, we show empirical evidence of state space collapse.

**Literature Survey.** There are two strands of literature that we briefly review. The first is on market microstructure and financial engineering, and focuses on the structure and behavior of limit order books. Apart from the classical market microstructure models, such as those proposed by Kyle (1985), Glosten and Milgrom (1985) and Glosten (1987), our work is related to several strands of work. First is the set of papers that report on empirical analyses of the dynamics of exchanges that operate as electronic limit order books, such as Bouchaud et al. (2004), Griffiths et al. (2000), and Hollifield et al. (2004) and the review article Parlour and Seppi (2008). Related to the above work, there is a body of literature that studies the effect of adverse selection, which factors in order placement decisions; c.f., Keim and Madhavan (1998), Dufour and Engle (2000), Holthausen et al. (1990), Huberman and Stanzl (2004), Gatheral (2010), and Sofianos (1995).

Second, there are several papers that study market fragmentation, exchange competition and their effect on market outcomes dating back to the work of Hamilton (1979), Glosten (1994, 1998), and, more recently, Bessembinder (2003) and Barclay et al. (2003). A number of papers, including O'Hara and Ye (2011), Jovanovic and Menkveld (2011), and Degryse et al. (2011), empirically study the impact of exchange competition on available liquidity and market efficiency. Biais et al. (2010) and Buti et al. (2011) consider the impact of differences in tick-size on exchange competition, while in the markets we consider, the tick-size is uniform. Foucault et al. (2005) describe a theoretical model to understand make-take pricing when monitoring the market is costly. Malinova and Park (2010) empirically study the introduction of make-take rebates and fees in a single market. Foucault and Menkveld (2008) studies the impact of smart order routing on market behavior in a setting with two

exchanges and focusing, however, on smart order routing decisions by traders submitting market orders aiming to optimize their execution price (i.e., in a setting where exchanges operate without a price protection mechanism, like Reg NMS that applies to the U.S. equities market, that would eliminate the opportunity from such routing decisions); their paper does not consider the routing decisions of limit orders, and disregards queueing effects. van Kervel (2012) considers the impact of order routing in a setting where market makers place limit orders on multiple exchanges simultaneously so as to increase execution probabilities. Their analysis ignores economic and execution delay differences between venues. Sofianos et al. (2011) discuss smart order placement decisions in relation to their all-in cost, introducing similar considerations to the ones explored in this chapter, and Cont and Kukanov (2013) formulates a smart order routing control problem.

Third, there is a growing body of work that develops models of limit order book dynamics and studies optimal execution problems. Obizhaeva and Wang (2013), Rosu (2009), Alfonsi et al. (2010), Parlour (1998), treat the market as one limit order book and use an aggregated model of market impact and abstracts away queueing effects. The high-frequency behavior of limit order books can probably be best modeled and understood as that of a queueing system. This connection has been explored in recent work, starting with Cont et al. (2010); see also Maglaras and Moallemi (2011), Cont and Larrard (2013), Lakner et al. (2013), Blanchet and Chen (2013), Stoikov et al. (2011), Guo et al. (2013), and Lakner et al. (2014).

The second strand of literature related to our work is on stochastic modeling and relates to the asymptotic analysis tools that motivate our method of analysis and the area of queueing systems with strategic consumers. So-called equivalent workload formulations and the associated idea of state space collapse arise in stochastic network theory in the context of their approximate Brownian model formulations. This idea has been pioneered by the work of Harrison (1988) and Harrison (2006). Workload fluid models were introduced in Harri-

son (1995). The condition that guarantees that parallel server systems exhibit SSC down to one-dimensional systems was introduced by Harrison and Lopez (1999), and two papers that establish SSC results with optimized routing of order arrivals are Stolyar (2005) and Chen et al. (2010). We model market order routing decisions via a reduced form state dependent service rate process. Mandelbaum and Pats (1995) derive fluid and diffusion approximations for such queues.

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

Most of the above papers look at static rules, where consumers make decisions based on steady-state expected delays. Chen et al. (2010) considers competing suppliers and arriving consumers making decisions based on real-time information, like in our model, but where each supplier has his own dedicated processing capacity; the resulting dynamics are different and only couple through order arrivals. The nature of the service completion process that emerges as the aggregation of infinitesimal self-interested contributions appears novel viz the existing literature. Finally, Plambeck and Ward (2006) study an assemble-to-order system, that involves a two-sided market fed by product requests on one side and raw materials on the other, but such systems allow queueing on both sides and the flow of material is controlled by the system manager.

## 2.2. Model

We propose a stylized model of a fragmented market consisting of $N$ distinct electronic limit order books simultaneously trading a single underlying asset. The model will take the form of a system of parallel FIFO queues; new price and delay sensitive jobs arrive over time and optimize their routing decisions; self-interested agents arrive over time and optimize where to route their market order that triggers an instantaneous service completion at the respective queue (i.e., this routing decision happens at the "end of the service time"). Our focus is to understand the effect of optimized order routing decisions on the interaction between multiple limit order books. We make a number of simplifying assumptions that aid the tractability of our model studied in Sections 4.2–2.3.

*One-sided market.* We model one side of the market, which, without loss of generality, choose to be the bid side, where investors post limit orders to buy the stock and wait to execute against market orders directed by sellers.

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

*Fluid model.* We consider a deterministic fluid model, or "mean field" model, where the discrete and stochastic order arrival processes are replaced by continuous and deterministic analogues, where infinitesimal orders arrive continuously over time at a rate that is equal to the instantaneous intensity of the underlying stochastic processes. This model can be justified as an asymptotic limit using the functional strong law of large numbers in settings where the rates of order arrivals grow large but the size of each individual order is small relative to the

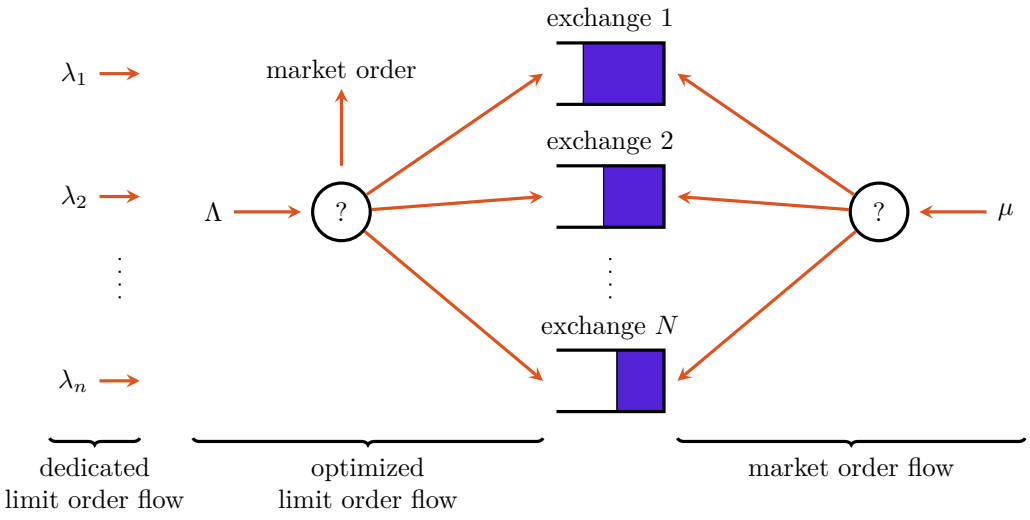



**Figure 2.1:** A one-sided, top-of-book model of multiple limit order books. Limit orders (i.e., jobs) arrive to each exchange (modeled by the respective queues) in a) dedicated streams and b) optimized limit order placement decisions. Liquidity is removed through the arrival of decentralized, self-interested market orders, acting as service completions.

overall order volume over any interval of time. It is well suited for characterizing transient dynamics in such systems, which is time scale over which queue lengths drain or move from one configuration to another; this is also the relevant time scale in order routing decisions. For liquid securities, orders arrive on a time scale measured in milliseconds to seconds, while queueing delays are of the order of seconds to minutes.

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

Our model is illustrated in Figure 2.1. For each of the $N$ exchanges, there is a (possibly empty) queue of resting limit orders at the national best bid price. The vector of queue lengths at time $t$ is denoted by $Q(t) \triangleq (Q_1(t), Q_2(t), \ldots, Q_N(t)) \in \mathbb{R}_+^N$.

### 2.2.1. Limit Order Routing

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

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

*Optimized limit order flow* arrives at a rate $\Lambda > 0$. Each infinitesimal investor observes the state of the market, $Q(t)$, and optimizes over where to route the associated infinitesimal order, or, if conditions are unfavorable, not to leave a limit order and to trade instead with a market order at the offered (other) side of the market; this option is denoted by $i = 0$.

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

*Expected delay.* All things being equal, an investor would prefer a shorter delay until an order gets executed. Apart from price risk considerations, this is often due to exogenous constraints on the speed at which the order needs to get filled; in many instances, a limit order may be a "child order" that is part of the execution plan of a larger "parent order," which itself needs to be filled within a limited time horizon and under some constraints on its execution trajectory defined by its "strategy." As will be seen in Section 3.6, the expected delays vary in the range of 1 to 1000 seconds.

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

$$\mathsf{ED}_i(t) \triangleq \frac{Q_i(t)}{\mu_i}. \tag{2.1}$$

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

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

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

In practice, order placement decisions depend on various factors in addition to the ones described above. For example, an investor may have explicit views on the short-term movement of prices ("short-term alpha"), and these can be relevant for the placement of limit orders; be sensitive to adverse selection, or the anticipated price movement after the execution of a limit order; etc. In order to maintain tractability, we will focus on the direct trade-off between financial benefits and delays. We will denote the financial benefit per share

traded associated with exchange $i$ by $\tilde{r}_i$ and refer to it as the *effective rebate*; this includes the direct exchange rebate but possibly incorporates other financial considerations. All else being equal, a higher effective rebate is preferable.

We denote the opportunity set of effective rebate and delay pairs encountered by an investor arriving at time $t$ by $\mathcal{E}(t) \triangleq \{(\tilde{r}_i, \mathsf{ED}_i(t)) \ : \ 0 \leq i \leq N\}$. Investors are heterogeneous with respect to their way of trading off rebate against delay. Each investor is characterized by its type, denoted by $\gamma \geq 0$, that is assumed to be an independent identically distributed (i.i.d.) draw from a cumulative distribution function $F(\cdot)$, that is differentiable and has a continuous density function, and selects a routing decision $i^*(\gamma)$ so as to maximize his "utility" according to the rule [5]

$$i^*(\gamma) \in \underset{i \in \{0,1,\ldots,N\}}{\operatorname{argmax}} \quad \gamma \tilde{r}_i - \mathsf{ED}_i(t). \tag{2.2}$$

In other words, $\gamma$ is a trade-off coefficient between price and delay, with units of time per dollar, that characterizes the type of the heterogeneous investors. Given the range of rebates and expected delays, this trade-off coefficient should roughly be in the range of 1 to $10^4$ seconds per \$.01.

An equivalent formulation, which is commonly used in the economic analysis of queues, is to convert the delay into a monetary cost by multiplying it with a delay sensitivity parameter. Yet another alternative interpretation would assume that investors differ in terms of their expected delay tolerance, i.e., the maximum length of time they are willing to wait for an order to be filled. Overall, while (2.2) is a simplified criterion, it captures the fundamental trade-off between time and money, and it will ultimately yield structural results that are consistent with our empirical analysis.

---

[5]The criterion (2.2) is "static." In practice, order routing decisions are "dynamic," i.e., done and updated over the lifetime of the order in the market.

### 2.2.2. Market Order Routing

Investors arrive to the market continuously at an aggregate rate $\mu > 0$, seeking to sell an infinitesimal quantity of stock instantaneously via a market order. For an investor who arrives to the market at time $t$ when the queue length vector is $Q(t)$, the routing decision is restricted to the set of exchanges $\{i \ : \ Q_i(t) > 0\}$. One important factor influencing this decision is that each exchange charges a fee for taking liquidity, and these fees vary across exchanges. Typically the fee at an exchange is slightly higher than the rebate, and the exchange pockets the difference as a profit. Fee and rebate data is given in Section 3.6. For the purposes of this discussion, we assume that the fee on exchange $i$ is equal to the rebate $r_i$. Since a market order executes without any delay, it is natural to route it to exchange $i^*$ so as to minimize the fee paid:

$$i^* \in \underset{i \in \{1,\ldots,N\}}{\operatorname{argmin}} \ \{r_i \ : \ Q_i(t) > 0\}. \tag{2.3}$$

In practice, routing decisions may differ from those predicted by fee minimization for a number of reasons: (a) Real order sizes are not infinitesimal, and to trade a significant quantity one may need to split an order across many exchanges. (b) If an investor observes that liquidity is available at an exchange, due to latency in receiving market data information or in transmitting the market order to the exchange, that liquidity may no longer be present by the time the investor's market order reaches the exchange. This is accentuated if there are only a few limit orders posted at an exchange. Both (a) and (b) create a preference for longer queue lengths. (c) If an exchange has very little available liquidity, "clearing" the queue of resting limit orders is likely to result in greater price impact. (d) There maybe other considerations involved in the order routing decision, such as different economic incentives between the agent making the order routing decision and the end investor. All of these effects point to a more nuanced decision process than the fee minimization suggested by (2.3), which

we will capture through a reduced form "attraction" model that is often used in marketing to capture consumer choice behavior. Specifically, given $Q(t)$, the instantaneous rate at which market orders to sell arrive at exchange $i$ is denoted by $\mu_i(Q(t))$ given by

$$\mu_i(Q(t)) \triangleq \mu \frac{f_i(Q_i(t))}{\sum_{j=1}^{N} f_j(Q_j(t))}. \tag{2.4}$$

Equation (2.4) specifies that the fraction of the total order flow $\mu$ that goes to exchange $i$ is proportional to the attraction function $f_i(Q_i(t))$., with $f_i(0) = 0$, i.e., market orders will not route to an exchange $i$ with no liquidity. The discussion above suggests that $f_i(\cdot)$ is an increasing function of the queue length $Q_i$, and a decreasing function of the size of the fee charged by the exchange.

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

$$f_i(Q_i) \triangleq \beta_i Q_i, \tag{2.5}$$

where $\beta_i > 0$ is a coefficient that captures the attraction of exchange $i$ per unit of available liquidity. We posit (but our model does not require) that the $\beta_i$'s be ordered inversely to the fees of the corresponding exchanges. We will revisit this empirically in Section 3.6.

### 2.2.3. Fluid Model

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

$$Q_i(t) = Q_i(0) + \lambda_i t + \Lambda \int_0^t \chi_i(Q(s))\, ds - \int_0^t \mu_i(Q(s))\, ds. \tag{2.6}$$

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

$$\chi_i(Q(t)) \triangleq \int_{\mathcal{G}_i(Q(t))} dF(\gamma), \tag{2.7}$$

where $\mathcal{G}_i(Q(t))$ denotes the set of optimizing limit order investor types $\gamma$ that would prefer exchange $i$, i.e., the set of all $\gamma \geq 0$ with $\gamma\tilde{r}_i - \mathsf{ED}_i(t) \geq \gamma\tilde{r}_j - \mathsf{ED}_j(t)$ for all $j \notin \{0, i\}$, and $\gamma\tilde{r}_i - \mathsf{ED}_i(t) \geq \gamma\tilde{r}_0$, given the expected delays $\mathsf{ED}_j(t) = Q_j(t)/\mu_j(Q(t))$, for $j = 1, \ldots, N$, implied[6] by $Q(t)$.

## 2.3. Equilibrium Analysis

Suppose that at some point in time a high rebate exchange has a very short expected delay relative to other exchanges. Then, the routing logic in (2.2) will direct many arriving limit orders towards this exchange, increasing delays and erasing its relative advantage viz the other exchanges. This type of argument suggests that queue lengths will evolve over time and eventually converge into some equilibrium configuration where no exchange seems to have a relative advantage with respect to its rebate/delay trade-off taking into account the investors' heterogeneous preferences.

Expressing (2.6) in differential form, we have that $\dot{Q}_i(t) = \lambda_i + \Lambda\chi_i(Q(t)) - \mu_i(Q(t))$, for $i = 1, \ldots, N$. Denoting such an equilibrium queue length vector by $Q^*$, we have that:

$$\lambda_i + \Lambda\chi_i(Q^*) = \mu_i(Q^*), \quad i = 1, \ldots, N. \tag{2.8}$$

[6]Here, we employ a "snapshot" estimate of expected delays that is consistent with our definition (2.1) 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.

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.

### 2.3.1. Equilibrium Definition

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

$$\pi_i(\gamma) \geq 0, \quad \forall\, i \in \{0, 1, \ldots, N\}; \qquad \sum_{i=0}^{N} \pi_i(\gamma) = 1. \tag{2.9}$$

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 $\pi$ where $\pi(\gamma)$ is feasible for (2.9), for all $\gamma \geq 0$, and where each $\pi_i(\cdot)$ is a measurable function over $\mathbb{R}_+$. We have suppressed the dependence of $\pi$ on the rate parameters $(\lambda, \Lambda, \mu)$ 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 $\gamma$ investors is an optimal solution for*

$$\begin{aligned} \underset{\pi(\gamma)}{\text{maximize}} \quad & \pi_0(\gamma)\, \gamma \tilde{r}_0 + \sum_{i=1}^{N} \pi_i(\gamma) \left( \gamma \tilde{r}_i - \frac{Q_i^*}{\mu_i(Q^*)} \right) \\ \text{subject to} \quad & \pi_i(\gamma) \geq 0, \quad \forall\, i \in \{0, 1, \ldots, N\}; \qquad \sum_{i=0}^{N} \pi_i(\gamma) = 1. \end{aligned} \tag{2.10}$$

*(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.,*

$$\lambda_i + \Lambda \int_0^\infty \pi_i^*(\gamma)\, dF(\gamma) = \mu_i(Q^*). \tag{2.11}$$

Assuming that queue lengths are constant and given by $Q^*$, the expected delay on each exchange $i$ is given by $Q_i^*/\mu_i(Q^*)$. The individual rationality condition (i) ensures that limit orders are routed in a way that is consistent with (2.2). The flow balance condition, (ii), ensures that inflows and outflows at each exchange are balanced and that the queue length vector $Q^*$ remains stationary. Definition 4 is consistent[7] with the informal system of equations (2.8) since $\chi_i(Q^*) = \int_0^\infty \pi_i^*(\gamma)\, dF(\gamma)$.

### 2.3.2. State Space Collapse

Given a vector of queue lengths $Q$, define the *workload* to be the scaled sum of queue lengths given by $W \triangleq \sum_{i=1}^N \beta_i Q_i$. The workload captures the aggregate market depth across all exchanges, weighted by the attractiveness of each exchange. Orders queued at attractive exchanges (high $\beta_i$, typically corresponding to low $\tilde{r}_i$) are weighted more since these orders have greater priority to get filled first, and, therefore, more greatly impact the delays experienced by arriving limit orders at all exchanges. In fact, from (2.1) and (2.4), the expected delay on exchange $i$ is given by

$$\mathsf{ED}_i = \frac{W}{\mu\beta_i}. \tag{2.12}$$

That is, the 1-dimensional workload is sufficient to determine delays at every exchange. Theorem 1 below establishes something stronger: in equilibrium, the queue length vector $Q^*$,

[7]Strictly speaking, the informal definition (2.8) may not deal properly with situations where agents are indifferent between multiple routing decisions, while the formal Definition 4 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.

which is the state of the $N$-dimensional system can be inferred from the equilibrium workload $W^*$. This is a notion of *state space collapse*.

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

*(i)* $\pi^*$ *is an optimal solution for*

$$\begin{aligned} \underset{\pi}{\text{maximize}} \quad & \int_0^\infty \left\{ \pi_0(\gamma)\, \gamma \tilde{r}_0 + \sum_{i=1}^N \pi_i(\gamma) \left( \gamma \tilde{r}_i - \frac{W^*}{\mu \beta_i} \right) \right\} dF(\gamma) \\ \text{subject to} \quad & \pi_i(\gamma) \geq 0, \quad \forall\ i \in \{0, 1, \ldots, N\}, \quad \forall\ \gamma \geq 0, \\ & \sum_{i=0}^N \pi_i(\gamma) = 1, \quad \forall\ \gamma \geq 0. \end{aligned} \tag{2.13}$$

*(ii)* $\pi^*$ *satisfies*

$$\sum_{i=1}^N \left( \lambda_i + \Lambda \int_0^\infty \pi_i^*(\gamma)\, dF(\gamma) \right) = \mu. \tag{2.14}$$

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

$$Q_i^* \triangleq \left( \lambda_i + \Lambda \int_0^\infty \pi_i^*(\gamma)\, dF(\gamma) \right) \frac{W^*}{\mu \beta_i}. \tag{2.15}$$

*Conversely, if* $(\pi^*, Q^*)$ *is an equilibrium, define* $W^* \triangleq \beta^\top Q^*$. *Then,* $(\pi^*, W^*)$ *satisfy (i)–(ii).*

**Proof**. Suppose that $(\pi^*, W^*)$ satisfy (i)–(ii). For $Q^*$ given by (2.15), 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,

$$\frac{W^*}{\mu \beta_i} = \frac{\beta^\top Q^*}{\mu \beta_i} = \frac{Q_i^*}{\mu_i(Q^*)}. \tag{2.16}$$

Combining this with the fact that optimization problem in (i) is separable with respect to $\gamma$ (i.e., it can be optimized over each $\pi(\gamma)$ separately), it is clear that $(\pi^*, Q^*)$ satisfies the individual rationality condition (2.10). Further, rewriting (2.15),

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

Thus, $(\pi^*, Q^*)$ satisfies flow balance condition (2.11), and $(\pi^*, Q^*)$ is an equilibrium.

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

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

Given that $(\pi^*, Q^*)$ satisfies (2.10), this implies that $(\pi^*, W^*)$ satisfy (i). Further, if we sum up all $N$ equations in (2.11), it is clear that $(\pi^*, W^*)$ satisfy (ii). ■

Condition (i) of Theorem 1 implies individual rationality when faced with delays implied by the workload $W^*$, cf. (2.10) and (2.12). Condition (ii), is a market-wide flow balance equation. Given a pair $(\pi^*, W^*)$ satisfying (i) and (ii), $Q^*$ is determined as a function of workload $W^*$ through the *lifting map* (2.15) that distributes the workload across exchanges in a way that takes into account rebates, delays, and investor heterogeneity through the distribution $F(\cdot)$ of the trade-off coefficient $\gamma$. The lifting map corresponds to Little's Law: each queue length is equal to the corresponding aggregate arrival rate (dedicated and optimized) times the equilibrium expected delay.

### 2.3.3. Equilibrium Characterization

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

sufficient for the existence of an equilibrium:

**Assumption 1.** *Assume that*

*(i) The cumulative distribution function $F(\cdot)$ over the price-delay trade-off coefficients $\gamma$ is non-atomic with a continuous and strictly positive density on the non-negative reals.*

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

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

The dedicated flow $\sum_{i=1}^{N} \lambda_i$ is not delay sensitive. Condition (ii) ensures that the queueing system is stable ($\sum_{i=1}^{N} \lambda_i < \mu$) and leads to a non-trivial equilibrium where queue lengths are non-zero ($\mu < \Lambda + \sum_{i=1}^{N} \lambda_i$). Condition (iii) says that if delays are zero, then the effective rebate of a limit order is always preferable to the cost of crossing the spread and paying a fee to trade with a market order, $\tilde{r}_0$. Returning to condition (ii), given that $\mu < \Lambda + \sum_{i=1}^{N} \lambda_i$, one would expect non-zero queue lengths to build up in the system to discourage some optimizing investors from placing a limit order and instead trade with a market order. Intuitively, one expects this to be the most impatient investors, i.e., those of type $\gamma \leq \gamma_0$, for some $\gamma_0$, chosen to satisfy (2.14), i.e.,

$$\Lambda(1 - F(\gamma_0)) + \sum_{i=1}^{N} \lambda_i = \mu. \tag{2.17}$$

Under conditions (i)–(ii) of Assumption 1, $\gamma_0$ satisfying (2.17) is uniquely determined by

$$\gamma_0 = F^{-1}\left(1 - \frac{\mu - \sum_{i=1}^{N} \lambda_i}{\Lambda}\right). \tag{2.18}$$

In order for all types $\gamma \leq \gamma_0$ not to submit limit orders, the routing criterion (2.2) requires that

$$\max_{i \neq 0} \ \gamma(\tilde{r}_i - \tilde{r}_0) - \frac{W^*}{\mu \beta_i} \leq 0, \tag{2.19}$$

for all $\gamma \leq \gamma_0$. Under Assumption 1(iii), the left side of (2.19) is increasing in $\gamma$. Hence, (2.19) is satisfied if we ensure that type $\gamma_0$ investors are indifferent between market orders and limit orders.

**Lemma 1.** *Under Assumption 1, suppose that* $(\pi^*, W^*)$ *is an equilibrium and define* $\gamma_0$ *by* (2.18)*. Then,*

$$\max_{i \neq 0} \ \gamma_0(\tilde{r}_i - \tilde{r}_0) - \frac{W^*}{\mu \beta_i} = 0. \tag{2.20}$$

*Further, suppose that for a given* $W^*$*,* (2.20) *holds, and for each exchange* $i$*, define*

$$\kappa_i \triangleq \beta_i(\tilde{r}_i - \tilde{r}_0). \tag{2.21}$$

*Then, an exchange* $i$ *achieves the maximum in* (2.20) *if and only if* $i \in \operatorname{argmax}_{j \neq 0} \ \kappa_j$*.*

(The proof of the Lemma is provided in Appendix.) The quantity $\kappa_i$ is related to the desirability of exchange $i$ from the perspective of a limit order investor; $\kappa_i$ is high when $\beta_i$ is high (resulting in low delay) or when $\tilde{r}_i$ is high (resulting in a high rebate). Lemma 1 suggests that maximizing $\kappa_i$ characterizes the behavior of type $\gamma_0$ (the marginal) investors that are indifferent between choosing between a market order and a limit order. We refer to exchanges that achieve this maximum as marginal exchanges. Thus, given a marginal exchange $\bar{i} \in \operatorname{argmax}_{j \neq 0} \kappa_j$, according to Lemma 1,

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

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

**Theorem 2** (Equilibrium Characterization)**.** *Under Assumption 1, define* $\gamma_0$ *by* (2.18)*. Suppose*

*that the pair* $(\pi^*, W^*) \in \mathcal{P} \times \mathbb{R}_+$ *satisfy*

$$W^* \triangleq \gamma_0 \mu \max_{i \neq 0} \ \kappa_i, \tag{2.22}$$

*and*

$$\begin{aligned} \pi_0^*(\gamma) &= 1, \qquad \textit{for all } \gamma < \gamma_0, \\ \pi_i^*(\gamma_0) &= 0, \qquad \textit{for all } i \notin \mathcal{A}^*(\gamma_0) \cup \{0\}, \\ \pi_i^*(\gamma) &= 0, \qquad \textit{for all } \gamma > \gamma_0,\ i \notin \mathcal{A}^*(\gamma), \end{aligned} \tag{2.23}$$

*where* $\mathcal{A}^*(\gamma) \triangleq \operatorname{argmax}_{i \neq 0} \ \gamma \tilde{r}_i - W^*/\mu\beta_i$. *Then,* $(\pi^*, W^*)$ *is an equilibrium, i.e., satisfies* (2.13)-(2.14).

*Conversely, suppose that* $(\pi^*, W^*) \in \mathcal{P} \times \mathbb{R}_+$ *is an equilibrium, i.e., satisfies* (2.13)-(2.14). *Then,* $W^*$ *must satisfy* (2.22) *and* $\pi^*$ *must satisfy* (2.23)*, except possibly for* $\gamma$ *in a set of* $F$*-measure zero.*

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

### 2.3.4. Convergence of Fluid Model to Equilibrium Configuration

We first establish uniqueness of the equilibrium queue length vector $Q^*$ in the next Theorem (its proof is available in the Appendix), under the following mild assumption:

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

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

Next we establish that the fluid model queue length vector $Q(t)$ converges to the unique equilibrium vector $Q^*$ as $t \to \infty$. As in Section 2.2.3, define $\mathcal{G}_i\big(W(t)\big) \subset \mathbb{R}_+$ to be the set of optimizing limit order investor types $\gamma$ that would prefer exchange $i$ given a workload level[8] of $W(t)$, i.e., the set of all $\gamma \geq 0$ with

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

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

$$\chi_i(W(t)) \triangleq \int_{\mathcal{G}_i\left(W(t)\right)} dF(\gamma). \tag{2.24}$$

Under Assumptions 1 and 2, (2.24) can be rewritten as

$$\chi_i(W) = \begin{cases} F\left(\dfrac{W\Gamma_i^+}{\mu}\right) - F\left(\dfrac{W\Gamma_i^-}{\mu}\right) & \text{if } \Gamma_i^+ \geq \Gamma_i^-, \\ 0 & \text{otherwise,} \end{cases} \tag{2.25}$$

where the constants $\Gamma_i^+, \Gamma_i^-$ are defined by

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

[8]Note that in Section 2.2.3, $\mathcal{G}_i(\cdot)$ and $\chi_i(\cdot)$ were defined to be functions of the vector of all queue lengths. However, since they depend on the queue length of each exchange only through the expected delay and therefore the workload, we will abuse notation and define these as functions of workload here.

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

$$\sum_{i=1}^{N} \beta_i \frac{d\chi_i(W)}{dW} < 0. \tag{2.26}$$

Assumption 3 is essentially a local stability drift condition[9] that is easy to verify, and takes the form of a tail condition on $F(\cdot)$. Specifically, using (2.25), we have that:

$$\sum_{i=1}^{N} \beta_i \frac{d\chi_i(W)}{dW} = \sum_{i=1}^{N} \left( \frac{\Gamma_i^+}{\mu} f\left( \frac{W\Gamma_i^+}{\mu} \right) - \frac{\Gamma_i^-}{\mu} f\left( \frac{W\Gamma_i^-}{\mu} \right) \right) \mathbb{I}_{\{\Gamma_i^+ \geq \Gamma_i^-\}}, \tag{2.27}$$

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

$$t\Gamma_i^+ f(t\Gamma_i^+) < t\Gamma_i^- f(t\Gamma_i^-), \tag{2.28}$$

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

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

Examining (2.28), it is clear that elastic distributions will always satisfy Assumption 3. As an example, note that decreasing generalized failure rate distributions; see, e.g., Lariviere (2006), are included in the class of elastic distributions.

---

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

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

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

(The proof of Theorem 4 is available in the Appendix.) For $N > 2$, condition (2.26) is necessary for local stability of the equilibrium $Q^*$. We conjecture that, as for $N = 2$, Assumption 3 is, in fact, also a sufficient condition when $N > 2$.

The state-space collapse result and its functional form hinge on the formulation of the order routing models described in Sections 2.2.1 and 2.2.2. The primary drivers of the dimension reduction are: (a) the desirability to place an order at a given queue is decreasing in its anticipated delay, and (b) that the attractiveness of an exchange for an incoming marker order is increasing in its queue length. Both drivers seem plausible even under different models of order routing optimization logic on both sides of the market, and would typically lead to some form of state space collapse: long queues would discourage new orders from joining while attracting more service completions, thus reducing queue size; small queues would attract more arrivals but fewer service completions, thus increasing queue size. For example, the same rationale holds if we replace the market order routing model (2.4) with a model of the form $\mu_i(Q) \triangleq M_i + f_i(Q)$, for each exchange $i$. Here, each $M_i \geq 0$ represents "dedicated" market order flow to exchange $i$ that does not react to the state of the system, while the $f_i(Q)$ term captures optimized order flow. The detailed form of the equilibrium of such a system would not coincide with the one derived here, however, at a high level, one would

expect similar results under different modeling assumptions that satisfy (a)–(b).

### 2.3.5. Pointwise-Stationary-Fluid-Model

In the preceding analysis we have assumed that the event rate parameters $(\Lambda, \lambda, \mu)$ are constant over time. In such a setting, our results show that the queue length configuration converges into an equilibrium state, which is a function of the underlying rate parameters. A quick look at the data will show that the underlying model parameters exhibit significant variation over time. This could result from the superposition of different institutional (parent) orders that enter and exit the market over time, or switch execution strategies and aggression. Such changes in the underlying order flow will translate into changes in the arrival rates of limit and market orders into the order book, which, in turn, will affect the resulting equilibrium state. Table 2.1 studies this issues for a sample of liquid stocks that comprise the Dow Jones Industrial Average (DJIA). Specifically, for each $\tau$-minute interval, we compute the trading rate $\mu_t$ and then test how often was $\mu_{t+1}$ within the confidence interval $\mu_t \pm k\sigma_t$ for $k = 2, 3$, and assuming that the arrival rate of market orders is Poisson with rate $\mu_t$ we set $\sigma_t = \sqrt{\mu_t}$. The data suggests that the trading rate exhibits significant fluctuations after 5 or 10 minutes out, but that is fairly consistent over the span of 1 to 3 minutes. Similar findings apply for the rates of limit order submissions. To interpret these results it is instructive to recognize that the average queueing delays across liquid stocks (we will study later on the 30 stocks that comprise the DJIA) is of the order of 1 minute, as illustrated by the summary statistics in Table 2.2.Queueing delays are of the order of magnitude of the queueing transients, so the parameters may be assumed to be constant in the time scale of the fluid transients but exhibit fluctuations over longer time horizons.

Based on the above one would suggest an analysis over two time scales. The fast time scale, which is is the nominal clock of the system at which orders arrive at the market, and

| | % obs. in $\pm 2\sigma_t$ | % obs. in $\pm 3\sigma_t$ | % obs. outside $\pm 3\sigma_t$ |
|---|---|---|---|
| 1 min | 63.33% | 79.23% | 20.77% |
| 3 min | 32.56% | 50.39% | 49.61% |
| 5 min | 27.27% | 35.06% | 64.94% |
| 10 min | 13.16% | 31.58% | 68.42% |

**Table 2.1**

| Mean | Std. dev. | 1st quantile | 3rd quantile |
|---|---|---|---|
| 1.0884 | 0.8223 | 0.5510 | 1.3756 |

**Table 2.2:** Summary statistics of expected delay across the 30 names that comprise the DJIA and across the 6 major exchanges listed in Table 2.4. Expected delay measurement is outlined as (2.29) in Section 2.4.1.

over which model parameters appear to be constant. And a slower time scale over which parameters fluctuate stochastically. Consider a market where event rates are proportional to some constant $n$ (e.g., the number of trades per minute). Moreover assume that

$$\mu^n(t) = n\mu(t/a_n), \quad \lambda^n(t) = n\lambda(t/a_n) \quad \text{and} \quad \Lambda^n(t) = n\Lambda(t/a_n)$$

where $a_n \to \infty$ and $a_n/n \to 0$ as $n \to \infty$. So, rates grow large, and fluctuate according to $\mu(\cdot), \lambda(\cdot), \Lambda(\cdot)$, but these fluctuations occur in slower time scale. These rate processes are themselves stochastic with sufficient continuity to allow for a tractable analysis. Let $Q^n(t)$ denote the queue length process associated with the market of speed $n$. An $n$ grows large, an argument based on the results by Kurtz (1977/78) would establish that the queue length process, when rescaled by $n$, $Q^n(t)/n$ converges to a deterministic limit that satisfies the fluid model equations we have analyzed thus far, and over which the model primitives $(\mu, \lambda, \Lambda)$ appear constant. If one where to study the same model on the slower time scale over which rates fluctuate, by studying the queue length process over stretches of time proportional to

$a_n$ one would see that the queue length process seems to always be equal to the equilibrium state that would correspond to the triple $(\mu(t), \lambda(t), \Lambda(t))$. The resulting model is a so-called "Pointwise-Stationary-Fluid-Model," which is stochastic but whose queue length configuration appears (on the slower time scale) to be always in its equilibrium state. In that sense, the SSC property established earlier can be shown to be a "pathwise" property as opposed to a point in time result. (We will empirically explore this pathwise result in the next section.) We will not fully flush out the requisite analysis for establishing the above assertions, as this would be lengthy and would not add to the emphasis of this chapter. Standard machinery for establishing such results either exploit the work by Bramson (1998) or Bassamboo et al. (2006). Our model seems to satisfy the key requirements that one would need to derive the PSFM and as a result the sample path version of the SSC property, but we will not pursue this in this chapter. An example can also be found in Besbes and Maglaras (2009).

## 2.4. Empirical Results

Motivated by our analysis and the fact that for liquid securities the markets experience high volumes of flow per unit time, one would expect the market to behave as if it is near its equilibrium state most of the time, which would manifest itself as a strong coupling between the quote depths and dynamics of competing exchanges. More precisely, the expected delay trajectories across exchanges and over time should exhibit strong linear relationships, and behave like a lower dimensional process. Moreover, the workload process (a measure of weighted aggregate depth) should offer accurate estimates of delays and queue depths at different exchanges, as stated in (2.12). The precise form of these predictions is, of course, predicated on the structure of (2.2) and (2.4)–(2.5) and the deterministic and stationary nature of the model. The sample of market data analyzed below captures more complex and

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

### 2.4.1. Overview of the Data Set

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

We consider the 30 component stocks of the Dow Jones Industrial Average over the 21 trading days in the month of September 2011. A list of the stocks and some basic descriptive statistics are given in Table 2.3.

We restrict attention to the $N = 6$ most liquid U.S. equity exchanges: NASDAQ, NYSE,[10] ARCA, EDGX, BATS, and EDGA. Smaller, regional exchanges were excluded as they account for a small fraction of the composite daily volume and are often not quoting at the NBBO level. The associated fees and rebates during the observation period of September 2011 are given in Table 2.4.

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

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

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

**Table 2.3:** Descriptive statistics for the 30 stocks over the 21 trading days of September 2011. The average bid-ask spread is a time average computed from our TAQ data set. The volatility is an average of daily volatilities over Sept 2011. All the other statistics were retrieved from Yahoo Finance.

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

**Table 2.4:** Rebates and fees of the 6 major U.S. stock exchanges during the September 2011 period, per share traded. †Rebates on NASDAQ and ARCA are subject to "tiering": higher rebates than the ones quoted may be available to traders that contribute significant volume to the respective exchange.

In contrast, the arrival rates $(\lambda, \Lambda, \mu)$ are time-varying. We will estimate these rates for each stock by averaging the event activity over one hour time intervals between 9:45am and 3:45pm (i.e., excluding the opening 15 minutes and the closing 15 minutes).[11] This yields $T = 126$ time slots over the 21 day horizon of our data set. For each time slot $t$, exchange $i$, stock $j$, and side $s \in \{\mathsf{BID}, \mathsf{ASK}\}$, we estimated the corresponding queue length as the average number of shares available at the NBBO, denote this by $Q_i^{(s,j)}(t)$. Similarly, denote by $\mu_i^{(s,j)}(t)$ the arrival rate of market orders to side $s$ on exchange $i$ for security $j$, in time slot $t$. The rates $\mu_i^{s,j}(t)$ are estimated by classifying trades to be bid or ask side of the market, by matching trade time stamps with the prevailing quote at the same time, i.e., using a zero time shift in the context of the well known Lee-Ready algorithm. Given these parameters, we compute the following measure of expected delay

$$\mathsf{ED}_i^{(s,j)}(t) \triangleq \frac{Q_i^{(s,j)}(t)}{\mu_i^{(s,j)}(t)}. \tag{2.29}$$

The above expression disregards the effect of order cancellations from the bid and ask queues,

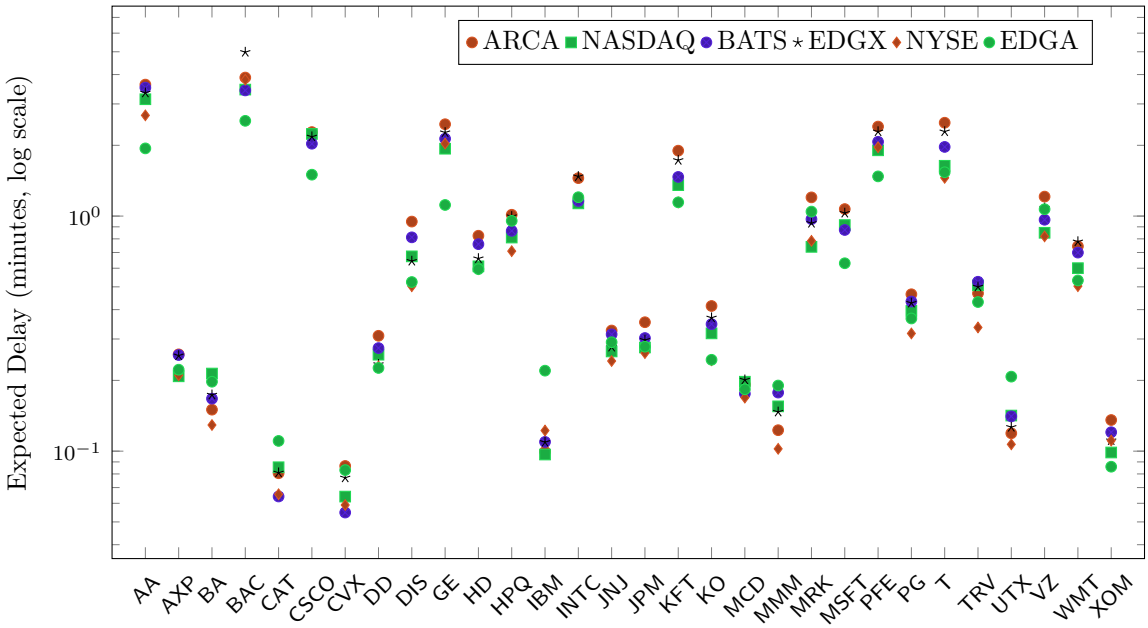
[11] The time intervals should be sufficiently long so as to get reliable estimates of the event rates, and also long when compared to the event inter-arrival times, so that one could expect that the transient dynamics of the market due to changes in these rates settle down during these time intervals.

as well as the non-infinitesimal nature of the order flow. It serves as a practical proxy for expected delay that is commonly used in trading systems. For each stock and each exchange, Figure 2.2(a) shows the expected delay, averaged across time slots and the bid and ask sides of the market. Delays range from 5 seconds to about 5 minutes across the 30 stocks we studied, and we observe 2x to 3x variation in the delay estimates at different exchanges for the same security. Similarly, for each stock and each exchange, Figure 2.2(b) shows the average queue lengths, or, the number of shares available at the NBBO, averaged across time slots and the bid and ask sides of the market. Queue lengths range from 10 to 100,000 shares across securities, and exhibit about a 10x variation in the queue sizes across exchanges for the same security. Deeper queues correspond to longer delays.

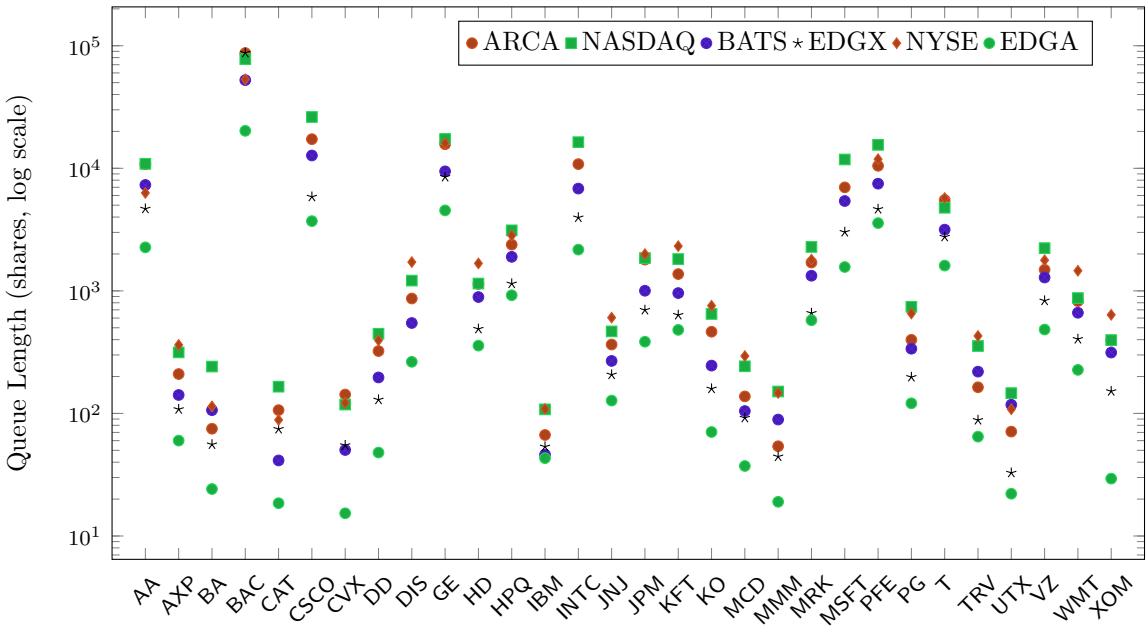
**Principle component analysis (PCA).** The state space collapse result of our model predicts that delays are coupled across exchanges and are restricted to a 1-dimensional subspace. Define the empirically observed expected delay vector trajectories

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

where $\mathsf{ED}^{(s,j)}(t)$ was estimated in (2.29) and the trajectories consider all one hour time slots in the 21 days of our observation period. A natural way to test the effective dimensionality of this vector of trajectories is via PCA by examining the number of principle components necessary to explain the variability of the expected delay trajectories across exchanges and over time. The output of the PCA analysis is summarized in Table 2.5: the first principle component explains around 80% of the variability of the expected delays across exchanges, and the first two principle components explain about 90%. This is consistent with the hypothesis of low effective dimension. In contrast, when we conduct PCA for the vector trajectories of observed queue lengths $\left\{Q^{(s,j)}(t) \ : \ t = 1, \ldots, T; \ s = \mathsf{BID}, \mathsf{ASK}\right\}$, we find relatively weaker evidence




(a) Average expected delay across stocks and exchanges.




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

**Figure 2.2:** Averages of hourly estimates of the expected delays and queue lengths for the Dow 30 stocks on the 6 exchanges during September 2011. Results are averaged over the bid and ask sides of the market for each stock. Queues do not include estimates of hidden liquidity at each of the exchanges.

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

**Table 2.5:** Results of PCA: how much variance in the data can the first two principle components explain.

for a low effective dimensionality. In this test, the first principle component explains about 65% of the variability of the queue lengths across exchanges, and the first two principle components explain less than 80%. A detailed report of the results can be found in Table A.1 in the Appendix.

Intuitively, in the high flow environment of our observation universe, i.e., where $\Lambda$ and $\mu$ are large, queue length deviations from the equilibrium configuration would be quickly erased by optimized arriving limit and market orders. The equilibrium state itself changes over time as the rates of events change, but the coupling across exchanges remains strong, and persists even if we shorten the time period over which market statistics are averaged from 1 hour down to 15 minutes.[12]

[12]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.

### 2.4.2. Estimation of the Market Order Routing Model

Define $\mu_i^{(s,j)}(t)$ to be the total arrival rate of market orders for security $j$ and side $s \in \{\mathsf{BID}, \mathsf{ASK}\}$ in time slot $t$ directed to exchange $i$, and let $\mu^{(s,j)}(t)$ be the total arrival rate across all exchanges for $(s,j)$ in time $t$. The attraction model of Section 2.2.2 for market orders suggests the relationship

$$\mu_i^{(s,j)}(t) = \mu^{(s,j)}(t) \ \frac{\beta_i^{(j)} Q_i^{(s,j)}}{\sum_{i'=1}^{N} \beta_{i'}^{(j)} Q_{i'}^{(s,j)}}, \tag{2.30}$$

where $\beta_i^{(j)}$ is the attraction coefficient for security $j$ on exchange $i$. Note that our market order routing model is invariant to scaling of the attraction coefficients, hence we normalize so that the attraction coefficient for each stock on its listing exchange is 1. Given that $\{\mu_i^{(s,j)}(t)\}$, $\{\mu^{(s,j)}(t)\}$, and $\{Q_i^{(s,j)}(t)\}$ are observable, we estimated the $\beta_i^{(j)}$'s using a nonlinear regression on (2.30). The results are given in Table 2.6. Note that all attraction coefficient estimates are statistically significant.

### 2.4.3. Empirical Evidence of State Space Collapse

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

**Verification of linear dependence of expected delays via regression analysis.** Define $W^{(s,j)}(t)$ to be the workload for side $s$ of security $j$ in time slot $t$, i.e.,

$$W^{(s,j)}(t) \triangleq \sum_{i=1}^{N} \beta_i^{(j)} Q_i^{(s,j)}(t), \tag{2.31}$$

| | 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 2.6:** Estimates of the attraction coefficients $\beta_i$ from nonlinear regression. Note that the attraction coefficient of the listing exchange is normalized to be 1.

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

$$\mathsf{ED}^{(s,j)}(t) = \frac{W^{(s,j)}(t)}{\mu^{(s,j)}(t)} \left( \frac{1}{\beta_1^{(j)}}, \ldots, \frac{1}{\beta_N^{(j)}} \right). \tag{2.32}$$

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

$$\mathsf{ED}_i^{(s,j)}(t) = \frac{\beta_{i'}^{(j)}}{\beta_i^{(j)}} \, \mathsf{ED}_{i'}^{(s,j)}(t), \tag{2.33}$$

for each time slot $t$. To test this prediction, we perform a linear regression of the left side of (2.33), which is the expected delay of that security on a particular exchange, as a function of the right side of (2.33), which is the expected delay on a benchmark exchange (ARCA) rescaled by the ratio of the attraction coefficients of the two exchanges. The regression is performed using the expected delay measurements outlined in (2.29), i.e., by dividing the average observed queue size in each exchange with its respective observed rate of trading, for all time slots, both sides of the market, and all the 30 component stocks of the Dow Jones Industrial Average. In addition, for the cross-sectional regression analysis, for each security, we normalize the expected delay measurements by dividing them by the median expected delay of that security on a benchmark exchange (ARCA) across all time slots and both sides of the market. That is, we use the following measure of expected delays in the linear regressions

$$\overline{\mathsf{ED}}_i^{s,j}(t) \triangleq \frac{\mathsf{ED}_i^{s,j}(t)}{\underset{\tau=1,\ldots,T;\ s=\mathsf{BID},\mathsf{ASK}}{\text{median}} \left( \mathsf{ED}_{\text{ARCA}}^{(s,j)}(\tau) \right)}, \tag{2.34}$$

where $\mathsf{ED}_i^{(s,j)}(t)$ was estimated in (2.29).

The results of these regressions are summarized in Table 2.7. The $R^2$ varies between 52% and 70% across the five exchanges. All of the regressions are statistically significant

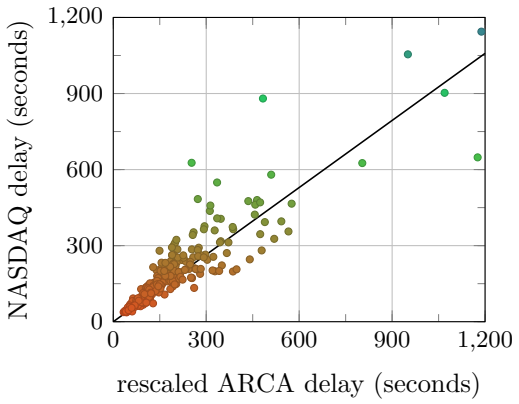
| | Dependent Variable: $\overline{\mathsf{ED}}_{\text{exchange}}$ | | | | |
|---|---|---|---|---|---|
| | NASDAQ OMX | BATS | DirectEdge X | NYSE | DirectEdge A |
| Intercept | $0.27^{***}$ | $0.28^{***}$ | $0.24^{***}$ | $0.28^{***}$ | $0.36^{***}$ |
| | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| Rescaled $\overline{\mathsf{ED}}_{\text{ARCA}}$ | $0.70^{***}$ | $0.72^{***}$ | $0.72^{***}$ | $0.63^{***}$ | $0.60^{***}$ |
| | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| $R^2$ | 70% | 70% | 52% | 60% | 52% |
| Note: | | | | | $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$ |

**Table 2.7:** Linear regressions of the normalized expected delay on a particular exchange, versus that of the benchmark exchange (ARCA) rescaled by the ratio of the attraction coefficients of the two exchanges.

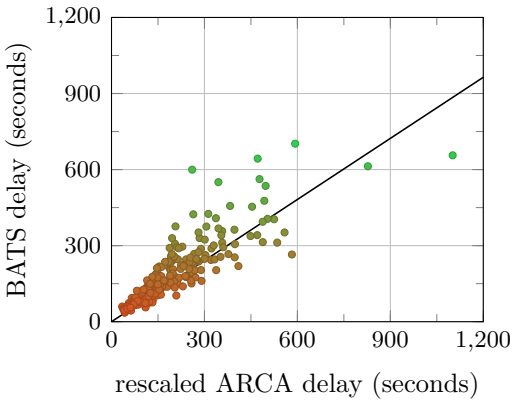
and we are able to reject the null hypothesis that the delay on a particular exchange has a zero regression coefficient relative to the rescaled delay on ARCA. These results statistically verify the linear dependence of delays across different exchanges suggested by (2.33). Note that (2.33) further predicts that the regression should have a zero intercept and the slope of the rescaled $\mathsf{ED}_{\text{ARCA}}$ term should be 1. These are not born in the regressions — the intercept is statistically different from 0 and the slope is statistically different from 1. Nevertheless, the intercept and slope are, respectively, quite close to 0 and 1. This is remarkable given the stylized nature of the routing model of Section 2.4.2 and the noise in the extensive market data sample.

While the regressions in Table 2.7 were performed cross-sectionally across all securities, similar results hold if the analysis is performed on a security by security basis. Figure 2.3 depicts the delay relationships in the case of Bank of America. It illustrates the strong linear relationship across all exchanges over time and across significant variations in prevailing market conditions; the latter is manifested in the roughly two orders of magnitude variation in estimated expected delays.

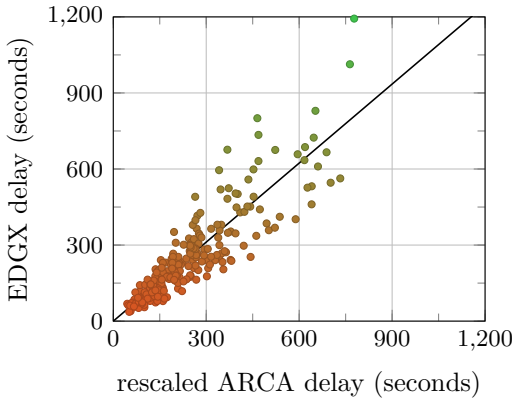
A competing hypothesis is that queue lengths across exchanges are linearly related, that



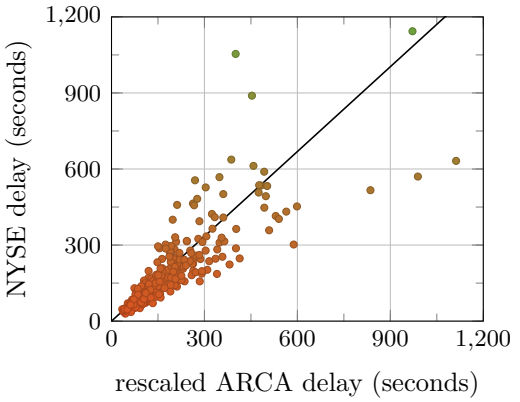

(a) slope $= 0.88$, intercept $= 6 \times 10^{-3}$, $R^2 = 84\%$



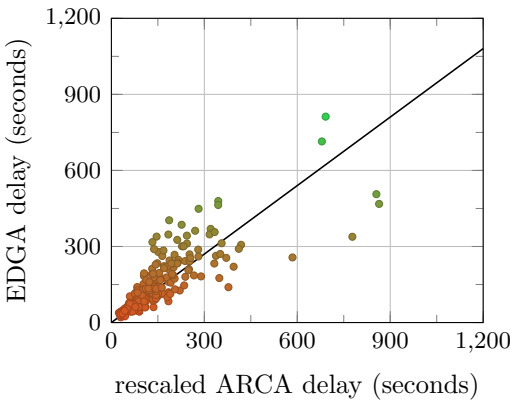

(b) slope $= 0.80$, intercept $= 9 \times 10^{-3}$, $R^2 = 79\%$




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




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




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

**Figure 2.3:** Scatter plots of the expected delay for Bank of America (BAC) on each exchange, versus the delay on ARCA rescaled by the ratio of the attraction coefficients of the two exchanges. The black lines correspond to linear regressions with intercept.

is, for each security $j$, exchanges $i$, $i'$, and market side $s$,

$$Q_i^{s,j}(t) = c_{ii'} Q_{i'}^{s,j}(t), \tag{2.35}$$

for each time slot $t$. The following test explores such an alternative hypothesis. According to (2.35), predicated on queue length estimates obtained in 2.4.1, i.e., $Q_i^{s,j}(t)$ as the average number of shares available at the NBBO for time slot $t$, exchange $i$, stock $j$, and side $s \in \{\mathsf{BID}, \mathsf{ASK}\}$, we perform a cross-sectional linear regression of the queue length of each security on a particular exchange, as a function of that on a benchmark exchange (ARCA). As before, we normalize the queue lengths by dividing them by the median queue length of that security on a benchmark exchange (ARCA) across all time slots and both sides of the market, i.e., we use $\overline{Q}_i^{s,j}(t) \triangleq Q_i^{s,j}(t)/\text{median}_{\tau=1,\ldots,T;\ s=\mathsf{BID},\mathsf{ASK}} \left(Q_{\text{ARCA}}^{(s,j)}(\tau)\right)$ as the queue length measure in regression. The results are provided in table 2.8. We observe that the $R^2$ is significantly lower than that in the previous table, varying only between 13% and 26%.

| | Dependent Variable: $\overline{Q}_{\text{exchange}}$ | | | | |
|---|---|---|---|---|---|
| | NASDAQ OMX | BATS | DirectEdge X | NYSE | DirectEdge A |
| Intercept | $0.84^{***}$ | $0.39^{***}$ | $0.25^{***}$ | $0.57^{***}$ | $0.05^{***}$ |
| | (0.02) | (0.01) | (0.01) | (0.02) | (0.01) |
| $\overline{Q}_{\text{ARCA}}$ | $0.74^{***}$ | $0.45^{***}$ | $0.29^{***}$ | $0.96^{***}$ | $0.24^{***}$ |
| | (0.02) | (0.01) | (0.01) | (0.02) | (0.00) |
| $R^2$ | 19% | 20% | 13% | 26% | 26% |
| Note: | | | | | $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$ |

**Table 2.8:** Linear regressions of the normalized queue length on a particular exchange versus that of the benchmark exchange (ARCA).

**Residual analysis and accuracy of delay estimates based on the aggregate workload.** The SSC result culminated in relationship (2.32) that makes expected delay predictions in each exchange based on the 1-dimensional aggregated workload process. Specifically, given the market model coefficients $\beta_i^{(j)}$ and a measurement of the queue sizes at the various

exchanges, $Q_i^{(s,j)}(t)$, one can compute the workload via (2.31), and then construct estimates for the expected delays at the various exchanges via (2.32). We denote the resulting delay estimates by $\widehat{\mathsf{ED}}^{(s,j)}(t)$, where the $\hat{}$ notation denotes in this context the estimate obtained via the one-dimensional workload process, as opposed to measuring the actual expected delay $\mathsf{ED}^{(s,j)}(t)$ via (2.29). This prediction can be tested again through a set of linear regressions between the workload delay estimate and the delay estimate that uses information about the state of the exchange (queue length and trading rate). All these regressions are statistically significant and are accompanied with high $R^2$ values. We do not report on these results, instead we pursue a more detailed analysis of the residuals, i.e., the errors between the workload and exchange-specific delay estimates, $\mathsf{ED}^{(s,j)}(t) - \widehat{\mathsf{ED}}^{(s,j)}(t)$. We define the quantity

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

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 (2.32), relative to the variability of the underlying expected delays. By its definition, when $R_*^2$ is close to 1, most of the variability of expected delays is explained by the relationship (2.32). Numerical results for $R_*^2$ across securities are given in Table 2.9. Typical values for $R_*^2$ are around 80%, highlighting the predictive power of the one-dimensional workload model as a means of capturing the state of the decentralized fragmented market.

Our analysis showed that optimized order routing couples the exchange dynamics in terms of their delay estimates as opposed to their queue depths. As mentioned earlier in discussing Figure 2.2(b), queue lengths across exchanges exhibit significantly more variation than their corresponding delays. One could repeat the above analysis, for example, starting

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

**Table 2.9:** The measure of performance $R^2_*$, which given the reduction of variability in expected delays explained by the workload relationship (2.32).

from trying to see whether the queue length processes live on a lower dimensional manifold, similarly to what we observed in studying the respective delay estimates. Not surprisingly, and as suggested through our analysis and the above comments, the PCA of the queue length trajectories yields weaker results, and similarly all of the subsequent tests lead to noticeably lower quality of fit. Our model suggests two explanations: (a) the limit order routing logic seems to rely on delay estimates as opposed to queue lengths; and, (b) the model capturing the routing of market orders is itself nonlinear. Both (a) and (b) hinge on some of our modeling assumptions that build on insight from practical smart order router optimization logic, where indeed limit order placement decisions depend crucially on delays or fill probabilities, and market order routing follow variants of fee minimization arguments that depend nonlinearly on the displayed quantities.

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

### 2.4.4. The Effects of Fee Change: evidence from the NASDAQ fee experiment

We finally illustrate how the predictions of our model match up to the observations made possible by the natural experiment of the NASDAQ exchange that made a significant reduction of its fee and rebate schedule for a subset of 14 stocks between February and May of 2015. [13] NASDAQ lowered the fees charged to liquidity takers from \$.0030/share to \$.0005/share, and correspondingly, lowered the rebates rewarded to liquidity providers from \$.0029/share to \$.0004/share.

To test the impact of this significant reduction in the make-take fee on NASDAQ, we analyze and compare trade and quote (TAQ) data of the 14 tested symbols in two separate time periods: the *pre-period* of 01/12/2015 - 01/30/2015 and the *post-period* of 02/09/2015 - 02/27/2015, that is, 3 weeks before and after the initiation of the program at the beginning of February 2015, respectively. Table 2.10 contains the fees charged on the 6 major exchanges in the tested periods, during which only that of NASDAQ has been modified.

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

**Table 2.10:** Fees of the 6 major U.S. stock exchanges, per share traded, in January-February 2015 around the time of the NASDAQ access fee experiment. Note that the fees here are different from previous figures in table 2.4 because they are in different time periods.

A change in the per share fee and rebate will affect the attractiveness of the exchange

[13] The test symbols are: AAL, BAC, FEYE, GE, GPRO, GRPN, KMI, MU, RAD, RIG, S, SIRI, TWTR, ZNGA.

for traders placing limit orders and traders sending aggressive market orders. There have been a few studies published thus far Hatheway (2015a,b) and Pearson (2015) on the some of these effects. Both studies were not predicated on a underlying model of order routing, and primarily focused on market share and depth comparisons, before and after the fee change. Our model of market and limit order routing yields some direct implications on market outcomes, which we explore in the sequel.

In the sequel, we will first estimate the exchange attraction coefficients before and after the fee change. We will propose a structural model for the attractiveness of each exchange that explicitly incorporates it's prevailing fee, from which one would expect that the attractiveness of Nasdaq increased after the fee reduction. We verify this prediction. We will then study the effect of the fee change in the routing of limit orders. In this case, our model predicts that the equilibrium expected delay for limit orders to get filled will decrease after the fee change. The empirical analysis will again verify this prediction.

Put together these observations suggest that the impact of the fee change is best understood through its structural impact to limit and market order routing policies, and complement the findings of the analyses reported in Hatheway (2015a,b); Pearson (2015). Our empirical findings will validate our model predictions.

**Attraction coefficient $\beta_{\mathbf{NASDAQ}}$.** The discussion in Section 2.2.2 had suggested that the attractiveness of an exchange for market orders is a decreasing function of its fee. Our preceding analysis focused on an observation period where fees were constant, which implied that the attractiveness coefficients of the exchanges were themselves constant throughout that time period. Nasdaq's fee experiment allows us to proceed with a more nuanced analysis, and examine what is the effect of the exchange fees on market order flow. We will postulate

the following structural model:

$$\mu_i^{(s,j)}(t) = \mu^{(s,j)}(t)\ \frac{e^{a_i^{(j)}+r_{i,t}\cdot b^{(j)}} Q_i^{(s,j)}(t)}{\sum_{k=1}^{N} e^{a_k^{(j)}+r_{k,t}\cdot b^{(j)}} Q_k^{(s,j)}(t)}. \tag{2.36}$$

That is, we are postulating the attractiveness coefficient $b_i^j$ take the form:

$$b_i^j = e^{a_i^{(j)}+r_{i,t}\cdot b^{(j)}}$$

We will estimate (2.36) using 3 weeks of data before and after the fee change. Our hypothesis is that the coefficients $b^j$ are negative, i.e., an higher fee makes an exchange less desirable, all other things being equal. The corresponding $\{\mu_i^{(s,j)}(t)\}$, $\{\mu^{(s,j)}(t)\}$, and $\{Q_i^{(s,j)}(t)\}$ are estimated from the two trade and quote data samples as outlined in Section 2.4.1.

We normalize the results so that $a_i$ of the benchmark exchange NYSE ARCA is 0. Finally, $b$ is estimated by using nonlinear regressions on (2.36), based on the combined sample for each security. Results are in Table 2.11. Indeed, the estimated $b$ coefficients are negative for all 14 tested securities, among which 12 are statistically significant. This agrees with our hypothesis.

On a related note we estimate the $\beta_{\text{NASDAQ}}$ coefficients before and after the fee change and compare; this estimation is non-parametric, specifically not predicated on (2.36), and estimates via nonlinear regressions the following:

$$\mu_i^{(s,j)}(t) = \mu^{(s,j)}(t)\ \frac{\beta_i^{(j)} Q_i^{(s,j)}}{\sum_{i'=1}^{N} \beta_{i'}^{(j)} Q_{i'}^{(s,j)}}, \tag{2.37}$$

based on the pre-period sample and the post-period sample, respectively. Our market order routing model is invariant to scaling of the attraction coefficients and in this section we normalize so that the attraction coefficient for each stock on the benchmark exchange NYSE

| | $b^{(j)}$ | $a^{(j)}_{\text{NASDAQ}}$ | $a^{(j)}_{\text{EDGX}}$ | $a^{(j)}_{\text{BATZ}}$ | $a^{(j)}_{\text{NYSE}}$ | $a^{(j)}_{\text{EDGA}}$ | 1-sided 95% test |
|---|---|---|---|---|---|---|---|
| AAL | -4.91546 | 0.1122 | -0.0179 | 0.26979 | | 0.09173 | NO |
| BAC | -179.49864 | 0.41989 | 0.19534 | 0.25584 | 0.12998 | 0.46171 | YES |
| FEYE | -98.64992 | -0.30427 | -0.20065 | 0.0999 | | -0.31724 | YES |
| GE | -93.50329 | 0.1524 | -0.03982 | 0.12865 | -0.0113 | 0.26895 | YES |
| GPRO | -50.15462 | -0.07134 | -0.20676 | 0.07204 | | -0.06651 | YES |
| GRPN | -94.41 | 0.1174 | -0.07473 | 0.2509 | | 0.0007071 | YES |
| KMI | -113.09557 | 0.08852 | 0.06439 | 0.13474 | -0.17634 | -0.02007 | YES |
| MU | -89.01952 | 0.06419 | -0.02114 | 0.13652 | | 0.18507 | YES |
| RAD | -138.45738 | -0.0452 | -0.21638 | 0.06565 | -0.29318 | -0.09038 | YES |
| RIG | -76.558 | -0.0187 | -0.01546 | 0.03923 | -0.14439 | 0.09803 | YES |
| S | -162.59365 | -0.21658 | -0.28509 | 0.03679 | -0.25726 | -0.37789 | YES |
| SIRI | -69.31381 | 0.12737 | -0.1322 | 0.17258 | | 0.33232 | YES |
| TWTR | -87.330238 | -0.141326 | -0.155197 | -0.009813 | -0.272932 | -0.227205 | YES |
| ZNGA | -17.66407 | 0.14102 | -0.27556 | 0.14014 | | 0.40745 | NO |

**Table 2.11:** Estimates of $b$ and $a_i$ in the attraction model (2.36) and hypothesis testing results on whether the coefficient $b$ is negative. For each stock, the results are based on a combined sample that includes both the pre-period and the post-period of the fee experiment.

ARCA is 1. Table 2.12 reports and compares the estimated attraction coefficients before and after the NASDAQ fee experiment for individual stocks. Note that $\beta^{(j)}_{\text{NASDAQ}}$ - post is greater than $\beta^{(j)}_{\text{NASDAQ}}$ - pre for 12 names among the 14 tested securities. For all of these 12 names, the increments are statistically significant under a one-tailed test and a 95% level of confidence. This is again with our model prediction.

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

As stated in (2.12),

$$\mathsf{ED}_i = \frac{W}{\mu\beta_i}. \tag{2.38}$$

We expect the workload $W$ to remain the same after NASDAQ reduces its make-take fee, as the equilibrium value of $W$ depends on the "marginal" exchange, which is likely to be the one

| | $\beta^{(j)}_{\text{NASDAQ}}$ - pre | std. dev. | $\beta^{(j)}_{\text{NASDAQ}}$ - post | std. dev. | INCREASE? | 1-sided 95% test |
|---|---|---|---|---|---|---|
| AAL | 1.1478 | 0.0152 | 1.0787 | 0.0189 | NO | NO |
| BAC | 1.4516 | 0.0297 | 2.5617 | 0.0647 | YES | YES |
| FEYE | 0.6958 | 0.0136 | 0.9543 | 0.0156 | YES | YES |
| GE | 1.1339 | 0.0244 | 1.5250 | 0.0386 | YES | YES |
| GPRO | 0.9285 | 0.0196 | 1.0557 | 0.0250 | YES | YES |
| GRPN | 1.1035 | 0.0253 | 1.4288 | 0.0258 | YES | YES |
| KMI | 1.0814 | 0.0163 | 1.4780 | 0.0267 | YES | YES |
| MU | 1.0314 | 0.0124 | 1.3862 | 0.0186 | YES | YES |
| RAD | 0.9515 | 0.0302 | 1.3530 | 0.0368 | YES | YES |
| RIG | 0.9775 | 0.0230 | 1.1925 | 0.0239 | YES | YES |
| S | 0.9905 | 0.0453 | 1.0957 | 0.0429 | YES | YES |
| SIRI | 1.1787 | 0.0265 | 1.3045 | 0.0347 | YES | YES |
| TWTR | 0.9093 | 0.0235 | 1.0623 | 0.0269 | YES | YES |
| ZNGA | 1.2111 | 0.0337 | 1.1939 | 0.0332 | NO | NO |

**Table 2.12:** Estimates of the attraction coefficient of individual stocks on NASDAQ before and after the fee experiment, and hypothesis testing results on whether the attraction coefficient increases under the fee change.

with the lowest rebate, which is not NASDAQ; we are assuming that the remaining model parameters remain the same. As described above, we anticipate the attraction coefficient $\beta_{\text{NASDAQ}}$ to increase after the fee change, which would result in a lower expected delay $\mathsf{ED}_{\text{NASDAQ}}$ after NASDAQ reduced its make-take fee. An alternative justification is that traders submitting orders into NASDAQ would expect lower expected delays given that they are compensated with a lower rebate when their orders trade. In equilibrium, patient traders will submit orders to higher rebate exchanges, which would result in a lower equilibrium delay at NASDAQ after the fee change.

To test this hypothesis we will compare the normalized expected delay at NASDAQ before and after the fee change. We first compute the measure of expected delay along the lines of Section 2.4.1, as

$$\mathsf{ED}_i^{(s,j)}(t) \triangleq \frac{Q_i^{(s,j)}(t)}{\mu_i^{(s,j)}(t)}, \tag{2.39}$$

for side $s$, security $j$, on exchange $i$, at time slot $t$. We then use these measures to calculate

an aggregate, normalized estimate of the expected delay on NASDAQ, as follows:

$$\tilde{\mathrm{ED}}^{(j)}_{\mathrm{NASADQ}} = \frac{1}{2T} \sum_{s\in\{\mathrm{bid,ask}\}} \sum_{t=1}^{T} \frac{\mathrm{ED}^{(s,j)}_{\mathrm{NASDAQ}}(t)}{\sum_{k=1}^{N} \mathrm{ED}^{(s,j)}_{k}(t)}. \tag{2.40}$$

For each stock, we can obtain two estimates $\tilde{\mathrm{ED}}^{(j)}_{\mathrm{NASADQ}}$ - pre and $\tilde{\mathrm{ED}}^{(j)}_{\mathrm{NASADQ}}$ - post based on the pre-period sample and the post-period sample, respectively. Table 2.13 reports on these two statistics for individual securities. We observe that the normalized expected delay on NASDAQ decreases for all 14 tested securities; in 13 of these 14 cases, the reduction is statistically significant. This agrees with the postulation of our model.

| | $\tilde{\mathrm{ED}}^{(j)}_{\mathrm{NASADQ}}$ - pre | std. err. | $\tilde{\mathrm{ED}}^{(j)}_{\mathrm{NASADQ}}$ - post | std. err. | DECREASE? | 1-sided 95% test |
|---|---|---|---|---|---|---|
| AAL | 0.1978 | 0.0023 | 0.1886 | 0.0027 | YES | YES |
| BAC | 0.1556 | 0.0022 | 0.0963 | 0.0020 | YES | YES |
| FEYE | 0.2282 | 0.0037 | 0.1964 | 0.0031 | YES | YES |
| GE | 0.1688 | 0.0026 | 0.1265 | 0.0023 | YES | YES |
| GPRO | 0.2262 | 0.0053 | 0.2145 | 0.0057 | YES | NO |
| GRPN | 0.2030 | 0.0034 | 0.1700 | 0.0035 | YES | YES |
| KMI | 0.1646 | 0.0018 | 0.1265 | 0.0018 | YES | YES |
| MU | 0.2062 | 0.0023 | 0.1683 | 0.0024 | YES | YES |
| RAD | 0.1750 | 0.0038 | 0.1027 | 0.0030 | YES | YES |
| RIG | 0.1721 | 0.0022 | 0.1401 | 0.0024 | YES | YES |
| S | 0.1765 | 0.0063 | 0.1305 | 0.0034 | YES | YES |
| SIRI | 0.2150 | 0.0072 | 0.1558 | 0.0077 | YES | YES |
| TWTR | 0.1813 | 0.0026 | 0.1453 | 0.0035 | YES | YES |
| ZNGA | 0.1831 | 0.0062 | 0.1515 | 0.0054 | YES | YES |

**Table 2.13:** Estimates of the normalized expected delay on NASDAQ of individual stocks before and after the fee experiment, and hypothesis testing results on whether the normalized expected delay decreases under the fee change.

**Linear relation $\mathbf{ED_{NASADQ} = \beta_{ARCA}/\beta_{NASADQ} \cdot ED_{ARCA}}$.** Last we examine how the fee change affects the linear relation (2.33) in Section 2.4.3, which is one of the major

| | $\tilde{\text{ED}}^{(j)}_{\text{NASADQ}}$ - pre | std. err. | $\tilde{\text{ED}}^{(j)}_{\text{NASADQ}}$ - post | std. err. | DECREASE? | 1-sided 95% test |
|---|---|---|---|---|---|---|
| AAL | 1.0029 | 0.0109 | 0.9681 | 0.0139 | YES | YES |
| BAC | 0.9293 | 0.0121 | 0.5895 | 0.0137 | YES | YES |
| FEYE | 1.2431 | 0.0202 | 1.0567 | 0.0164 | YES | YES |
| GE | 1.0267 | 0.0160 | 0.7863 | 0.0169 | YES | YES |
| GPRO | 1.1619 | 0.0336 | 1.0842 | 0.0360 | YES | NO |
| GRPN | 1.0468 | 0.0191 | 0.8941 | 0.0168 | YES | YES |
| KMI | 0.9938 | 0.0100 | 0.7622 | 0.0108 | YES | YES |
| MU | 1.0318 | 0.0114 | 0.8699 | 0.0127 | YES | YES |
| RAD | 1.1165 | 0.0303 | 0.6757 | 0.0195 | YES | YES |
| RIG | 1.0361 | 0.0128 | 0.8614 | 0.0142 | YES | YES |
| S | 1.2665 | 0.0711 | 0.8699 | 0.0218 | YES | YES |
| SIRI | 1.4101 | 0.1685 | 1.4559 | 0.4852 | NO | NO |
| TWTR | 1.1077 | 0.0166 | 0.9034 | 0.0260 | YES | YES |
| ZNGA | 1.0337 | 0.0404 | 0.8984 | 0.0308 | YES | YES |

**Table 2.14:** Results in parallel to those in Table 2.13 when the expected delays are normalized by median delay instead of by sum of delays.

conclusions arising from our model,

$$\mathsf{ED}_i^{(s,j)}(t) = \frac{\beta_{i'}^{(j)}}{\beta_i^{(j)}}\, \mathsf{ED}_{i'}^{(s,j)}(t). \tag{2.41}$$

Specially, we want to test that when considering the above linear relation between NASDAQ and the benchmark exchange, ARCA, the slope of that linear relation before and after the fee change will decrease, since we expect that the attraction coefficient $\beta_{\text{NASADQ}}$ should increase in response to that change. We perform linear regressions for each security between the expected delays on NASDAQ against that on the benchmark exchange ARCA before and after the fee change:

$$\text{ED}_{\text{NASADQ}} = \beta_0 + \beta_1 \cdot \text{ED}_{\text{ARCA}}, \tag{2.42}$$

Results are in Table 2.15. We observe that the resulting slopes decrease for all 14 tested securities, among which 8 are statistically significant under a one-sided test and a 95% confidence level. In addition, cross-sectionally, in the linear regression before the fee change

$\beta_1 = 0.78564^{***}(0.01571)$, $R^2 = 58\%$; in the linear regression after the fee change $\beta_1 = 0.66384^{***}(0.01221)$, $R^2 = 62\%$. That is, we also find a statistically significant drop in $\beta_1$. Again, this finding agrees with our model prediction.

| | $\beta_1$ (before) | std. err. | $\beta_1$ (after) | std. err. | $\beta_1$ DECREASE? | 1-sided 95% test |
|---|---|---|---|---|---|---|
| AAL | 0.6742 | 0.0308 | 0.5981 | 0.0392 | YES | NO |
| BAC | 0.6176 | 0.0299 | 0.3245 | 0.0269 | YES | YES |
| FEYE | 0.8773 | 0.0761 | 0.7950 | 0.0543 | YES | NO |
| GE | 0.7851 | 0.0384 | 0.4859 | 0.0349 | YES | YES |
| GPRO | 0.3744 | 0.0458 | 0.3314 | 0.0624 | YES | NO |
| GRPN | 0.9570 | 0.0418 | 0.8223 | 0.0234 | YES | YES |
| KMI | 0.8764 | 0.0289 | 0.5400 | 0.0255 | YES | YES |
| MU | 0.8313 | 0.0277 | 0.5741 | 0.0274 | YES | YES |
| RAD | 0.9439 | 0.0611 | 0.3090 | 0.0482 | YES | YES |
| RIG | 0.6571 | 0.0286 | 0.5360 | 0.0282 | YES | YES |
| S | 0.5740 | 0.1151 | 0.5406 | 0.0395 | YES | NO |
| SIRI | 1.3337 | 0.2686 | (0.0001) | 0.0135 | YES | YES |
| TWTR | 0.8406 | 0.0679 | 0.7470 | 0.0724 | YES | NO |
| ZNGA | 0.0546 | 0.0087 | 0.0425 | 0.0174 | YES | NO |

**Table 2.15:** Linear regression results of equation (2.42) and hypothesis testing results on whether the slope decreases after the fee change on NASDAQ.

| | $\beta_1$ (before) | std. err. | $\beta_1$ (after) | std. err. | $\beta_1$ DECREASE? | 1-sided 95% test |
|---|---|---|---|---|---|---|
| AAL | 0.7960 | 0.0149 | 0.8121 | 0.0188 | NO | NO |
| BAC | 0.6774 | 0.0154 | 0.3789 | 0.0138 | YES | YES |
| FEYE | 1.3292 | 0.0428 | 1.0828 | 0.0288 | YES | YES |
| GE | 0.8285 | 0.0221 | 0.5942 | 0.0194 | YES | YES |
| GPRO | 0.7147 | 0.0375 | 0.6852 | 0.0453 | YES | NO |
| GRPN | 0.9706 | 0.0294 | 0.8040 | 0.0207 | YES | YES |
| KMI | 0.9253 | 0.0150 | 0.6436 | 0.0123 | YES | YES |
| MU | 0.9034 | 0.0152 | 0.6440 | 0.0140 | YES | YES |
| RAD | 1.0516 | 0.0440 | 0.4285 | 0.0366 | YES | YES |
| RIG | 0.7851 | 0.0199 | 0.6453 | 0.0177 | YES | YES |
| S | 0.6519 | 0.1113 | 0.6650 | 0.0299 | NO | NO |
| SIRI | 1.4071 | 0.2435 | 0.0017 | 0.0134 | YES | YES |
| TWTR | 1.1070 | 0.0354 | 0.8970 | 0.0390 | YES | YES |
| ZNGA | 0.0584 | 0.0091 | 0.0550 | 0.0178 | YES | NO |

**Table 2.16:** Results in parallel to those in Table 2.15 when the linear regressions are performed without intercept.

# Chapter 3

# Optimal Execution in a Limit Order Book and an Associated Microstructure Market Impact Model

## 3.1. Introduction

Modern equity markets have, to a large extent, become computerized technological systems. Market participants, including institutional investors, market makers, and opportunistic investors, interact within today's high-frequency marketplace with the use of electronic algorithms. These algorithms differ across participants and trading styles. At a high level, they dynamically optimize where, how often, and at what price to trade taking into account the state of the exchanges and other real-time market information. Our goal in this chapter is

to develop models based on queueing theory for the dynamics of an electronic market over short time scales, and to understand how features of the market microstructure impact the execution costs that market participants face.

We will focus on markets that are organized as so-called *electronic limit order books* (LOBs). This is the dominant market structure among, for example, exchange-traded U.S. equities. In an electronic limit order book, traders may provide liquidity by submitting limit orders to buy or sell specific quantities of stock at a specified price, or remove liquidity by sending market orders to buy or sell at the best available prices. When a market order arrives, it will be matched by the exchange to a contra-side resting limit order. These resting orders are first prioritized by price, and then, within each price level, prioritized by their time of arrival. In this way, each price level can be associated with a queue of resting limit orders that await execution according to a first-in-first-out (FIFO) service discipline, and an electronic limit order book can be naturally modeled as a multi-class queueing system.

A simplified view of a typical portfolio manager is as an agent that makes high-level decisions to buy or sell quantities of securities. The outcomes of these investment decisions are then delegated to a 'trader' that executes them, often making use of a so called 'algorithmic trading' system. These systems are developed internally by large institutional investors or, alternatively, offered as a service by a multitude of banks or brokers. Broadly speaking, such algorithmic trading strategies are designed hierarchically. First, they decide how to schedule the parent order, at a high level, over the course of its execution horizon. For example, if an investor seeks to buy a block of shares over the course of a trading day, this might involve scheduling target quantities for purchase in 5-minute intervals. In this way, the trade scheduling phase involves strategic decisions that consider trade-offs that are realized over minutes or hours. Second, they consider each such sub-interval of the longer horizon, and decide how to execute the target quantity over the sub-interval by dividing it into smaller

child orders that are tactically directed to the market either as market or limit orders at optimized price levels and time points. This second phase is often referred to as the micro-trader or slicer, and involves tactical decisions that consider trade-offs on the time scale of seconds to minutes; the queueing delay incurred by limit orders is an important consideration in this step.

An essential input to both the portfolio selection decision as well as the algorithmic trade execution process is the so-called *market impact model*. This model estimates the anticipated cost of a trade and takes into account the adverse effect of one's own trading activity to the price of the security — i.e., how much will the price move against a trader that is buying or selling a block of a specific stock over a specified time horizon. The market impact model depends on the characteristics of the security, such as its liquidity, volatility and typical bid-ask spread, as well as the size and timing of the trade itself. In portfolio construction, a market impact model is often used as a penalty term to capture the trading frictions and resulting costs associated with a portfolio transition. In trade scheduling, it is used in the context of deciding how aggressively to trade — aggressive execution will result in high expected execution costs over shorter trading horizons but reduce execution risk due to exposure to fluctuating market prices. In the micro-trader, a market impact model is used in the tactical optimization of order placement decisions.

In this chapter, we first formulate and solve a stylized version of the optimal execution problem faced by the micro-trader described above that takes the form of optimally buying (or selling) a pre-specified quantity of stock over a fixed short time horizon, typically in the order of a few minutes. Then, leveraging the solution of the execution problem, we construct a market impact model that explicitly takes into account the microstructure information that describes the state and queueing dynamics of the limit order book. Specifically, the key contributions of the chapter are the following: (a) We develop a model of the LOB

as a multi-class queueing network. Using a fluid (deterministic, mean-field) model of the queueing system, we solve the resulting optimal execution problem, that describes what fraction of the trade quantity will be executed using limit and market orders and at what price levels. (b) Our optimal execution problem yields an estimate for the (optimized) execution costs, which suggests a functional form for a market impact model and identifies relevant microstructure variables (e.g., queue lengths, arrival rates, etc.) that impact trading costs. The microstructure market impact model seems to be novel viz the extensive literature on this topic and to be of practical interest in estimating transaction costs and optimizing trading decisions over short time horizons of the order of a few minutes. (c) Finally, we calibrate the microstructure market impact model using a proprietary data set of algorithmic trades and contemporaneous real-time measurements of limit order book variables. We compare the quality of the statistical fit of the microstructure model to what can be achieved using a typical macroscopic market impact model that estimates costs without consideration of limit order book variables. We find that our microstructure impact model yields a factor of four improvement in out-of-sample explanatory power. We further test the robustness of our model over its specification and over the problem primitives. We find our model has the most explanatory power for larger orders (measured as a percentage of overall volume) and for assets with greater market depth (measured through queues sizes capturing available liquidity). These correspond to settings where our fluid model assumptions are most realistic. Further, we note conventional macro models are also more successful in settings with greater market depth, a fact that seems unobserved thus far in the literature.

**Literature review.** This work is related to the growing literature that lies on the interface of queueing and the study of limit order book markets. This connection was first illustrated by Cont et al. (2010); see also Cont and Larrard (2013), Lakner et al. (2013), Blanchet and Chen (2013), Stoikov et al. (2011), and Lakner et al. (2014). Our model builds on Cont

et al. (2010), recognizing the multiple price levels in a limit order book can be modeled as a multi-class queue. We work directly with the fluid model representation and do not study the stochastic dynamics of the multi-class queue. The majority of the papers above focus on characterizing the performance of the limit order book, in many cases involving fluid or diffusion approximations. Our emphasis is on optimization of tactical trading decisions, and specifically in optimizing how to execute a block of shares in a limit order book over a predetermined time horizon that is of the same order as that of queueing delays in the order book, and as such modeling of queueing effects becomes important. Related work includes that of Guo et al. (2013), who study a problem of optimizing when to send limit orders and market orders in the market, taking into account, in a stylized manner, the limit order book dynamics but excluding a careful consideration of queueing delays and order cancellation effects. Cont and Kukanov (2013) study the smart order routing problem, specifically taking into account the fact that there are multiple exchanges to which one can post a limit order, so the control decision becomes how much to post and to which exchange. Our work considers one consolidated limit order book, like Guo et al. (2013), but models explicitly the queueing dynamics, order cancellations, and the ability to trade aggressively on multiple price levels with market orders. Apart from optimizing limit order placement, we find that the optimized routing of market orders over the optimization horizon is an important ingredient that affects the overall execution cost; in particular, it is typically not optimal to send all market orders to trade at the end of the time horizon. The resulting execution cost motivates the microstructure market impact model.

A separate set of papers deal with the longer horizon trade scheduling problem. Bertsimas and Lo (1998) solved this problem when optimizing the expected cost, and Almgren and Chriss (2001) considered the mean-variance criterion; see also Almgren (2003) and Huberman and Stanzl (2005). These papers use a market impact model to capture the cost of the

execution expressed as a function of the speed of trading, but do not explicitly model the interaction in a limit order book, or the state variables of the order book. Obizhaeva and Wang (2013), Rosu (2009), Alfonsi et al. (2010) treat the market as one limit order book and use an aggregated and stylized model of market impact to capture how the price moves as a function of trading intensity. These references address the trade scheduling problem, whose longer time horizon allows one to abstract away the queueing effects that are inherent in the limit order book.

Market impact models estimate the expected transaction cost of a trade. They take various functional forms, and typically deconstruct the price impact into temporary and permanent components, and further specify the decay behavior of the temporary contribution. They depend on specific characteristics of the stock as well as the speed of trading, often assumed to be a constant participation rate – e.g., an order executed at 10% participation rate would aim to trade 100 shares for every 1,000 shares traded in the market across all participants. Huberman and Stanzl (2004) showed using a no-arbitrage argument that the permanent price impact must be a linear function of the quantity traded; see also Gatheral (2010). The functional form and decay kernel of the temporary impact term is not as simple to characterize analytically. The simplest assumption treats that decay as being instantaneous. Other alternatives typically allow for exponential or power decay functions. The functional form that specifies the magnitude of the temporary cost is itself typically assumed to be linear or sub-linear function of the speed of trading; stylized analytical arguments and statistical evidence suggest a sub-linear functional form. For example, Chacko et al. (2008) provide empirical evidence that the expected price impact is proportional to the square root of the quantity traded; see also Bouchaud et al. (2008).

We refer to the class of models described above as macroscopic (or "macro") models in the sense that they do not take into account microstructure variables that can be gleaned from the

limit order book, and typically try to give cost estimates over long time durations, minutes to hours to days. These models are typically estimated through large scale cross-sectional regressions based on the realized costs of a proprietary set of algorithmically executed trades. Almgren et al. (2005) describe an econometric approach for that problem, while Rashkovich and Verma (2012) provide important insights that improve the estimation procedure, and allow for more accurate de-trending of the trade data. Moallemi et al. (2014) extend the above approach to include a short term alpha fixed effect associated with the identity of the trader.

In contrast to the above mentioned papers, our analysis proposes a temporary price impact model that explicitly depends on limit order book variables. It is best suited over short time horizons of the order of minutes (the same order of magnitude as that of queueing delays encountered by limit orders until they execute in the market). Recently, Cont et al. (2014) studied a price impact model expressed as a function of the so-called order flow imbalance that measures the difference between events (arrivals, trades and cancellations) on the two sides of the limit order book. Imbalance should be normalized by the queue depth, which is something that emerges in our work as well in capturing the effect of market orders; limit orders have a different relation to depth that we also identify. Cont et al. (2014) did not suggest a model that could be used to explain and predict trading costs, but such an extension may be possible.

The remainder of the chapter is organized as follows. Section 3.2 models the operation of a limit order book as a multi-class queueing system and studies its fluid dynamics. Section 3.3 states the optimal execution problem. Section 3.4 charactererizes the optimal strategy, on which a microstructure cost function we provide in Section 3.5 is predicated. Section 3.6 reports on the empirical performance of our model and provides a comparison with some benchmark models in the literature.

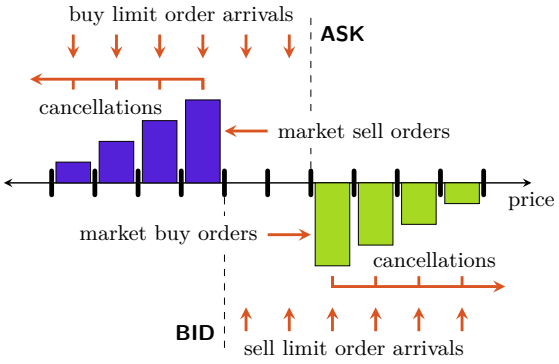



**Figure 3.1:** An illustration of an electronic limit order book.

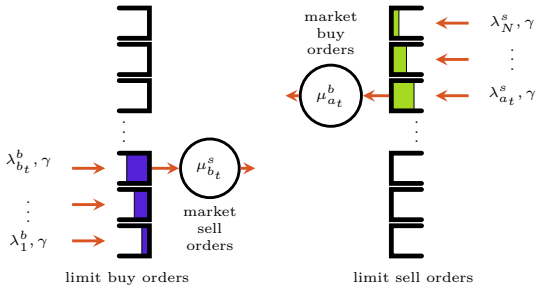



**Figure 3.2:** An illustration of the coupled, multi-class priority queueing network associated with an electronic limit order book and its fluid dynamics.

## 3.2. The Limit Order Book

An electronic limit order book (LOB) can be modeled as a multi-class queueing system. In broad terms, we will associate queues at each price point where buy or sell limit orders can wait until executed or canceled by the respective traders. We model and track cumulative arrivals of limit orders into the various queues, model the arrival and execution behavior of market orders, and subsequently discuss the dynamics of this queueing system. Figure 3.1 provides a useful schematic to visualize the various aspects of the LOB.

This chapter studies an optimal execution problem and explores how this provides the basis of a microstructure-based transaction cost function. The specific problem that we analyze is one of optimally buying $C$ shares of a security at the lowest possible price over a given time horizon $T$. In our setting, we typically imagine $T$ to be of the order of a few minutes.

This transient optimal control problem motivates the use of a deterministic fluid model (sometimes known as a "mean field" model) for the evolution of the LOB, where the discrete

and stochastic primitive processes (e.g., order arrivals, cancellations) are replaced by continuous and deterministic analogues, where infinitesimal orders arrive continuously over time at a rate that is equal to the instantaneous intensity of the underlying stochastic processes. This model can be justified as an asymptotic limit using the functional strong law of large numbers in settings where the rates of order arrivals grow large but the size of each individual order is small relative to the overall order volume over any interval of time.[1] It is well-suited for characterizing transient dynamics in such systems, which roughly correspond to the time scale over which queues drain or move from some initial configuration to an equilibrium state; this is also the relevant time horizon for our optimal execution problem. Indeed, our model is oriented towards liquid securities, where orders arrive on a time scale measured in milliseconds to seconds, while we will consider a time horizon on the order of minutes.

### 3.2.1. Multiclass Queueing Network

Our multiclass queueing network model of the LOB is defined as follows:

**Prices.** We will consider a discrete price grid indexed by $i \in \{1, \ldots, N\}$, refer to the $i$th price point by $p_i$, and assume that prices are labeled so that $p_1 < p_2 < \cdots < p_N$; it is natural to think that this price sequence is in uniform increments of an underlying minimum tick-size.

**Queues.** At each price point $p_i$ we associate two queues for buy and sell limit orders, respectively. Specifically, at each time $t \geq 0$, denote by $Q_i^b(t), Q_i^s(t) \in \mathbb{R}_+$ the total quantity of shares available for purchase or sale, respectively, at price level $p_i$. We define the *best-bid*

[1] Mandelbaum and Pats (1995) provides a framework that could be adapted into this setting to justify such a limit.

queue $b_t \in \{1, \ldots, N\}$ to be the non-empty queue of buy orders of highest price, i.e.,

$$b_t \triangleq \min \left\{ 1 \leq i \leq N \ : \ Q_j^b(t) = 0, \text{ for all } i < j \leq N \right\},$$

and the *best-ask* queue $a_t \in \{1, \ldots, N\}$ to be the non-empty queue of sell orders of lowest price, i.e.,

$$a_t \triangleq \max \left\{ 1 \leq i \leq N \ : \ Q_j^s(t) = 0, \text{ for all } 1 \leq j < i \right\}.$$

We denote the overall state of the LOB by $Q(t) \triangleq (Q^b(t), Q^s(t)) \in \mathbb{R}_+^N \times \mathbb{R}_+^N$, where

$$Q^b(t) \triangleq \left(Q_1^b(t), \ldots, Q_N^b(t)\right) \in \mathbb{R}_+^N \quad \text{and} \quad Q^s(t) \triangleq \left(Q_1^s(t), \ldots, Q_N^s(t)\right) \in \mathbb{R}_+^N.$$

We will require that queue length vectors satisfy $b_t < a_t$, or, equivalently, that $p_{b_t} < p_{a_t}$, i.e., the best-bid price is strictly less than the best-ask price. This will be made clearer through the equations of dynamics. Further, we require that both sides of the limit order book be non-empty, i.e., the best bid and best ask levels are well defined and $Q_{b_t}^b(t) \neq 0$ and $Q_{a_t}^s(t) \neq 0$. Denote by $\mathcal{Q} \subset \mathbb{R}_+^N \times \mathbb{R}_+^N$ the set of such feasible queue length vectors.

**Limit order arrivals.** *Limit orders* seek to buy (resp., sell) a certain quantity of shares at any price up to and including a limit price that is below (resp., above) the best-bid (resp., best-ask) price in the market.[2] Limit orders cannot be filled upon their arrival, but instead join FIFO queues associated with their limit prices and wait until they are filled or canceled.

**Market order arrivals.** *Market orders* seek to buy (resp., sell) a certain quantity of shares at the "best" available price. Market orders trade instantaneously against posted limit orders on the contra-side of the order book according to a *price-time* priority rule: when matching

[2]These are commonly known as non-marketable limit orders. In our setting, limit orders that do not satisfy this price condition (i.e., marketable limit orders) are equivalent to market orders and thus considered as such.

a market order to buy (resp., sell) against resting limit orders to sell (resp., buy), the resting orders are first considered in increasing (resp., decreasing) order of price; within each price level, resting limit orders are considered in a first-in-first-out (FIFO) order. The resting limit order shares that are matched to and filled by a market order are subsequently removed from the order book.

**Limit order cancellations.** Resting limit orders may be canceled at any point. When a cancellation occurs, the canceled shares are removed from their corresponding queue in the order book.

In queueing parlance, a limit order book corresponds to a coupled multiclass queueing network; cf. Figure 3.2. Job arrivals correspond to the arrival of limit orders, service completions correspond to the arrival of market orders, and abandonments correspond to the arrival of limit order cancellations. The price-time priority rule creates a service discipline where queues are assigned priority classes based on their prices and where each queue is served in FIFO.

### 3.2.2. Fluid Model Dynamics

The *fluid model* approximation of the LOB replaces stochastic and discrete arrival and cancellation processes by continuous and deterministic flows.

**Limit order arrivals.** At time $t$, we assume that buy and sell limit orders arrive at each price level $p_i$ with rates $\lambda_i^b \cdot \mathbf{1}(i \leq b_t)$ and $\lambda_i^s \cdot \mathbf{1}(i \geq a_t)$, respectively, given two vectors $\lambda^s, \lambda^b \in \mathbb{R}_+^N$. In other words, limit orders arrive at price levels that are at the top-of-the-book, i.e., at the current best-bid and best-ask, or at prices inside the book, i.e., buy orders at prices below the best-bid and sell orders at prices higher than the best-ask.[3]

[3]The rates $\lambda_i^b$ and $\lambda_i^s$ are specified as functions of the price level $p_i$, and and these limit order flows turn off depending on the price level as compared to the prevailing best-bid and best-ask prices. A more complex

**Market order arrivals.** Market orders to sell or to buy arrive at rates that are dependent on the current best-bid and best-ask prices, respectively, denoted by $\mu^s_{b_t}$ and $\mu^b_{a_t}$. The two vectors $\mu^s, \mu^b \in \mathbb{R}^N_+$ define the market order arrival rates at different price levels for the best-bid and best-ask, respectively.

**Limit order cancellations.** We assume that resting limit orders are canceled at a uniform rate $\gamma > 0$, which implies that the cancellation rate per unit time in a queue of size $Q$ is $\gamma Q$.

Combining the above, we obtain the following ODEs for the order book state process:

$$\dot{Q}^b_i(t) = \lambda^b_i \cdot \mathbf{1}(i \leq b_t) - \mu^s_i \cdot \mathbf{1}(i = b_t) - \gamma Q^b_i(t), \qquad \forall\ 1 \leq i \leq N, \tag{3.1}$$

$$\dot{Q}^s_i(t) = \lambda^s_i \cdot \mathbf{1}(i \geq a_t) - \mu^b_i \cdot \mathbf{1}(i = a_t) - \gamma Q^s_i(t), \qquad \forall\ 1 \leq i \leq N. \tag{3.2}$$

We will make the following assumption regarding the arrival rate parameters:

**Assumption 4.** *The arrival rate of limit orders at any price level exceeds the arrival rate of contra-side market orders associated with that price level. That is, $\lambda^s_i \geq \mu^b_i$ and $\lambda^b_i \geq \mu^s_i$ for all $1 \leq i \leq N$.*

The following lemma characterizes the unique stationary point of the fluid dynamics (3.1)–(3.2).

**Lemma 2.** *Given an arbitrary initial condition $Q(0) \in \mathcal{Q}$, there exists a unique solution $Q\colon [0, \infty) \to \mathcal{Q}$ to the fluid model ODEs (3.1)–(3.2). This solution satisfies:*

*(i) $b_t = b_0$, $a_t = a_0$, for all $t \geq 0$,*

---

model would allow for the rates at $p_i$ to depend on the distances of $p_i$ from $b_t$ and $a_t$, and possibly on the queue lengths, especially these at the best-bid and best-ask. Given our end goal of extracting a transaction cost model which is parsimonious and easily estimable using data, we will not consider these extensions herein.

*(ii) As $t \to \infty$, $Q(t) \to q^*$, where $q^* \triangleq (q^{*,b}, q^{*,s})$ is given by*

$$q_i^{*,b} \triangleq \begin{cases} \lambda_i^b/\gamma & \text{if } 1 \le i < b_0, \\ \dfrac{\lambda_i^b - \mu_i^s}{\gamma} & \text{if } i = b_0, \\ 0 & \text{if } b_0 < i \le N, \end{cases} \qquad q_i^{*,s} \triangleq \begin{cases} 0 & \text{if } 1 \le i < a_0, \\ \dfrac{\lambda_i^s - \mu_i^b}{\gamma} & \text{if } i = a_0, \\ \lambda_i^s/\gamma & \text{if } a_0 < i \le N, \end{cases}$$

(All proofs can be found in the Appendix.) Part (i) of Lemma 2 states that starting from any initial condition, the best-bid and best-ask prices remain constant. This is a direct consequence of Assumption 2.[4] Part (ii) of Lemma 2 identifies the long-run equilibrium configuration of the limit order book in terms of the rate parameters and the initial condition.

## 3.3. The Optimal Execution Problem

We consider a trader that seeks to buy $C$ shares over a given time interval $[0, T]$ by posting limit and market orders over time and at various price levels in the limit order book. The trader's objective is to minimize the average buying price. We describe this problem in detail as follows:

**Limit orders.** Given Lemma 2, any limit orders posted at price levels $p_i$ with $i < b_t$ (i.e., strictly below the best-bid price) will never trade and can therefore be excluded from consideration, without loss of generality. The following assumption also disallows limit orders strictly above the best-bid price:

**Assumption 5** (No Limit Orders Inside Spread)**.** *We restrict attention to execution policies that, at each time $t$, submit no limit orders at price level $i$, if $i > b_t$. In other words, no limit*

[4] If Assumption 2 is relaxed, then there may be a short term transient that one would need to consider, e.g., the event rates $\lambda_i, \mu_i$ may be imbalanced in a way that the best-bid or the best-ask would change.

*orders are submitted inside the current best-bid and best-ask prices.*

We make this assumption for tractability reasons. It disallows the trader from setting a new best-bid price. Under Assumption 5, the limit order placement decision is reduced to selecting how much quantity to submit at the best-bid price level $p_{b_t}$. In our model, again without loss of generality, we can assume that all limit orders are placed in a single block at time $t = 0$.[5] Thus, we will restrict attention to policies which place all limit orders (if any) at time $t = 0$ at the best-bid price level $b_0$. We denote by $S_L$ the aggregate size of this limit order, and require that $0 \leq S_L \leq C$.

**Market orders.** The trader may also place market orders. We denote by $S(t)$ the cumulative number of market orders placed over the interval $[0, t]$.

**Assumption 6** (Regularity of Market Orders)**.** *The market order process $S(\cdot)$ must satisfy:*

*(i) $S(\cdot)$ is nondecreasing and right continuous with left limits. Denote by $S(t^-)$ the left limit of function $S(\cdot)$ at $t \in (0, T]$ and define $S(0^-) \triangleq 0$.*

*(ii) $S(\cdot)$ has finitely many jump discontinuities and is absolutely continuous on the intervals between jumps.*

Given the above assumption, the process $S(\cdot)$ can be rewritten as a combination of discrete jumps or "block" trades, and continuously emitted orders or "flow" trades. Specifically, denote the times of the jump discontinuities by $0 \leq t_1 \leq \cdots \leq t_K \leq T$. Denote by $J_k$ the size of the $k$th jump or block trade. Then, there exists a Lebesgue integrable instantaneous rate

[5] We will not provide a proof of that assertion. Intuitively, any policy that submits limit orders at some time $t > 0$ can be weakly improved by submitting the same quantity of limit orders at $t = 0$, which due to the FIFO priority rule, will now execute sooner.

function $r\colon [0,T] \to \mathbb{R}_+$ such that

$$S(t) = \sum_{k=1}^{K} \mathbf{1}\{t_k \le t\} \cdot J_k + \int_0^t r(s)\, ds, \quad \forall\, t \in [0,T]. \tag{3.3}$$

**Constraints on the policy.** An execution policy is specified via a quantity of limit orders $S_L$ and a market order process $S(\cdot)$ that comprises of block trades $\{J_k\}$ and flow trades $r(\cdot)$.

**Definition 3** (Admissible Policy). *Given an initial order book state $Q(0^-) \in \mathcal{Q}$, an execution policy $(S_L, S(\cdot))$ with representation (3.3) is said to be admissible if it satisfies*

*(i) A total of $C$ shares is purchased by the end of the time horizon.*

*(ii) For each block trade $J_k$ occurring at time $t_k$, with $k = 1, \cdots, K$, the sizes of block trade does not exceed the available liquidity on the ask side of the order book, i.e.,*

$$J_k \le \sum_{i=a_{t_k^-}}^{N} Q_i^s(t_k^-).$$

Denote by $\mathcal{P}(Q(0^-))$ the set of admissible policies given an initial condition $Q(0^-) \in \mathcal{Q}$. For simplicity, we will further assume that ask queues outside of the best-ask price start at their stationary queue lengths specified in Lemma 2. Specifically:

**Assumption 7** (Initial Conditions). *$Q(0^-) \in \mathcal{Q}^{eq}$, where*

$$\mathcal{Q}^{eq} := \{q \ :\ q \in \mathbb{R}_+^N,\ q_i = \lambda_i^s/\gamma \text{ for } i = a_0 + 1, \cdots, N\}.$$

**Price movement and the effect on book dynamics.** We need to augment the dynamics specified in Section 3.2, to incorporate the effect of the trader's actions:

(a) Buy market orders submitted by the trader may empty queues on the ask side of the LOB, which would induce a price change in the order book. We will assume that the the order book maintains a constant bid-ask spread after a price shift, formalized in Assumption 8.

(b) Buy limit orders submitted by the trader to the best-bid price must be tracked separately from other limit orders at the best-bid price, so as to maintain their queue position and priority to execute relative to other orders at the same price level. Specifically, the total quantity of buy limit orders $Q^b_{b_t}(t)$ at the best-bid price level at time $t$ can be decomposed as follows

$$Q^b_{b_t(t)} = Q^0(t) + Q_L(t) + Q^1(t),$$

where $Q^0(t)$ is quantity of limit orders still in the queue that were submitted at $t = 0^-$; $Q_L(t)$ is quantity of limit orders still in the queue submitted by the trader at $t = 0$; and $Q^1(t)$ is quantity of limit orders submitted by other participants after $t = 0$. These orders are placed in the queue as illustrated in Figure 3.3: $Q^0(t)$ is in the front of the queue, followed by $Q_L(t)$ and then by $Q^1(t)$.

The trader's market order policy may deplete price levels on the ask side of the book. Let $\tau_i$ be time when the aggregate queue lengths up to price $p_i$, for $i = a_0, \ldots, N$, are depleted, i.e.,

$$\tau_i := \inf \left\{ t \in [0, T] \mid Q^s_j(t) = 0, \ \forall\ j = 0, \ldots, i \right\}, \tag{3.4}$$

and set $\tau_i = \infty$ if the condition is not satisfied at any time in $[0, T]$.

Note that we have suppressed the dependence of these times on the initial conditions and the execution policy in our notation. By their definition, $0 \leq \tau_{a_0} \leq \cdots \leq \tau_N$. The best ask process $a_t$, for $t \in [0, T]$, can be expressed in terms of these depletion times by

$$a_t = a_0 + \sum_{i=a_0}^{N} \mathbf{1} \left\{ \tau_i \leq t \right\}. \tag{3.5}$$

The next assumption describes the order book behavior when an ask queue is depleted. We assume that the bid-side queues shift to higher price points as needed to ensure that the bid-ask spread $a_t - b_t$ is constant over time.

**Assumption 8** (Constant Bid-Ask Spread). *Denote by $k_t \triangleq a_t - a_{t-}$ the price jump at the ask at a time $t \in \{\tau_{a_0}, \ldots, \tau_N\}$. We assume that the bid-side of the book shifts up by the same amount $k_t$ at each such time $t$. In other words,*

$$Q_i^b(t) = \begin{cases} Q_{i-k_t}^b(t^-) + \mathbf{1}\{t = 0, i = b_0\} \cdot S_L & \text{for } i = 1 + k_t, \ldots, b_t, \\ \lambda_i^b/\gamma & \text{for } i = 1, \cdots, k_t, \end{cases} \tag{3.6}$$

*for $t \in \{\tau_{a_0}, \ldots, \tau_N\}$. Further, queue priority at the best-bid price level is not affected by the price change, i.e., $Q^0(t) = Q^0(t^-),\ Q_L(t) = Q_L(t^-),\ Q^1(t) = Q^1(t^-)$, for $t \in \{\tau_{a_0}, \ldots, \tau_N\}$.*

**System dynamics.** Under Assumptions 4–8, and for an admissible policy the evolution of buy limit orders at the best-bid price are as follows:

$$Q^0(0) = Q_{b_0}^b(0^-), \qquad \dot{Q}^0(t) = \begin{cases} -\mu_{b_t}^s - \gamma Q^0(t) & \text{if } Q^0(t) > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.7}$$

$$Q_L(0) = S_L, \qquad \dot{Q}_L(t) = \begin{cases} -\mu_{b_t}^s \cdot \mathbf{1}\{Q^0(t) = 0\} & \text{if } Q_L(t) > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.8}$$

$$Q^1(0) = 0, \qquad \dot{Q}^1(t) = \lambda_{b_t}^b - \mu_{b_t}^s \cdot \mathbf{1}\{Q^0(t) = Q_L(t) = 0\} - \gamma Q^1(t). \tag{3.9}$$

Specifically, the orders submitted by other participants before $t = 0$ or after $t = 0$ may get canceled at rate $\gamma$, whereas the block of orders submitted by the trader at $t = 0$ will not get canceled. At times $t \in \{\tau_{a_0}, \ldots, \tau_N\}$, the bid-side queues will shift price levels according to

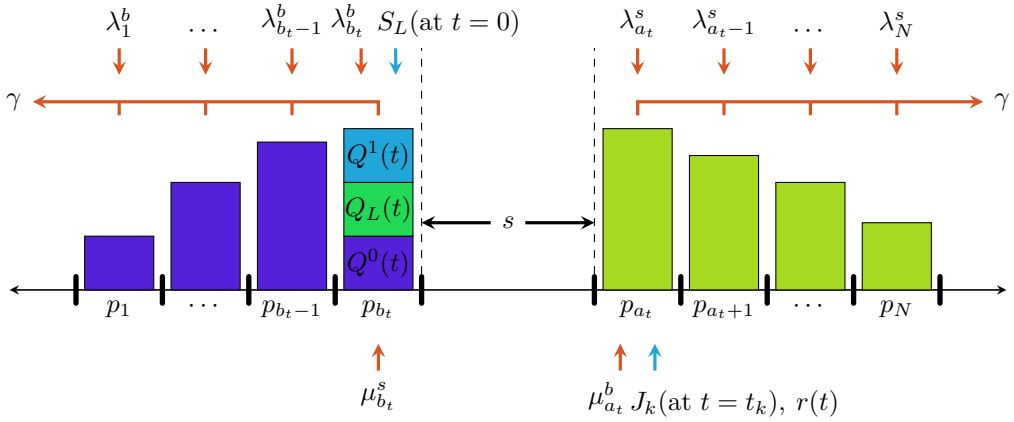



**Figure 3.3:** Illustration of system dynamics.

(3.6). Further,

$$\dot{Q}^b_i(t) = \lambda^b_i \cdot \mathbf{1}(i < b_t) - \gamma Q^b_i(t) \quad \text{for } 1 \le i < b_t,\ t \notin \{\tau_{a_0}, \ldots, \tau_N\}.$$

The ask-side queues evolve, for $1 \le i \le N$ as follows: for $t \in \{t_1, \ldots, t_K\}$,

$$Q^s_i(t) = \begin{cases} \left( Q^s_i(t^-) - \left( J_k - \sum_{j=a_{t^-}}^{i-1} Q^s_j(t^-) \right)^+ \right)^+ & \text{if } i \ge a_{t^-}, \\ 0 & \text{otherwise,} \end{cases}$$

and for $t \notin \{t_1, \ldots, t_K\}$,

$$\dot{Q}^s_i(t) = \lambda^s_i \cdot \mathbf{1}\{i \ge a_t\} - \left( \mu^b_i + r(t) \right) \cdot \mathbf{1}\{i = a_t\} - \gamma Q^s_i(t) \quad \text{for } a_t \le i \le N.$$

**Objective function.** The optimal execution problem is to pick an admissible policy

$(S_L, S(\cdot))$ to minimize the total purchase cost

$$P(S_L, S(\cdot)) \triangleq \int_0^T p_{b_t} \cdot \mu^s_{b_t} \mathbf{1}\left\{Q^0(t) = 0, Q_L(t) > 0\right\}\, dt + \int_0^T p_{a_t} \cdot r(t)\, dt + \sum_{k=1}^{K} \left( \sum_{j=a_{t_k^-}}^{a_{t_k}-1} p_j Q^s_j(t_k^-) + p_{a_{t_k}} \left( J_k - \sum_{j=a_{t_k^-}}^{a_{t_k}-1} Q^s_j(t_k^-) \right) \right), \tag{3.10}$$

under Assumptions 4–8, and where the first term is the cost of the executed limit orders, and the second and third terms are the costs due to the flow and block market order trades, respectively.

## 3.4. The Optimal Execution Policy

The characterization of the optimal execution policy involves three steps: (a) We identify the execution policy that uses only market orders and minimizes the time needed to fill a target quantity at a given price level. (Lemma 3.) (b) We characterize the optimal execution policy that would complete a target quantity within the specified time horizon again using only market orders. (Lemma 4.) (c) Steps (a)–(b) will ultimately guarantee that the market order execution path will maintain the current price level $(b_0, a_0)$ for all $t < T$, and then push the price at $T$ as needed to complete the target quantity. This property allows us to compute the maximum number of shares that can be executed via limit orders at the best bid, $b_0$, taking into account the queue priority of orders posted into that best-bid queue prior to $t = 0$ and their respective cancellations over the execution horizon. (Lemma 5.) Jointly these results characterize the optimal policy in Theorem 5.

We first consider the problem of executing in minimum time a target quantity $C_{a_0}$ using market orders only at $p_{a_0}$, i.e., the (highest priority) best-ask queue that is non-empty at time $t = 0$. In studying this problem we impose the constraint that the queue cannot be depleted

prior to finishing the target quantity, and, specifically, that the queue length stays above some arbitrary value $\varepsilon > 0$. This is imposed for mathematical tractability and to guarantee the existence of an optimal policy; without that minimum quantity, the control will strive to take the queue length arbitrarily close to zero, yet without actually depleting the queue that would trigger a price change. This assumption is useful in deriving the structural insight of the next lemma, and will be relaxed later on.

**Lemma 3** (Market Orders at One Price). *Without loss of generality we focus at the price level $p_{a_0}$. Let $C_{a_0}$ be the target number of shares to trade using market orders only at $p_{a_0}$ and let $Q^s_{a_0}(0^-) > 0$ be the initial queue length. Consider the minimum time control problem:*

$$\textit{minimize} \quad \{\tau : \ S(\tau) = C_{a_0}\}, \tag{3.11}$$

*over admissible market order control trajectories $\{S(t) : t \in [0, \tau]\}$ that satisfy the following constraints*

$$Q^s_{a_0}(t) \geq \varepsilon, \quad t \in [0, \tau) \quad \textit{and} \quad S(\tau) - S(\tau^-) \leq Q^s_{a_0}(\tau^-). \tag{3.12}$$

*The optimal control trajectory $\{S^*(t), t \in [0, \tau]\}$ for (3.11)–(3.12) is the following:*

$$S^*(0) = \begin{cases} Q^s_{a_0}(0^-) - \varepsilon, & \textit{if } C_{a_0} > Q^s_{a_0}(0^-), \\ C_{a_0}, & \textit{otherwise}, \end{cases} \tag{3.13}$$

*and*

$$\dot{S}^*(t) = r^*(t) = \kappa_{a_0}, \quad S^*(t) - S^*(t^-) = 0, \quad \textit{for } t \in (0, \tau), \quad \tau = \frac{\left(C_{a_0} - Q^s_{a_0}(0^-)\right)^+}{\kappa_{a_0}}, \tag{3.14}$$

*where* $\kappa_{a_0} := \lambda^s_{a_0} - \mu^b_{a_0} - \gamma\varepsilon$, *and*

$$S^*(\tau) - S^*(\tau^-) = \begin{cases} \varepsilon, & \text{if } C_{a_0} > Q^s_{a_0}(0^-), \\ C_{a_0}, & \text{otherwise.} \end{cases} \tag{3.15}$$

The intuition behind the lemma is simple: we trade as much as possible without depleting the queue at $t = 0$ to avoid the effect of order cancellations at the best-ask queue; if the order is not completed, we trade with a continuous submission of market orders until we fill $C_{a_0} - \varepsilon$ shares; we finish the trade with a small block trade of size $\varepsilon$. Note that the value of $\kappa_{a_0}$ is such that the queue length will remain constant at $\varepsilon$ during $(0, \tau)$. The total duration of the execution is 0 if the target quantity is less than the displayed depth, and is otherwise determined by the length of the interval that is needed to continuously trade at rate $\kappa_{a_0}$ until the order is completed.

Based on Lemma 3, the length of the execution interval $l_i := \tau_i - \tau_{i-1}$ to execute $C_i$ shares at price $p_i$, for $i = a_0, \cdots, N$, is

$$l_i = \frac{(C_i - Q^s_i(0^-))^+}{\lambda^s_i - \mu^b_i - \gamma\varepsilon} \approx \frac{(C_i - Q^s_i(0^-))^+}{\kappa_i}, \tag{3.16}$$

where we redefine $\kappa_i := \lambda^s_i - \mu^b_i$, and the approximation occurs when $\varepsilon$ is small; recall that $Q^s_i(0^-) = \bar{Q}^s_i$ for $i > a_0$. We adopt the above approximation for the remainder of this chapter. Let $C_{a_0}, C_{a_0+1}, \cdots, C_N$ denote the amount of market orders to execute at prices $p_{a_0}, p_{a_0+1}, \cdots, p_N$, respectively. Given the relationship in equation (3.16), the optimal execution problem described in Section 3.3 can be simplified into the following control problem:

$$\underset{S_L, C_{a_0}, \cdots, C_N}{\text{minimize}} \quad \int_0^T p_{b_t} \cdot \mu^s_{b_t} \mathbf{1}\left\{Q^0(t) = 0, Q_L(t) > 0\right\} \, dt + \sum_{i=a_0}^{N} C_i \cdot p_i, \tag{3.17}$$

subject to

$$S_L + \sum_{i=a_0}^{N} C_i = C, \quad S_L, C_{a_0}, \cdots, C_N \geq 0, \tag{3.18}$$

$$\int_0^T \mu_{b_t}^s \mathbf{1}\left\{Q^0(t) = 0\right\} dt \geq S_L, \qquad \text{(limit order time)} \tag{3.19}$$

$$b_t = b_0 + \min\left\{0 \leq j \leq N - a_0 : \sum_{i=a_0}^{a_0+j} l_i > t\right\}, \qquad \text{(limit order dynamics)} \tag{3.20}$$

$$Q^0(t) \text{ satisfies (3.7)}, \ Q_L(t) \text{ satisfies (3.8)}, \ \text{for } t \in [0, T], \qquad \text{(limit order dynamics)} \tag{3.21}$$

$$\sum_{i=a_0}^{N} l_i \overset{(3.16)}{\approx} \sum_{i=a_0}^{N} \frac{(C_i - Q_i^s(0^-))^+}{\kappa_i} \leq T, \qquad \text{(market order time)} \tag{3.22}$$

$$C_i \geq Q_i^s(0^-), \quad \text{for } i < n, \qquad \text{(market order dynamics)} \tag{3.23}$$

$$n = \min\left\{a_0 \leq j \leq N : C_k = 0 \text{ for all } k > j\right\}. \qquad \text{(market order dynamics)} \tag{3.24}$$

Constraint (3.19) upper bounds the number of shares that can be traded using limit orders within time $T$, taking into account the execution priority of limit orders resting in book before time $t = 0$. Constraint (3.22) ensures that the total time taken trading using market orders at different price levels is upper bounded by the specified time horizon $T$. Condition (3.24) identifies the highest price queue in which market orders will be executed, indexed by $n$, at price $p_n$, and (3.23) ensures the time-price priority rule that ensures that all lower priced queues (that have higher priority) will be depleted.

For the remainder the chapter we make the following simplifying assumption on $\kappa_i$:

**Assumption 9.** *Assume that* $\kappa_i = \lambda_i^s - \mu_i^b = \kappa$ *for all* $i$.

$\kappa_i$ captures the rate at which the trader can continuously execute with market orders when the best-ask is at price $p_i$, and without causing a price change. One would expect the continuous trading rate $\kappa_i$ increases as the price moves up, because more limit orders to sell get submitted at these more favorable price levels. The solution of the optimal execution problem is more involved in that case, and we will not consider it in this chapter, given our ultimate interest in specifying a parsimonious microstructure market impact model.

Lemma 4 studies a subproblem of (3.17)–(3.24) that seeks to optimize over how to execute $C'$ shares over a time horizon of length $T$ at minimum cost using only market orders, allocated according to $C_{a_0}, \cdots, C_N$ across price levels.

**Lemma 4** (Market Orders Across Price Levels)**.** *Given initial queue lengths* $Q^s_{a_0}(0^-) > 0$ *and* $Q^s_k(0^-) = \bar{Q}^s_k$ *for* $k = a_0 + 1, \cdots, N$ *as assumed in Section 3.3. Consider the problem of minimizing the total execution cost of* $C'$ *shares of market orders over a time horizon of length* $T$

$$\begin{aligned} \min_{\{C_k \geq 0,\ k=a_0,\ldots,N\}} \quad & \sum_{k=a_0}^{N} C_k \cdot p_k \\ s.t. \quad & \sum_{k=a_0}^{N} C_k = C', \quad \sum_{k=a_0}^{N} l_k \overset{(3.16)}{=} \sum_{k=a_0}^{N} \frac{(C_k - Q^s_k(0^-))^+}{\kappa} \leq T \\ & C_i \geq Q^s_i(0^-), \quad \textit{for } i < n, \\ & n = \min\left\{a_0 \leq j \leq N : C_k = 0 \textit{ for all } k > j\right\}. \end{aligned} \tag{3.25}$$

*Then, the optimal solution to* (3.25) *is* $\{C_k^*, k = a_0, \ldots, N\}$ *given by*

$$\begin{aligned} C_{a_0}^* &= \min\left\{Q_{a_0}^s(0^-) + \kappa T, C'\right\} \\ \text{and } C_k^* &= \min\left\{Q_k^s(0^-),\ \left(C' - \sum_{m=a_0}^{k-1} C_m^*\right)^+\right\},\ k = a_0+1, \ldots, N. \end{aligned} \tag{3.26}$$

Under Assumption 9, the above problem admits a simple solution where the trader only applies this continuous submission of market orders at rate $\kappa$ at the best-ask queue at price $a_0$, and then submits a block order (as needed) to deplete higher price level queues at $T$. This is the cheapest price at which the trader can accumulate up to $\kappa T$ shares. A consequence of Lemma 4 is that the best-bid and the best-ask remain equal to $(b_0, a_0)$ for all $t \in [0, T)$, which simplifies the determination of the limit order placement decision, $S_L \in [0, C]$.

**Lemma 5** (Limit Orders). *In the optimal solution of problem* (3.17)–(3.24)*,*

$$S_L = \min\left\{\mu_{b_0}^s \left(T - \frac{1}{\gamma}\log\left(1 + \frac{\gamma}{\mu_{b_0}} Q^0(0)\right)\right)^+, C\right\}. \tag{3.27}$$

The above expression is intuitive, and crucially depends on the quantity $t_{\text{drain}}$, which is defined as $t_{\text{drain}} := \frac{1}{\gamma}\log\left(1 + \frac{\gamma}{\mu_{b_0}} Q^0(0)\right)$. It is derived from a transient analysis of a fluid queue with abandonments and is equal to the length of time required for the initial queue length $Q^0(0)$ to get depleted either due to cancellations or trades (service completions); this is increasing in the initial queue length and decreasing in the trading rate $\mu_{b_0}$ and the cancellation rate $\gamma$.

The next theorem characterizes the optimal strategy.

**Theorem 5** (Optimal Policy). *Fix the target size* $C > 0$*, execution horizon* $T > 0$*, and consider an arbitrary initial condition* $Q(0) \in \mathcal{Q}^{eq}$*. The optimal execution policy for* (3.17)–(3.24) *is the following:*

*(a) set the limit order execution quantity $S_L$ according to* (3.27)*;*

*(b) for $C' = C - S_L$, set the market order execution quantities $C_{a_0}, C_{a_0+1}, \ldots, C_N$ according to* (3.26)*;*

*(c) for $i = a_0$ and $C_{a_0}$ specified above, set the market order execution trajectory $\{S(t) : t \in [0, \tau_{a_0})\}$ according to* (3.13)–(3.15)*;*

*(d) for $i = a_0 + 1, \cdots, N$, according to Lemma 4, $\tau_i = \tau_{a_0} \leq T$. That is, market order executions at higher prices happen with block trades at $t = \tau_{a_0}$. We will refer to this aggregate block as the "cleanup" trade.*

In Part (c), the solution uses the infinitesimal $\varepsilon > 0$ to denote the minimum queue length to be maintained in $Q^s_{a_0}$ while submitting a continuous stream of market orders (i.e., service completions) at rate $\kappa$.

## 3.5. A Microstructure Market Impact Cost Model

In this section, we exploit the solution of the execution problem studied thus far in order to propose a microstructure market impact model. Such a model estimates the trading cost of an order as a function of microstructure limit order book variables, including, for example, real-time measurements of queue lengths and trading rates. We will propose a series of approximations that will yield a parsimonious microstructure market impact model that can be easily and robustly estimated through trade data.

The optimal value of the control problem studied in the previous two sections provides

an estimate of the cost of purchasing $C$ shares in $T$ time units. given by

$$\begin{aligned}\text{Total cost} &= p_{b_0} \cdot S_L + p_{a_0} \cdot C_{a_0} + \sum_{i=a_0+1}^{N} p_i \cdot C_i \\ &= (p - s/2) \cdot S_L + (p + s/2) \cdot C_{a_0} + \sum_{k=1}^{N-a_0} (p + s/2 + k\delta) \cdot C_{a_0+k} \\ &= (p + s/2) \cdot C - s \cdot S_L + \sum_{k=1}^{N-a_0} k\delta \cdot C_{a_0+k},\end{aligned} \tag{3.28}$$

where $p$ is the arrival price, i.e., the mid-price at the start time of the execution, and the last expression accounts for the execution cost relative to the (contra side or far side) price $p + s/2 = p_{a_0}$. The implementation shortfall, or average purchase price relative to the arrival price, is

$$\overline{IS} \triangleq \frac{\text{Total cost}}{C} - p = s/2 - s \cdot \frac{S_L}{C} + \sum_{k=1}^{N-a_0} k\delta \cdot \frac{C_{a_0+k}}{C}. \tag{3.29}$$

In this formula, the first term accounts for the cost relative to the best-ask price $p_{a_0}$ (the far side), which is half the spread ($s/2$) above the mid-price $p$. The second term then subtracts the spread for the shares traded using limit orders at the lower price $p_{b_0} = p_{a_0} - s$. The final term adds price increments (a multiple of the tick size) for the higher priced queues that were used in the cleanup trade. In order to simplify the subsequent empirical analysis, we will make several approximations to the final two terms:

(i) The limit order cost compensation term depends on $S_L = \min\left\{\mu^s_{b_0}(T - t_{\text{drain}})^+, C\right\}$. We will disregard cancellations and approximate the draining time $t_{\text{drain}}$ of the orders posted on the near side of the market prior to $t = 0$ by $t_{\text{drain}} \approx Q^0(0)/\mu^s_{b_0}$. Subsequently, we approximate $S_L$ as follows

$$S_L \approx \min\left\{\left(\mu^s_{b_0} T - Q^b_{b_0}(0)\right)^+, C\right\}.$$

(ii) For the cleanup cost term, we will first assume that the stationary queue lengths $\bar{Q}_i^s$, $a_0 < i \leq N$, as defined in Assumption 7, are all equal to some value $\bar{Q}^s$.[6] In that case, it follows from Lemma 4 that $C_{a_0+k} = \bar{Q}^s$ for $0 < k < n$, where

$$n := \left\lceil \frac{(C' - C_{a_0})^+}{\bar{Q}^s} \right\rceil = \left\lceil \frac{(C - S_L - Q^s_{a_0}(0) - \kappa T)^+}{\bar{Q}^s} \right\rceil \tag{3.30}$$

denotes the number of additional price levels needed in the cleanup trade. We will further simplify the expression by dropping $S_L$ from its calculation, i.e., we set $n \approx (C - Q^s_{a_0}(0) - \kappa T)^+ / \bar{Q}^s$, and subsequently approximate the average price penalty per share due to market order executions relative to the far side to be

$$\frac{\sum_{i=0}^{n} i\delta \cdot \bar{Q}^s}{C_{a_0} + n\bar{Q}^s}. \tag{3.31}$$

The effect of $C_{a_0}$ diminishes as $n$ increases. When $n$ is large, the average price per share in (3.31) can further be approximated by

$$\frac{n+1}{2}\delta \approx \frac{\delta}{2} \cdot \frac{(C - Q^s_{a_0}(0) - \kappa T)^+}{\bar{Q}^s} + \frac{\delta}{2}.$$

Combining (i)-(ii), the resulting simplified expression of the implementation shortfall is

$$\overline{IS} = s/2 - s \cdot \frac{\min\left\{\left(\mu^s_{b_0} T - Q^b_{b_0}(0)\right)^+, C\right\}}{C} + \frac{\delta}{2} \cdot \frac{(C - Q^s_{a_0}(0) - \kappa T)^+}{\bar{Q}^s} + \frac{\delta}{2}. \tag{3.32}$$

This expression depends on the microstructure variables such as trading rates on either side

[6] This is certainly an idealization. Typically, one would expect to see the limit order arrival rates $\lambda_i^s$ increase with price levels $i$, which then suggests $\bar{Q}_i^s := \lambda_i^s/\gamma$ should also increase with $i$. Nevertheless, we find in the empirical tests that using a uniform estimate of the stationary queue lengths performs reasonably well.

of the book, queue depths, spread, tick size, as well as the trade quantity and time horizon. Specifically,

(a) *Effect of limit orders:* The execution cost is decreasing in $S_L$, the volume that can be traded using limit orders. The latter is decreasing in the queue length on the near side of the book, $Q^b_{b_0}(0)$ (the bid side when buying, or ask side when selling), and is increasing in the the arrival rate of market orders to the near side (market orders to trade against the trader's posted limit orders), and in the execution horizon $T$. The expression for $S_L$ also indicates that the execution cost will be decreasing in the cancellation rate, although this dependence has been suppressed in the simplified cost formula. The limit order effect is independent of the trade quantity $C$ (assuming the latter is larger than $S_L$).

(b) *Market order effect at the top-of-book:* This depends on $C - S_L$, the residual quantity to be traded using market orders, and on $Q^s_{a_0}(0) + \kappa T$. The latter is increasing in the displayed depth $Q^s_{a_0}(0)$, the time horizon $T$, and the continuous trading rate $\kappa$ that, as discussed earlier, captures the rate at which one can continuously trade with market orders at a given price level without depleting the respective queue and moving the price.

(c) *Market orders at higher prices:* the residual quantity that needs to get executed at higher price levels is decreasing in $S_L$ (see (a)), $Q^s_{a_0}(0)$, $\kappa$, and $T$ (see (b)). Its effect is inversely proportional to the equilibrium depth $\bar{Q}^s$ in each of these queues, since that is used to compute the number of price levels $n$ that the trader will have to deplete.

## 3.6. Empirical Results

The microstructure market impact model of equation (3.32) identifies several important microstructure variables that may affect execution costs. While this model was based on a number of simplifying assumptions, it is our belief that these variables are nevertheless important. In order to demonstrate this, in the remainder of this chapter, we will calibrate this model using a proprietary dataset of algorithmic trades executed in the US equities market in the third quarter of 2013. Specifically, we will calibrate weights for the different microstructure variables identified in (3.32) via a regression analysis, and then validate that the resulting microstructure market impact model can help to explain more of the variability in observed trading costs.

Our data set consists of short time horizon slices of executions arising from algorithms based on TWAP, VWAP, and POV[7] policies. The execution logic used in those trades differs from the optimal policy derived in our stylized analysis in Section 3.4. Nevertheless, our findings will indicate that the microstructure market impact model leads to improved statistical fits, specifically in explaining the realized costs of execution in this dataset (attribution), when compared with conventional "macro" market impact models. Moreover, the coefficients of the explanatory variables postulated by our analysis are significant and have the right signs. The microstructure market impact model also exhibits improved predictive statistical accuracy, e.g., when used to make real-time predictions of future trading costs based on available information at the beginning of each trade.

[7]See, for example, Sotiropoulos (2013) for a description of these policies.

### 3.6.1. The Dataset

We use a proprietary dataset of US equities trades from July to September of 2013. This dataset is itself a random sample of a larger set of algorithmic orders executed over that time period. For each parent order (e.g., a full day execution according to the VWAP strategy), the data is summarized in 1-minute intervals. For each such interval we have execution statistics as well as measurements of various limit order book variables. The data has 980,000 active trade records (i.e., 1-minute summaries of execution activity), and represents a sample of 1,800 different securities.

Most of the analysis is performed in rolled-up 5-minute slices. Parent orders that lasted less than 5 minutes or parent order residuals that lasted less than 5 minutes are discarded. Intervals over which there were no executions are also discarded. We further filter according to the following criteria: (a) keep only slices that correspond to VWAP, TWAP, and POV strategies;[8] (b) remove orders for illiquid securities that have an average daily trading volume lower than 300,000 shares; (c) discard the last slice of each parent order to avoid special considerations and cleanup logic associated with the respective algorithmic strategy, apart from POV orders; (d) discard slices in the opening 15 minutes of the trading day, 9:30am–9:45am, and the last 15 minutes of the day, 3:45pm–4:00pm; (e) discard slices for which the realized implementation shortfall exceeds 200 basis points, where the daily volatility within the period exceeds 4%, or where the trade volume exceeded 5 times the volume of the immediately preceding slice; (f) restrict attention to slices with realized participation rate[9] between 1% and 30%. Table 3.1 reports monthly descriptive statistics of the filtered dataset.

[8]Such strategies tend to follow a fairly consistent rate of trading over short periods of time. The composition of the sample after the various filters were applied was roughly uniform across the three strategies and across months.

[9]The participation rate is the ratio of the execution quantity of the slice over the total volume traded in the corresponding time interval by all market participants.

| | JUL 2013 | AUG 2013 | SEP 2013 |
|---|---|---|---|
| **Sample Size** | | | |
| 5min Slices | 27,760 | 30,054 | 29,226 |
| Parent Orders | 3,396 | 3,607 | 3,882 |
| Distinct Securities | 988 | 896 | 885 |
| **Characteristics** | | | |
| Average Daily Volume (shares) | | | |
| mean | 3,014,000 | 2,595,000 | 2,509,000 |
| 3rd quantile | 2,585,000 | 2,689,000 | 2,626,000 |
| 1st quantile | 554,300 | 578,500 | 544,000 |
| Size of 5min Slices (shares) | | | |
| mean | 1,294 | 1,043 | 849 |
| 3rd quantile | 1,000 | 1,000 | 700 |
| 1st quantile | 81 | 100 | 82 |
| # 5min Slices in Parent Order | | | |
| mean | 8.2 | 8.3 | 7.5 |
| 3rd quantile | 10 | 9.5 | 8 |
| 1st quantile | 1 | 1 | 1 |
| Average Queue Length | | | |
| mean | 10,280 | 21,730 | 17,750 |
| 3rd quantile | 2,278 | 4,078 | 5,148 |
| 1st quantile | 434 | 477 | 536 |
| Realized Participation Rate | | | |
| mean | 9.60% | 9.40% | 8.39% |
| 3rd quantile | 17.70% | 16.20% | 14.19% |
| 1st quantile | 2.20% | 2.26% | 1.90% |
| Price ($) | | | |
| mean | 46.80 | 38.16 | 41.41 |
| 3rd quantile | 57.41 | 52.23 | 51.64 |
| 1st quantile | 15.35 | 13.31 | 13.33 |
| Spread ($) | | | |
| mean | 0.031 | 0.025 | 0.025 |
| 3rd quantile | 0.032 | 0.028 | 0.024 |
| 1st quantile | 0.010 | 0.010 | 0.010 |
| Daily Volatility | | | |
| mean | 2.23% | 1.90% | 1.94% |
| 3rd quantile | 2.39% | 2.31% | 2.34% |
| 1st quantile | 1.03% | 0.97% | 0.90% |
| Implementation Shortfall (bps) | | | |
| mean | 3.04 | 3.09 | 3.48 |
| 3rd quantile | 7.25 | 7.86 | 7.19 |
| 1st quantile | (2.62) | (2.53) | (1.84) |

**Table 3.1:** Descriptive statistics of the filtered dataset, aggregated into 5-minute slices. *Average queue length* represents the aggregated per side, time-averaged queue length at the best-bid or best-ask over the 5-minute interval. *Price* is the average trading price. *Implementation Shortfall (bps)* = (average trading price - arrival price)*side/arrival price$*10^4$; arrival price is the mid-price at the beginning of the respective 5-minute slice. The above are straight arithmetic averages as opposed to volume or notional weighted. (See Section 3.6.3)

### 3.6.2. Calibration of Auxiliary Model Parameters

There are three quantities in the market impact equation (3.32) that are not directly observable in the data: the equilibrium queue length $\bar{Q}^s$, the effective tick size $\delta$, and the rate of continuous trading $\kappa$.

The parameter $\kappa$ captures the rate at which one can execute with a continuous stream of market orders at the best-ask without causing any price change. Motivated by Assumption 9 and the discussion after it, we will think of $\kappa$ as a constant multiple of market order rate $\mu^b$. Specifically, we postulate that $\kappa$ can be expressed in the form of $\theta \cdot \mu$, where $\mu$ is the nominal trading rate and $\theta$ is a parameter between 0 and 1. We assume that $\theta$ is the same on the bid side and ask side of the book, and across all securities.

Returning to our dataset, we identify the set of slices for which: (a) the average queue length on the far side (i.e., the ask when buying and the bid when selling) was small, specifically less than or equal to 1/3 of the nominal queue length for the corresponding security; and (b) there was no price impact, i.e., the respective price level did not change. For each such slice we know the quantity that was executed as part of that order. We also generate a forecast for the nominal trading rate $\mu$. We first estimate the fraction of the total daily volume that is forecast to trade over the corresponding time interval, and then re-scale by the average daily volume of the corresponding security.[10] The trading rate estimate $\mu$ is set equal to half the forecast volume. The ratio of the executed quantity by the slice and of the corresponding forecast provides a point estimate for $\theta$ that is normalized relative to stock-specific characteristics. We average these estimates for each month and report the sample estimates together with the standard errors in Table 3.2. The estimated parameter can be interpreted as follows: over short time durations, one could trade at a rate that is 10% of

[10]The forecast makes use of a cross-sectional liquidity profile depicted in Figure B.1 in the Appendix.

| | JUL 2013 | AUG 2013 | SEP 2013 |
|---|---|---|---|
| Critical ratio $\theta_{\text{month}}$ | 0.112 (0.006) | 0.104 (0.004) | 0.091 (0.006) |

**Table 3.2:** Estimates of the critical ratio of trading rate to nominal volume for July-September 2013.

the bid volume or ask volume, respectively, or, equivalently, at a 5% participation rate while avoiding any price impact. The order of magnitude of this estimate seems plausible but its precise value is likely to be slightly optimistic, especially for less liquid securities as well as securities that trade with few shares at the best-bid and best-ask.

For the equilibrium queue length $\bar{Q}^s$ and the effective tick size $\delta$, we proceeded as follows. Our dataset contains execution information for the trades described earlier, and we also have access to Trade-And-Quote (TAQ) data for each of the securities included in the dataset over the period of July to September of 2013. Our dataset does not include depth of book information, i.e., information about the price levels and the corresponding queue lengths at the price levels that are not at the best-bid and best-ask price levels at a given point in time. As a result we did not have access to information that would allow us to estimate directly the queue length $\bar{Q}^s$, but instead we approximated it as the average of the queue lengths at the best-bid and best-ask, time averaged over the time interval of each 5-minute execution slice. Similarly, the effective tick size $\delta$ is meant to capture the change in price necessary to accumulate $\bar{Q}^s$ shares in the limit order book. Since this was not observable, we will use the volatility, $\sigma^*$ as a proxy for the tick size $\delta^*$; $\sigma^*$ is the volatility estimate based on intraday data for the time interval of the respective slice and accounts for the strong time-of-day pattern exhibited by the intraday volatility profile.

### 3.6.3. Estimation of the Microstructure and "Macro" Market Impact Models

**Microstructure Market Impact Model (In-Sample Regressions).** We start by estimating the microstructure market impact model in equation (3.32) using a linear regression analysis. Let $IS_k$ denote the implementation shortfall of the $k$th observation (5-minute slice) in the trade data described in Section 3.6.1. Implementation shortfall is defined as the normalized difference between the average execution price and the arrival price, denoted as $P_k$ and $P_k^0$, respectively. It is expressed in basis points. The arrival price is defined as the mid-price, i.e., the average between the best-bid and best-ask prices at the start time of the slice. The start and end times include millisecond timestamps. Specifically,

$$IS_k := (P_k - P_k^0)/P_k^0 \cdot d_k \cdot 10^4,$$

where the trade direction $d_k = 1$ for orders to buy and $d_k = -1$ for orders to sell. Normalizing both sides of (3.32) by the arrival price we get that

$$IS = \frac{1}{2} \cdot s^* - \frac{\min\left\{C, \left(\mu_{b_0}^s T - Q_{b_0}^b(0)\right)^+\right\}}{C} \cdot s^* + \frac{1}{2} \cdot \frac{(C - Q_{a_0}^s(0) - \kappa T)^+}{\bar{Q}^s} \cdot \delta^* + \frac{1}{2} \cdot \delta^*, \quad (3.33)$$

where $s^* := s/p \cdot 10^4$, $\delta^* := \delta/p \cdot 10^4$ are the normalized spread and tick size, respectively. Define

$$R^L := \frac{\min\left\{C, \left(\mu_{b_0}^s T - Q_{b_0}^b(0)\right)^+\right\}}{C}, \quad R^M := \frac{(C - Q_{a_0}^s(0) - \kappa T)^+}{\bar{Q}^s}, \quad (3.34)$$

for the price adjustments due to limit order executions and market orders at higher price levels, respectively. Expressions (3.33)–(3.34) are written for buy orders. The corresponding expressions for sell orders would replace in the first term $\mu_{b_0}^s$ with $\mu_{a_0}^b$ and $Q_{b_0}^b(0)$ with $Q_{a_0}^s(0)$,

| | JUL 2013 | AUG 2013 | SEP 2013 |
|---|---|---|---|
| (intercept) | | | |
| coefficient | -0.6888*** | -0.6941*** | -0.5832** |
| std. error | 0.1232 | 0.1140 | 0.1076 |
| spread (bps): $s^*$ | | | |
| coefficient | 0.3187*** | 0.3905*** | 0.3950*** |
| std. error | 0.0069 | 0.0077 | 0.0070 |
| limit order: $R^L s^*$ | | | |
| coefficient | -0.3027*** | -0.3415*** | -0.3658*** |
| std. error | 0.0107 | 0.0100 | 0.0099 |
| add. tick to pay: $R^M \sigma^*$ | | | |
| coefficients | 0.0991*** | 0.1480*** | 0.1486*** |
| std. error | 0.0234 | 0.0225 | 0.0348 |
| tick size: $\sigma^*$ | | | |
| coefficients | 2.3238*** | 1.8508*** | 2.4290*** |
| std. error | 0.1098 | 0.0997 | 0.0996 |
| R-squared | 9.91% | 10.62% | 13.48% |

Significance: *** p<0.001, ** p<0.01, * p<0.05

**Table 3.3:** Monthly linear regression results for microstructure market impact model of (3.35).

in the second term $Q^s_{a_0}(0)$ with $Q^b_{b_0}(0)$ and $\bar{Q}^s$ with $\bar{Q}^b$. We will estimate the following linear model:

$$IS = \beta_0 + \beta_1 \cdot s^* + \beta_2 \cdot (R^L s^*) + \beta_3 \cdot (R^M \delta^*) + \beta_4 \cdot \delta^*. \tag{3.35}$$

The regression results can be found in Table 3.3. We find consistently good performance for our model, represented by the high $R^2$ values (see discussion in next section), the fact that the coefficients are all statistically significant, and that the signs of the coefficients are all in line with our predictions. The month-to-month variability is partially due to the modest sample size and variations in the set of securities and parent orders included in our data set as well as variations in market conditions. If, instead of lower bounding the realized participation rate by 1%, we only allowed slices whose realized participation rate was greater than 3%, then the explanatory power of the model increased to an $R^2$ of 12.30%, 11.94% and 15.45% for July, August and September, respectively.

| | JUL 2013 | AUG 2013 | SEP 2013 |
|---|---|---|---|
| (intercept) | | | |
| coefficient | 0.3204*** | 0.5495*** | 0.7799*** |
| std. error | 0.1238 | 0.1148 | 0.1091 |
| (percent of market vol.)$\cdot\sigma^*$ | | | |
| coefficients | 10.3835*** | 9.0038*** | 9.5916*** |
| std. error | 0.6445 | 0.6067 | 0.6922 |
| volatility: $\sigma^*$ | | | |
| coefficients | 1.5498*** | 1.4778*** | 1.9781*** |
| std. error | 0.1127 | 0.1026 | 0.1046 |
| R-squared | 3.24% | 3.02% | 3.75% |

Significance: *** p<0.001, ** p<0.01, * p<0.05

**Table 3.4:** Monthly linear regression of benchmark model in (3.36) with $\alpha = 1$ (linear).

**Benchmark "Macro" Market Impact Model.** Most transient market impact models in the literature express the execution cost as a function of the normalized size of the order, expressed as a percentage of the overall volume that trades in the market in the respective time interval, and suggest the use of functions of the form:

$$IS = \beta_0 + \beta_1 \cdot (\text{Percent of Market Vol.})^{\alpha} \sigma^* + \beta_2 \cdot \sigma^*, \tag{3.36}$$

where typically $\alpha = 0.5$ or $1$.[11]

Table 3.4 and 3.5 illustrate the quality of these fits. Note that, as for the microstructure market impact model estimate, $\sigma^*$ is the volatility estimate using intraday data for the time interval of the respective slice. A simpler model would use a static volatility estimate, again prorated to the duration of the slice, but independent of the time-of-day. This reduces the explanatory power of the "macro" models from around 3% to about 1%, underscoring the

[11] We have examined a finer grid of $\alpha = 0.1, 0.2, \ldots, 1$. The performance does not vary much with the selection of $\alpha$, and $\alpha = 0.5$ or $\alpha = 1$ oftentimes have the best performance. We focus on explaining the market impact of short duration slices and we will disregard the decay kernel that is sometimes included in transient market impact models.

| | JUL 2013 | AUG 2013 | SEP 2013 |
|---|---|---|---|
| (intercept) | | | |
| coefficient | 0.3235** | 0.5480*** | 0.7839*** |
| std. error | 0.1238 | 0.1148 | 0.1091 |
| (percent of market vol.)$^{0.5} \cdot \sigma^*$ | | | |
| coefficients | 6.4110*** | 5.5267*** | 5.8011*** |
| std. error | 0.3913 | 0.3685 | 0.4132 |
| volatility: $\sigma^*$ | | | |
| coefficients | 0.7626*** | 0.8033*** | 1.2844*** |
| std. error | 0.1429 | 0.1320 | 0.1367 |
| R-squared | 3.27% | 3.04% | 3.77% |

Significance: *** p<0.001, ** p<0.01, * p<0.05

**Table 3.5:** Monthly linear regression of benchmark model in (3.36) with $\alpha = 0.5$ (square root).

importance of incorporating this effect.

**Cross-Validation.** Next we compare the out-of-sample performance of our model against that of the benchmark models. We perform a 3-fold cross-validation using the three monthly samples of data from July to September in 2013.[12] We proceed as follows: in each round, we select one monthly sample among the three as the testing data. On the data of the other two months, our model, the linear benchmark model, and the square root benchmark models are fit. Then, the calibrated models are applied to the test set to evaluate how much of the variability in market impact can be explained by each model. Three rounds of training and testing are performed by rotating through the different months as the test set. Finally, the prediction performance of each model takes an average among the three rounds of cross-validation.

When evaluating the out-of-sample accuracy of the different models, we compare their

[12]Usually a $k$-fold cross-validation requires dividing all data randomly into equal size subsets. Here we take the natural monthly division of data instead. We expect the result, in particular, the comparison between the two models, be of similar quality when we trisect randomly.

mean squared error with that of the mean predictor to define a generalized $R^2$ as:

$$\text{generalized } R^2 := 1 - \frac{\text{Mean Squared Error (selected model)}}{\text{Mean Squared Error (mean predictor)}}. \tag{3.37}$$

For that purpose, there are two candidate mean predictors to use: the mean of the train set, or the mean of the test set. Using the mean of the train set is more popular in the literature and has the interpretation that the mean predictor itself is a model that is trained together with other models on the train dataset in each round. In Table 3.6, we report the average generalized $R^2$ values based on both mean predictors.

We find that the microstructure market impact model has an average out-of-sample $R^2$ of around 11%, explaining a factor of 2.5 more of the out-of-sample variability in realized trading costs relative to the "macro" models when compared to the mean predictor; the "macro" market impact models had an average out-of-sample $R^2$ of around 3.1%. The performance improvement is consistent across the three separate test sets, and, as we will see below, fairly robust to various changes to the way we construct and estimate the microstructure market impact model. The microstructure model treats separately the limit order effect on the execution cost and suggests that measuring trade size as a multiple of queue depth is useful in explaining execution costs. The latter suggests a further segmentation of the data by security characteristics, which we will explore in the next subsection.

[13]The above analysis could be repeated to include orders that are traded at lower participation rates, i.e., below 1% which we used as a filter thus far. When including slices with realized participation greater or equal to .25%, the $R^2$ of the microstructure market impact model drops to 9%; the "benchmark" linear and square root models exhibit an $R^2$ of about 3%. When we fit a model exclusively to lower participation rates, say in the interval $[.25\%, 1\%]$, the microstructure model explains 4.4% of the realized cost variability, while the benchmark models explain 1% of the variability.

| | Model eq. (3.35) | Benchmark model eq. (3.36) | |
|---|---|---|---|
| | | $\alpha = 1$ | $\alpha = 0.5$ |
| Avg. out-of-sample $R^2$ (vs. predicted mean) | 11.03% | 3.11% | 3.12% |
| relative improvement | 0.00% | 255% | 254% |
| Avg. out-of-sample $R^2$ (vs. current mean) | 10.97% | 3.04% | 3.06% |
| relative improvement | 0.00% | 261% | 258% |

**Table 3.6:** Average out-of-sample $R^2$ and relative improvements for a 3-fold cross-validation comparison between our model and the linear/square root benchmark models under two mean predictors. [13]

### 3.6.4. Robustness Checks

**Order & Security Segmentation.** First, we grouped the dataset into three sets depending on their realized participation rate. We used the following segments: $[1\%, 10\%]$, $(10\%, 20\%]$, $(20\%, 30\%]$. Table 3.7 reports the out-of-sample performance[14] of the microstructure model and the linear/square root benchmark models in each of these segments. The microstructure model continues to statistically outperform the "macro" benchmark models for all of these trade groups, but the explanatory power of all models improves as the participation rate increases, since, as expected, in these settings the statistical signature of the trading slice is likely to be a key driver of the price movement.

Second, following on the observation of the previous subsection, we segmented the trade observations according to the stock characteristics, and specifically, their average daily volume (ADV) and average queue length. We divided the dataset according to the 33% and 66% ADV percentiles, and further segmented according to average queue length at the 30%, 60%, and 90% percentiles. Table 3.8 reports the out-of-sample results based on these 12 segments of the data. For 9 out of the 12 segments we have enough observations to perform cross-

[14] Out-of-sample results in this section are with respect to the predicted mean unless otherwise indicated.

| | Model eq. (3.35) | Benchmark model eq. (3.36) | | Sample size |
|---|---|---|---|---|
| | | $\alpha = 1$ | $\alpha = 0.5$ | |
| Percent of market vol. | | | | |
| [1%,10%] | 8.82% | 1.87% | 1.89% | 55,337 |
| (10%,20%] | 14.10% | 5.34% | 5.21% | 19,974 |
| (20%,30%] | 15.08% | 4.23% | 4.24% | 11,729 |
| overall: [1%,30%] | 11.03% | 3.11% | 3.12% | 87,040 |

**Table 3.7:** Out-of-sample performance when clustering by market participation rate.

validation tests. Again, within each of these segments, the average out-of-sample $R^2$ of our model has consistently significant improvement over those of the "macro" models. Moreover, we see (as one would expect) that model accuracy improves as queue depth increases that correspond to settings where the queueing model used in our analysis may be more relevant. The results are qualitatively similar if we segment with respect to queue lengths expressed in notional dollars rather than shares.

Last, we examined the subset of the data trading in security names with low daily volumes. We restricted attention to securities with an average daily volume between 50,000 shares and 300,000 shares. Table 3.9 reports the out-of-sample performance of our model and the benchmark models based on the sample of these less liquid names. The explanatory power of all models improves. Moreover, the performance improvement of our model against the benchmark models becomes more significant.

**Effect of Nonlinearity.** The structural form of the microstructure model involves two non-linear terms that are not a concern when using the model to produce cost estimates or in attributing trade execution performance, but they may affect computational tractability in the context of an optimization model, either for stock selection or for scheduling how to execute a large trade during the course of a longer time horizon. A drastic simplification of

| | | | | | |
|---|---|---|---|---|---|
| | | Low depth | Mid depth | High depth | Ultra deep |
| Model eq. (3.35) | Low ADV | 6.26% | 10.23% | 17.14% | too few obs. |
| | Mid ADV | 5.38% | 8.12% | 12.62% | too few obs. |
| | High ADV | too few obs. | 5.56% | 10.32% | 24.84% |
| | | Low depth | Mid depth | High depth | Ultra deep |
| Model eq. (3.36) ($\alpha = 1$) | Low ADV | 2.37% | 3.28% | 5.10% | too few obs. |
| | Mid ADV | 2.23% | 2.64% | 4.62% | too few obs. |
| | High ADV | too few obs. | 3.03% | 3.84% | 6.64% |
| | | Low depth | Mid depth | High depth | Ultra deep |
| Model eq. (3.36) ($\alpha = 0.5$) | Low ADV | 2.39% | 3.25% | 5.13% | too few obs. |
| | Mid ADV | 2.27% | 2.63% | 4.59% | too few obs. |
| | High ADV | too few obs. | 3.10% | 3.90% | 6.68% |
| | | Low depth | Mid depth | High depth | Ultra deep |
| Sample size | Low ADV | 14,775 | 9,503 | 4,589 | 133 |
| | Mid ADV | 9,712 | 10,617 | 8,083 | 614 |
| | High ADV | 1,625 | 5,992 | 13,440 | 7,957 |

**Table 3.8:** Out-of-sample performance when clustering by (average daily volume, average queue length).

| | Model eq. (3.35) | Benchmark model eq. (3.36) | |
|---|---|---|---|
| | | $\alpha = 1$ | $\alpha = 0.5$ |
| Avg. out-of-sample $R^2$ (vs. predicted mean) | 23.26% | 4.72% | 4.91% |
| relative improvement | 0.00% | 393% | 374% |

**Table 3.9:** Out-of-sample performance for the sample of securities with low daily volumes.

the model would remove the non-linearities, as in

$$IS = \beta_0 + \beta_1 \cdot s^* + \beta_2 \cdot \frac{\left(\mu^s_{b_0} T - Q^b_{b_0}(0)\right)}{C} \cdot s^* + \beta_3 \cdot \frac{\left(C - Q^s_{a_0}(0) - \kappa T\right)}{\bar{Q}^s} \cdot \delta^* + \beta_4 \cdot \delta^*. \quad (3.38)$$

Using this simplified model in (3.38) in the cross-validation tests, we see that the out-of-sample $R^2$ of the microstructure model drops to an average of 8.19%, yet still outperforming the "macro" models; this comparison held across segments of the data by participation rates or security characteristics.

**Effect of Time Horizon.** The microstructure variables fluctuate over time, and one could expect that the model accuracy depends on the time horizon of the trade slices. Queue length measurements are likely to be more representative over shorter time intervals, but trading rate measurements will be more noisy over short time intervals. Table 3.10 summarizes our statistical results when instead of using 5-minute trade slices we organize the data sample in 1-minute slices, and illustrate that the statistical significance (out-of-sample) of the microstructure model improves in shorter horizons that may be relevant in the context of dynamic execution algorithms used to optimize over tactical order placement decisions. Tables 3.11–3.12 report the out-of-sample performance in segmented data samples of the 1-minute slices, and should be contrasted to Tables 3.7–3.8.

The explanatory power of these models improves if one adds lagged residuals of the past

| | Model eq. (3.35) | Benchmark model eq. (3.36) | |
|---|---|---|---|
| | | $\alpha = 1$ | $\alpha = 0.5$ |
| Avg. out-of-sample $R^2$ (vs. predicted mean) | 16.57% | 2.67% | 2.81% |
| relative improvement | 0.00% | 521% | 490% |
| Avg. out-of-sample $R^2$ (vs. current mean) | 16.52% | 2.61% | 2.75% |
| relative improvement | 0.00% | 533% | 501% |

**Table 3.10:** Out-of-sample performance for the sample of 1-min trade slices.

| | Model eq. (3.35) | Benchmark model eq. (3.36) | | Sample size |
|---|---|---|---|---|
| | | $\alpha = 1$ | $\alpha = 0.5$ | |
| Percent of market vol. | | | | |
| [1%,10%] | 13.53% | 0.94% | 0.96% | 73,166 |
| (10%,20%] | 19.24% | 2.26% | 2.26% | 40,631 |
| (20%,30%] | 21.51% | 3.59% | 3.59% | 19,830 |
| overall: [1%,30%] | 16.57% | 2.67% | 2.81% | 133,627 |

**Table 3.11:** Out-of-sample performance when clustering by market participation rate (1-min trade slices).

two periods (where each residual is the difference between the realized cost and the predicted cost). Their respective coefficients are positive and statistically significant, and they seem to capture short-term price momentum. The explanatory power improves by about 2% when explaining realized costs of 1-minute trading slices, and by about 0.6% for 5-minute slices. The "macro" model also improves by about 1% in terms of its explanatory power if one includes the lagged residual variables. One expects that similar improvements may be realized if one included short-term price signals that essentially added a short-term drift component in the regression models.

**Cost prediction versus attribution.** Market impact models are often used to compute pre-trade cost estimates that may be used as part of a portfolio selection process, or as part

| | | Low depth | Mid depth | High depth | Ultra deep | Overall |
|---|---|---|---|---|---|---|
| Model eq. (3.35) | Low ADV | 12.18% | 13.81% | 23.12% | too few obs. | |
| | Mid ADV | 9.41% | 10.84% | 18.78% | too few obs. | 16.57% |
| | High ADV | too few obs. | 3.91% | 20.74% | 28.98% | |

**Table 3.12:** Out-of-sample performance when clustering by (average daily volume, average queue length) (1-min trade slices).

of a dynamic trade execution algorithm. In such settings, the models are used to make cost predictions, e.g., at the beginning of a trading slice, and they use information available at that time, as opposed to contemporaneous information that is available in explaining realized costs. This includes snapshots of the queue lengths as well as trailing averages of the queue lengths and the bid side and ask side volume. Specifically, when making a prediction for a trading slice that commences at some time $t$, we will use exponentially smoothed trailing averages of the relevant limit order book variables computed over the duration of the previous 5-minute (or 1-minute) trading slice. We discard the first slice of each parent order in our dataset when we study the predictive accuracy of the market impact model, since itself was missing prior information needed for the above estimation; this removes 6.5% of the sample of 5-minute trade slices and 5.6% of the sample of 1-minute slices.

Table 3.13 reports the resulting average out-of-sample $R^2$ in comparison with the attributive models in Section 3.6.3. The drop in explanatory power is more significant in the microstructure model as opposed to the macro models, given that the former is using real-time information in a more nuanced way. However, in absolute terms, the microstructure model continues to significantly outperform the two benchmark models.

A similar comparison is reported in Table 3.14 where the various microstructure variables are replaced with historical forecasts, which may be practical in settings where real-time information is not readily available. We use the average monthly queue depth and spread for

| | Model eq. (3.35) | | Model eq. (3.36) ($\alpha = 1$) | | Model eq. (3.36) ($\alpha = 0.5$) | |
|---|---|---|---|---|---|---|
| | predictive | attributive | predictive | attributive | predictive | attributive |
| 5min | 8.20% | 11.07% | 2.26% | 2.82% | 2.25% | 2.84% |
| 1min | 11.93% | 16.80% | 1.99% | 2.62% | 2.27% | 2.76% |

**Table 3.13:** Out-of-sample performance using predictive estimates of average queue length, market volumes, and spread, based on the sample of 5-minute trade slices and the sample of 1-minute trade slices. "Predictive" refers to the model that is using information available at the beginning of each trade slice to estimate its cost. "Attributive" is the model that uses information over the slice, such as the realized participation rate, or the realized bid-side and ask-side volume. The attributive results differ from those in Tables 3.6–3.10 due to the additional filtering of the first trading slice of each parent order; similarly in Table 3.14.

| | Model eq. (3.35) | | Model eq. (3.36) ($\alpha = 1$) | | Model eq. (3.36) ($\alpha = 0.5$) | |
|---|---|---|---|---|---|---|
| | historical | attributive | historical | attributive | historical | attributive |
| 5min | 7.35% | 11.03% | 2.44% | 3.11% | 2.56% | 3.12% |
| 1min | 9.54% | 16.57% | 1.61% | 2.67% | 1.73% | 2.81% |

**Table 3.14:** Out-of-sample performance using monthly estimates of average queue length, market volumes, and spread, based on the sample of 5-minute trade slices and the sample of 1-minute trade slices.

the bid and ask side queues and the spreads, and we use 1/2 of the forecast interval volume for the bid and ask side rate of market orders. We continue to use the volatility forecast that corresponds to the time interval of each trading slice in our data set.

# Chapter 4

# Dynamic Matching Markets and an Application to Residential Real Estate

## 4.1. Introduction

In the residential real estate market, sellers arrive dynamically over time to put their units up for sale. These assets may differ in their attributes, including location, size, style, acreage, etc. Sellers themselves differ in their own financial constraints, carrying costs, and their delay tolerances, i.e., how long they are willing to wait until they sell their unit. Buyers arrive dynamically over time, differing in their preferences of house attributes, their budgets, and their delay tolerances. This market evolves sequentially, and is subject to other frictions, such as the fact that sellers and buyers can consider only a fraction of the entire market at any time, e.g., buyers cannot be simultaneously bidding for too many units. Both phenomena imply sellers and buyers face search friction and probably will experience delay, which results

in inventory of sellers that are incurred explicit carrying costs and inventory of buyers whose utilities are decreased as they spend more time searching. Buyer and seller decisions, i.e., how to price, which bid to accept, and whether to wait for better outcomes in the future, depend on the available inventory, its characteristics, the heterogeneity of buyers and sellers, the potential mismatch between buyers and sellers, and their beliefs for potential future arrivals of better units or less patient buyers, etc., and vice versa.

This chapter studies a microstructure model of this market, explicitly accounting its dynamics and the heterogeneity of buyers, sellers, and inventory. It investigates the tactical pricing/bidding decisions of the sellers and buyers in a dynamic, heterogeneous, and decentralized marketplace, with specific bearing on operational questions. It strives to answer systemic questions such as what explains the fact that similar units sell for different prices; how does the depth of the market and the time spent on the market by sellers and buyers depend on the supply and demand imbalance, the financial constraints, and the search friction; and tactical questions such as how much faster will a house sell if you lower the price by 5%; how should a seller or buyer interpret market conditions and quantify risks/rewards from observations, e.g., transaction history.

Specifically, we propose a stylized microstructure model of the residential real estate market. We analyze the market dynamics and its equilibrium under the simplifying approximation where buyers and sellers use linear bidding strategies. We motivate and characterize this near closed-form approximation of the market equilibrium, and show that it is asymptotically accurate. We provide numerical evidence in support of this approximation. Then with the gained tractability, we characterize steady-state properties such as market depth, price dispersion, and anticipated delays in selling or buying a unit. We characterize congestion and matching patterns for sellers and buyers, taking into account market dynamics, heterogeneity, and supply and demand imbalance manifested in the competition among buyers and sellers.

Furthermore, we show the effects of market primitives with comparative statics results.

In the sequel, we provide some detail about our modeling approach and results that we obtain.

We propose a sequential meeting, Nash bargaining microstructure model of the dynamic, heterogeneous, and decentralized market. The model evolves in discrete time over an infinite horizon. In each period, a new batch of random size of buyers and sellers arrive to the market, and consider whether to enter the market on the demand or supply side, respectively, as a function of the state of the market upon their arrival. Active buyers and sellers also make a decision whether to continue and stay in the market or leave (abandon). In our setting, the matching of buyers and sellers and the price formation occurs as follows: in each period, after buyers and sellers make entry/exit decisions, they contact potential matches under market frictions such as limited monitoring capability. The number of meetings between a buyer and a seller taking place in each period is determined by an aggregate matching function, which is an often used modeling tool in the economics literature, and is based on inventory of the two sides of the market. Under the standard assumptions of linear searching technology and random meeting mechanism, each active agent meets one potential match from the pool of active agents on the other side of the market in each period with a probability determined by the ratio of active buyers to active sellers in the current period. Then, when a meeting is formed, buyer and seller decides whether to trade and at what transaction price by Nash bargaining based on their dynamic valuations. These not only incorporate their heterogeneity in the nominal valuation of the good/service, but also their dynamic opportunity cost or continuation value of market participation in the future. After transacted buyers and sellers exit the market, all remaining agents suffer a delay penalty because of their failure to transact in the current period, and then carry over to the next period. Both buyers and sellers are trying to maximize discounted surplus.

With the aforementioned microstructure model, we obtain the equilibrium steady-state characteristics of market depth, price dispersion, and the optimal strategies for buyers and sellers in such dynamic matching markets. Specifically,

*Linear strategies.* We show that when the range of the valuations for buyers and sellers is small, bidding strategies are linear. Motivated by this result we study a market under an assumption that buyers and sellers always employ linear strategies. This assumption leads to a more tractable model, and is motivated by markets with "almost homogeneous" goods and agents, e.g., market of apartments in the same neighborhood. Numerical tests show the approximation is close to the true market equilibrium when the range of valuations is moderate.

*Equilibrium characterization.* Given the arrival features on the supply and demand sides and the market primitives, the strategies of buyers and sellers and the distribution of their types within the market's steady state is endogenous. Under linear strategies, we solve for the steady state equilibrium. For uniformly distributed valuation distributions, we obtain near closed-form characterization of the market equilibrium, in which delay increases with sellers' costs and decreases with buyers' valuations as a power law, or exponentially in some cases.

Besides insights on congestion, our results also characterize the matching pattern between buyers and sellers in settings where dynamics play a role, and clarify the effect of attractiveness (meaning having a low cost as a seller or having a high valuation as a buyer) in several aspects that would be otherwise ambiguous. The resulting assortativeness in dynamic matching patterns and surplus under the aligned preference structure add to the search and matching literature in economics.

*Symmetric market.* If the market is "symmetric" with respect to buyers' and sellers' primitive parameters, we can characterize the steady state equilibrium in closed form, as well

as establish its existence and uniqueness. As a result, we can provide a series of comparative statics results on how the equilibrium would react to different kinds of changes in market primitives. For example, how would the meeting technology, the prevailing interest rate, or the participation cost affect depth, price dispersion, and distribution over types in the inventory.

To conclude, this chapter makes the following key contributions: (a) We provide a stylized microstructure formulation of the residential real estate market that leads to tractable analysis of such a stochastic and dynamic matching market. It integrates the heterogeneity in the market, the dynamic entry/exit of the sellers and buyers, and their transaction behavior - that is, their mechanism of fragmented meeting, price formation, and self-interested decision making. (b) We propose a linear approximation method that yields tractable equilibrium analysis and in some cases near closed-form characterization of the market's steady state equilibrium. We justify this approximation by demonstrating that it is asymptotically accurate and is also numerically close to the true market equilibrium when the market is "almost homogeneous". For downstream analysis, the modeling and analytical insights described above provide the essential ingredients for formulating and solving optimal search/pricing problems for buyers and sellers in such dynamic matching markets. (c) We gain several insights into how such dynamic matching markets operate. Most importantly, we answer the questions of who would join/exit the market, what are the determinants of the depth of market, who would wait for how long in the market, and what is the distribution of valuations for buyers and sellers in the market. More subtly, we also investigate into the matching pattern between sellers and buyers when dynamics play a role. Finally, from a marketwise perspective we provide a series of comparative statics results in the symmetric case to shed light on how primitives such as meeting technology, interest rate, or participation cost, etc., would affect the operational metrics in the market's equilibrium.

**Literature review.** The modeling and analysis of a dynamic matching market lies in the interface of the economics literature on market design, the CS/OR literature on matching, the finance/OR literature on market microstructure, and the stochastic networks literature. This chapter leverages modeling approaches and tools from each of these areas, as well as the area of quantitative pricing and revenue management.

The matching and assignment literature is originated by the classical paper on marriage and college admission by Gale and Shapley (1962). This chapter relates to one strand of the matching literature that is pioneered by Becker (1973) and followed by Diamond (1982), Mortensen (1982), Shimer and Smith (2000), and others. These studies focus on matching models with structured preferences, e.g., aligned preferences. With such modeling advantage, important questions such as market efficiency (Hosios (1990), Shimer and Smith (2001a)), mechanism design (Shi (2001)), matching patterns (Shimer and Smith (2001b)), and comparative statics results on the effect of various market factors can be studied. See Rogerson et al. (2004) for a survey. In comparison, papers in the other strand of matching literature model preference in matching in more general ways, and mainly study the design of mechanisms to ensure stability and strategyproofness, for example, Roth (1985), Bogomolnaia and Moulin (2004), Abdulkadiroğluand et al. (2005), Pathak and Sethuraman (2011), etc.

There is an extensive body of modern developments that studies stylized dynamic matching and is closer to this work. First, there is a literature on the effect of search friction on matching markets that are particularly relevant to this paper, including for example, Mortensen and Pissarides (1994), Atakan (2010), Shimer and Smith (2000). They study models of markets with random matchings that are similar to the model studied in this chapter. In particular, Genesove and Han (2012) examines search and matching in the housing market. Again, Rogerson et al. (2004) provides an oversight of work until 2004. Second, Rubinstein and Wolinsky (1985), Gale (2000), and Satterthwaite and Shneyerov (2007) study

dynamic matching and bargaining in a general equilibrium framework. Finally, also relevant to the dynamic matching problem that we are looking at is the literature on dynamic auctions. See Bergemann and Said (2011) for a survey. In contrast to the economics literature mentioned here and above, the analysis of dynamic matching markets in this chapter focuses more on the operational issues, concerning for example the depth of market, delay in transaction, congestion pattern across types, etc., instead of on the economic issues such as efficiency, stability, or convergence to general equilibrium.

With this emphasis, our modeling and methodological work in this chapter adds some of the operational/tactical context to the previous dynamic matching models, guided by the following areas of studies in operations research/management.

This work is related to the growing literature on market microstructure studies of dynamic markets, which have aggravated toward the limit order book markets in the financial industry so far. The first example in operations research is Bertsimas and Lo (1998). Together with later developments by Almgren and Chriss (2001), Obizhaeva and Wang (2013), etc., this noticeable set of papers explicitly model the dynamics of the limit order book markets and address the problem of how to optimally execute a trade by dividing it into smaller child orders that are tactically directed to the market at optimized price levels and time points. More detailed reference can be found in the literature reviews from the other two chapters.

When dealing with the markets we look at in this chapter where dynamics and delay considerations play a role, we in broad terms see such markets as queueing systems with economics and service rules determined by the decentralized transaction behavior of buyers and sellers therein. Cont et al. (2010) first made the connection of queueing and limit order book markets. Maglaras et al. (2014) (chapter 2) and Maglaras et al. (2015a) (chapter 3) have developed models of the limit order books as queueing networks and have shown that the analytical tools from queueing theory can be utilized to study in such markets optimal

execution decisions, determinants of price impact of trades, and marketwise liquidity fluctuations across platforms. As such this work borrows tools and insights from the broad literature on economics of queues. See Mendelson and Whang (1990), Afeche (2013), Maglaras et al. (2015b), and Hassin and Haviv (2003) for a survey. Finally, for other examples of matching problems in operations management, see Garnett and Mandelbaum (2000), Whitt (2006), Ostrovsky (2008), and René et al. (2009).

The remainder of this chapter is organized as follows. Section 4.2 describes our microstructure model of sequential meeting and Nash bargaining. We define the mean-field steady state equilibrium and provide solution to a system of ordinary differential equations characterizing the flow balance conditions in Section 4.3. We solve for the equilibrium by linearization of the dynamic value function and prove its asymptotic accuracy in Section 4.4. In Section 5, focusing on the simpler case of symmetric markets, we illustrate the effects of various market primitives; moreover, numerical experiments under the symmetric case suggest that the linear approximation is close to the true market equilibrium when valuation range is moderate.

## 4.2. The Dynamic Matching Market

We propose a sequential meeting, Nash bargaining microstructure model of a dynamic matching market, e.g., the residential real estate market. We first discuss our modeling of the heterogeneity in the market and assign buyers and sellers into different type-based groups. We then model and track the cumulative arrivals of various types (§4.2.1), model their transaction behavior as a combination of sequential meetings between potential matches and price formation upon meeting according to Nash bargaining between the two parties (§4.2.2), and subsequently describe the market dynamics. Based on these we, in broad terms, see such a market as a queueing system and use a deterministic fluid model (a "mean field" model) to

analyze it. Moreover, we focus on steady states and discuss this assumption later on (§4.2.3).

**Heterogeneity in nominal valuations.** We consider a market where buyers and sellers have heterogeneous (nominal) willingness to buy or sell an indivisible unit of a homogeneous good, and are penalized for delay at rates that are common on either side when they fail to transact. In the market time progresses in discrete periods of length $\delta$, over an infinite horizon.

Specifically, we assign buyers and sellers into different type-based groups according to their nominal valuations. We consider a continuum of valuation types indexed by $V_B, V_S \in [L, U] \subset \mathbb{R}^+$ on the buying side and the selling side, respectively. In every period, active buyers and sellers each pay a side-specific participation fee of $c_S, c_B > 0$, charged on the selling side and the buying side, respectively. Furthermore, the time value of money is common for every one with a discount factor $\beta > 0$.

Such a model can be seen as representative of a situation in which preferences are aligned on both sides and participation is not significantly discriminated among agents on the same side.

One comment concerning this setup is on the dimensions of heterogeneity to consider in the model. Buyers and sellers can prefer one match over another for various reasons. For example, sellers can be arriving at the market endowed with heterogeneous supplies, buyers can form idiosyncratic preferences and desires over different units, buyers and sellers can vary in time sensitivity in terms of carrying cost or search patience, etc. When dynamics play a role in the matching of heterogeneous buyers and sellers in the market, they tactically decide on whether to transact with a match now or to wait for a more preferred match in the future. It is the purpose of this model to capture the tension in this tradeoff while generating tractable analysis. We therefore only incorporate one dimension of heterogeneity in matching - that in the nominal valuation of the good/service in the market, as a starting point. In

this setup, a buyer in general prefers to match with a seller having low cost, and similarly, a seller prefers to match with a buyer having high valuation, and these nominal preferences are incorporated into their dynamic valuations on top of their dynamic opportunity cost or continuation value of market participation in the future. We leave it for later research to introduce into the model other interesting dimensions of heterogeneity, or to explore more involved preference structures.

### 4.2.1. Dynamic Arrival, Entry, and Exit

**Arrivals:** There are sequential arrivals of sellers and buyers, each supplying or demanding one unit of good/service. Each seller or buyer is exogenously given her valuation type $V_S, V_B \in [L, U] \subset \mathbb{R}^+$, which is an independent identically distributed (i.i.d.) draw from a population wise type distribution $G_S, G_B : [L, U] \mapsto [0, 1]$ for the sellers and the buyers, respectively. For sellers seeking to sell at a price at least higher than their willingness to sell, the higher is their type of valuation, the less room of profit there will be toward a certain buyer. Similarly, for buyers seeking to buy at a price at most as high as their willingness to buy, the lower is their type of valuation, the less profit they can present for the other side. Therefore, the sellers having low cost and the buyers having high valuation are the relatively more *attractive* participants of the market.

Besides the draws of valuation types, the interarrival times are also stochastic. We model the arrivals of sellers and buyers as Poisson processes with rates $\lambda$ and $\alpha\lambda$, respectively, where $\alpha, \lambda > 0$.

Upon arrival each seller or buyer can decide either to join the market or to take an outside alternative with a normalized zero payoff according to her own interest. These join/exit decisions are based on the tradeoff between the potential profit from finding a match in the market and the uncertain cost of search given the frictions therein. Therefore such decisions

are type dependent. The joining populations will be different from the arriving populations on both sides.

**Entry/exit:** We consider a model with free entry/exit in which each agent in the market can decide either to stay or to exit in their own interests. We assume agents are infinitely-lived and do not leave for exogenous reasons.

At the beginning of periods, a newly arrived agent decides to join, and an agent that has already been participating for at least one period decides to stay, if and only if her continuation value of participating in the market exceeds the zero payoff of an alternative outside option. Each agent incurs a side-specific participation fee $c_S$ or $c_B$ for each period she decides to join/stay. Such agents form the population of active agents in the period. Again the active populations on the two sides will be different from either the arriving populations or the joining populations.

### 4.2.2. Sequential Meeting, Nash Bargaining

In our setting, the matching of buyers and sellers and the price formation occurs as follows:

**Meeting:** In each period, after buyers and sellers make entry/exit decisions, they contact potential matches. We assume that there exist certain market frictions in meeting, e.g., imperfect information about potential trading partners, slow mobility, large markets with limited monitoring capacities, etc. So, though under perfect condition the number of total meetings should be the product of the numbers of active agents on two sides, the truth is the actual chance of meeting will be far more limited. That is, meeting is fragmented and each agent can only see a small fraction of the other side within one period.

A usual modeling tool in economics literature used to capture the influence of such frictions without explicit reference to the complex sources is an aggregate matching function (see Petrongolo and Pissarides (2001) for a survey). It determines the number of meetings of

buyer and seller happening in one period based on inventory of the two sides of the market.

In particular, we assume agents meet with potential matches from the other side in a sequential, one-to-one manner. Each period will see a certain number of meetings formed, each between a different pair of one seller and one buyer. The number of meetings is determined by a matching function $M = m(T_S, T_B)$ where $T_S, T_B$ are the number of active sellers and buyers, respectively, at the beginning of the period.

Furthermore, we assume a random meeting mechanism. The chance of meeting will be randomly allocated among agents on either side, so each seller in one period has a meeting probability of $q(T_S, T_B) := M/T_S = m(T_S, T_B)/T_S$ while each buyer in one period has a meeting probability of $h(T_S, T_B) := M/T_B = m(T_S, T_B)/T_B$.

Finally, with the standard assumption of a linear searching technology, the matching function $m(T_S, T_B)$ would be homogeneous of degree 1 or featuring constant return to scale. In this case, each active agent meets one potential match from the pool of active agents on the other side of the market in one period with a probability determined by the ratio of active buyers to active sellers $\theta := T_B/T_S$ in the current period. Hence, we have $q(\theta) := m(1, \theta) = q(T_S, T_B)$ and $h(\theta) := m(1/\theta, 1) = h(T_S, T_B)$. Moreover, the total number of meetings in one period should be consistent regardless of the side counted from, so the following relationships hold

$$m(T_S, T_B) = T_S q(\theta) = T_B h(\theta), \quad q(\theta) = \theta h(\theta). \tag{4.1}$$

In the economic literature on search equilibrium analysis, the search intensities are strategic decisions made by the self-interested agents and are thus endogenous. However, here we focus on the matching and dynamic interactions between heterogeneous agents while simplifying the search aspect. In general we assume $q(\theta) : \mathbb{R}^+ \mapsto [0, 1]$, is nondecreasing and $q(0) = 0$, $q(\infty) = 1$.

Besides the assumption that agents on either side are randomly selected to participate in meetings, we also assume that the one-to-one pairings between these sellers and buyers to be random. These two assumptions determine that the type of agent that one meets, if any, is random and has a distribution consistent with that of the stocks on the other side of the market when the period starts.

In summary, in one period each seller (buyer) has a probability of $q(\theta)$ ($h(\theta)$) to meet with one buyer (seller) whose type has a probability distribution as that of the initial population of active buyers (sellers).

**Bargaining:** Then, when a meeting is formed, buyer and seller decides whether to match and at what transaction price by Nash bargaining based on their dynamic valuations, i.e., their nominal valuation and their continuation value that takes into account future trade opportunities and costs.

Specifically, upon meetings the pair of seller and buyer first sees whether there exists a surplus of transaction between them, then bargains on how to split that surplus to form price, and last compares the surplus of trading now with the expected future payoff of holding off.

Most importantly, the bargaining of seller and buyer should be based on their dynamic valuations. These not only incorporate their heterogeneity in the nominal valuation of the good/service, but also their dynamic opportunity cost or continuation value of market participation in the future. We denote the dynamic valuations, of type $V_S$ sellers and type $V_B$ buyers, that adjust for future opportunities as

$$\tilde{V}_S(V_S) = V_S + e^{-\beta\delta} W_S(V_S), \quad \tilde{V}_B(V_B) = V_B - e^{-\beta\delta} W_B(V_B), \tag{4.2}$$

where $W_S(V_S), W_B(V_B)$ are their dynamic continuation values. The dynamic continuation values are type dependent since heterogeneous agents expect to see different gains from poten-

tial future matches; and are stationary when we consider steady state in our infinite horizon model. The dynamic valuations $\tilde{V}_S(V_S), \tilde{V}_B(V_B)$ of each type can be seen as representing the 'true' valuation/type of the agents, which for a seller sums up her nominal cost plus the discounted continuation value or opportunity cost, while for a buyer it is the nominal valuation of the unit to buy minus the value of her lost future capacities. They reflect the actual willingness to sell/pay of the sellers and buyers during their dynamic interactions. Therefore the paired up sellers and buyers should calculate surplus of transactions and make matching decisions based on these future adjusted dynamic valuations instead of their nominal types.

Suppose within a meeting the seller has a dynamic valuation $\tilde{V}_S(V_S)$ less than that of the paired buyer $\tilde{V}_B(V_B)$. The difference $\tilde{V}_S(V_S) - \tilde{V}_B(V_B)$ is the potential surplus. We assume non-symmetric Nash bargaining between the seller and buyer on how to split the surplus, which is a classical axiomatic solution approach for the bargaining problem, characterized by desirable properties including Pareto optimality and independence of equivalent utility representation. These requirements on the bargaining solution is captured in the maximization of the Nash product $(p - \tilde{V}_S(V_S))^{\gamma}(\tilde{V}_B(V_B) - p)^{(1-\gamma)}$ (see Jr (1950), Roth (1979)) where $\gamma \in (0, 1)$ depicts asymmetries in bargaining procedure, bargaining ability, desire to maximize utility, etc (see Binmore et al. (1986) for detail). Note that the asymmetry in payoffs of alternatives outside transaction and thus in bargaining power has already been modeled and incorporated in the dynamic valuations $\tilde{V}_S(V_S), \tilde{V}_B(V_B)$. The resulting bargaining solution forms price

$$p = \tilde{V}_S(V_S) + \gamma \left( \tilde{V}_B(V_B) - \tilde{V}_S(V_S) \right), \tag{4.3}$$

at which the seller obtains $\gamma$ of the surplus while the buyer gathers $1 - \gamma$ of it. In addition, the asymmetry parameter $\gamma \in (0, 1)$ is assumed to be common in every meeting and reflects a marketwise imbalance in bargaining sophistication between sellers and buyers.

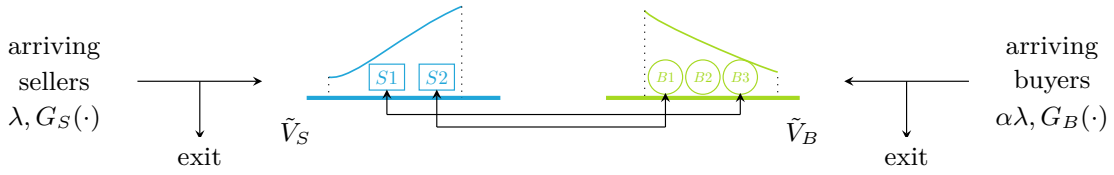



**Figure 4.1**

Then, after transacted buyers and sellers exit the market, all remaining agents suffer a delay penalty because of their failure to transact in the current period, and then carry over to the next period. Both buyers and sellers are trying to maximize discounted surplus. Figure 4.1 provides a schematic to visualize some aspects of the sequential meeting, Nash bargaining dynamics in the microstructure model.

### 4.2.3. Remarks

Two aspects need further remarks concerning the model before we move on to its analysis, first on the strategy game between buyers and sellers, and second on the mean-filed (fluid model) approximation of the market dynamics under the buyer/seller strategies.

**Dynamic cooperative game.** In dynamic matching markets, under the sequential meeting, Nash bargaining model as laid out, agents' strategies consist of (1) whether to enter or to exit the market (2) which ones on the opposite side to accept (match) upon meeting. We assume agents have perfect information in this dynamic cooperative game and do not strategize on type reporting.

We argue that the second component of the strategies, i.e., matching decisions, can be suppressed in the dynamic setting, as a result of the assumption of Nash bargaining based on dynamic valuations. For each meeting with positive surplus $\tilde{V}_S(V_S) - \tilde{V}_B(V_B) > 0$, according to Nash bargaining, the price formed as in (4.3) will be higher than the dynamic cost of the seller and lower than the dynamic value of the buyer. Therefore, the set of equilibria in which

agents bilaterally decide to match in a meeting as long as there is a positive surplus strictly dominate others. We focus on such equilibria and hence suppress the matching decisions as they can be derived from the dynamic valuations. Specifically, there are two situations of lost profitable matches that are ruled out: first, the seller and buyer lack coordination and believe the other would not decide to match; second, the seller or buyer mix between matching and declining.

In the remainder of the chapter, we define and characterize the decentralized steady state equilibrium, which will be a profile of individually rational, time-invariant entry/exit strategies. The assumption of steady state is not unreasonable when characterizing decentralized equilibrium, where the objective is to find fixed point of the best response correspondence, and the best response to time-invariant strategies of others should contain a time-invariant strategy as well. However, when solving the social planner's problem in centralized settings, examples such as Shimer and Smith (2001b) have shown that the optimal dynamic matching policy may feature nontrivial limit cycles when the social planner manipulates meeting chances and matching decisions.

**Fluid (mean-field) model.** We consider the dynamics of such marketplaces where sellers and buyers stochastically and dynamically arrive, meet, and match as a double-sided multiclass queueing network with economics and service rules as modeled above. We approximate the complicated queueing operations by its deterministic fluid limit to obtain first-order insights on the operational performance of the dynamic matching markets and its interaction with the economics. The use of fluid model can be motivated by the consumer-to-consumer market applications of dynamic matching, which feature large volumes of arrivals of infinitesimal agents on both the selling side and the buying side. In the remainder of the chapter we will define and characterize the market equilibrium that is not only economically rational in the individual decisions, but also operationally balanced in the population flows.

## 4.3. The Mean-Field Steady State Equilibrium

In this section we define the mean-field steady state equilibrium of the sequential meeting, Nash bargaining model of dynamic matching markets. As aforementioned, we focus on steady state equilibria that are free from the matching coordination problem. We characterize an equilibrium by 3 sets of conditions: (1) First, while the agents dynamically decide to enter/exit and trade/hold, their decision making and the dynamic continuation values should be characterized by recursive Bellman equations corresponding to each type. (2) Moreover, the equilibrium strategy of each type of agent should be utility maximizing and thus individually rational. (3) Finally, given agents' entry/exit and trading decisions in the equilibrium, together with the dynamics of the exogenous arrivals of new agents, the inflow and outflow of each type should be balanced so that the population of active agents of each type is in steady state. In the following sections we will talk about each of these equilibrium conditions.

### 4.3.1. Dynamic Value Functions

Recall that in Section 4.2.2 we denote $W_S(\cdot)$ and $W_B(\cdot)$ as the dynamic continuation values, i.e., the expected values of future payoffs of certain types of sellers and buyers, respectively. These dynamic continuation values are lost when any pair of agents decides to trade and exit the market. Therefore, the true cost of selling and the true value of buying of an agent should take into consideration the dynamic opportunity cost in addition to her nominal valuation, i.e., her type. We denote $\tilde{V}_S(\cdot), \tilde{V}_B(\cdot) : [L, U] \mapsto \mathbb{R}$ as the dynamic value functions which map the agents' types to their dynamic valuations that account for the opportunity costs of trading, defined by,

$$\tilde{V}_S(V_S) := V_S + e^{-\beta\delta} W_S(V_S), \quad \forall V_S \in [L, U], \tag{4.4}$$

$$\tilde{V}_B(V_B) := V_B - e^{-\beta\delta} W_B(V_B), \quad \forall V_B \in [L, U]. \tag{4.5}$$

The true cost of selling is the current cost plus the lost dynamic value, while the true value of buying is the current value minus the lost dynamic value.

We characterize the dynamic continuation values $W_S(\cdot), W_B(\cdot)$ recursively by steady state Bellman equations. We first introduce some notations. Denote by $T_S, T_B$ the steady state volume of active sellers and buyers in the market at the beginning of each period, respectively, which include both the newly arrived sellers/buyers who decide to join and those who have already participated for at least one period and decided to stay. In §4.2.2 we have assumed that both the selection and the pairing of the agents that meet in each period are random. Under these two assumptions the type of agent that one meets in a period, if any, is random and has a distribution consistent with that of the stock on the other side. We denote the steady state distributions of the dynamic valuations of active sellers and buyers at the beginning of each period as $F_S(\cdot), F_B(\cdot)$, respectively. In addition, define $\overline{F}_B(\cdot)$ as the tail distribution of the dynamic valuations of active buyers, $\overline{F}_B(\tilde{V}_B) := 1 - F_B(\tilde{V}_B)$, $\forall \tilde{V}_B \in \{x \in \mathbb{R} : x = \tilde{V}_B(V_B), V_B \in [L, U]\}$.

For the sellers, recall from §4.2.2 that within each period any seller has probability $q(\theta)$ to meet with one buyer, where $\theta = T_B/T_S$ is the steady state ratio of buyers to sellers in the market. The type of buyer that a seller meets, if any, has distribution $F_B(\cdot)$ as aforementioned. Upon meeting, say between a type $V_S$ seller and a type $V_B$ buyer, trade takes place and the seller collects her Nash bargaining revenue of

$$\begin{aligned} p - V_S &= \tilde{V}_S(V_S) + \gamma(\tilde{V}_B(V_B) - \tilde{V}_S(V_S)) - V_S \\ &= e^{-\beta\delta} W_S(V_S) + \gamma(\tilde{V}_B(V_B) - \tilde{V}_S(V_S)), \end{aligned} \tag{4.6}$$

if and only if $\tilde{V}_B(V_B) > \tilde{V}_S(V_S)$, which occurs with probability $\overline{F}_B(\tilde{V}_S(V_S))$ given that the

type $V_S$ seller participates in a meeting in the period. Otherwise, either the seller meets no one or fails to trade in meeting. Her steady state strategy would be to stay in the market following the current period, since she would face the same market status as when she decided to join at the first place. In this case she would again expect a future payoff of $W_S(V_S)$. At the same time, whether a seller meets, trades, or not in one period, she pays a fee of $c_S$ for market participation, as discussed at the beginning of Section 4.2. Therefore, in steady state, for any $V_S \in [L, U]$,

$$\begin{aligned} W_S(V_S) =& q(\theta) \int_{\tilde{V}_S(V_S)}^{\infty} \left( e^{-\beta\delta} W_S(V_S) + \gamma \left( s - \tilde{V}_S(V_S) \right) \right) dF_B(s) \\ &+ \left( 1 - q(\theta) \overline{F}_B(\tilde{V}_S(V_S)) \right) e^{-\beta\delta} W_S(V_S) - c_S \\ =& q(\theta)\gamma \int_{\tilde{V}_S(V_S)}^{\infty} \left( s - \tilde{V}_S(V_S) \right) dF_B(s) + e^{-\beta\delta} W_S(V_S) - c_S. \end{aligned} \tag{4.7}$$

Similarly, the dynamic continuation values of buyers in the steady state can also be characterized by Bellman equations. For any $V_B \in [L, U]$,

$$W_B(V_B) = h(\theta)(1-\gamma) \int_{-\infty}^{\tilde{V}_B(V_B)} \left( \tilde{V}_B(V_B) - t \right) dF_S(t) + e^{-\beta\delta} W_B(V_B) - c_B. \tag{4.8}$$

The above Bellman equations characterize the dynamic continuation values that sellers and buyers face at the beginning of each period in steady state. Sellers and buyers then decide whether to participate (enter or stay) by comparing them to the normalized zero benefit of the outside options. As a result, the seller types $\mathcal{A}_S := \{V_S \in [L, U] : W_S(V_S) \geq 0\}$ and the buyer types $\mathcal{A}_B := \{V_B \in [L, U] : W_B(V_B) \geq 0\}$ are active in the marketplace. [1] The

[1] We assume for simplicity of exposition that when agents are indifferent between market participation and the outside option they would decide to participate, which do not alter much of the nature of our analysis as we consider continuum of traders.

dynamic valuations of the active agents should then satisfy

$$\tilde{V}_S(V_S) = V_S + e^{-\beta\delta} W_S(V_S) \geq V_S \geq L, \quad \forall V_S \in \mathcal{A}_S, \tag{4.9}$$

$$\tilde{V}_B(V_B) = V_B - e^{-\beta\delta} W_B(V_B) \leq V_B \leq U, \quad \forall V_B \in \mathcal{A}_B. \tag{4.10}$$

Therefore, the steady state distributions $F_S(\cdot), F_B(\cdot)$ are constrained by the boundary conditions $F_S(L) = \overline{F}_B(U) = 0$. In this case, the infinite upper limit and lower limit of the integrals in (4.7) and (4.8) can be replaced by $U$ and $L$, respectively.

In view of the relationship between the dynamic continuation values and the dynamic valuations in equations (4.4) and (4.5), we can rewrite the Bellman equations in (4.7) and (4.8) with respect to $\tilde{V}_S, \tilde{V}_B$ alone, and obtain the following implicit characterization equations of the dynamic value functions,

$$\tilde{V}_S(V_S) = V_S + \frac{e^{-\beta\delta}}{1 - e^{-\beta\delta}} \left( q(\theta)\gamma \int_{\tilde{V}_S(V_S)}^{U} \left( s - \tilde{V}_S(V_S) \right) dF_B(s) - c_S \right), \quad \forall V_S \in [L, U], \tag{4.11}$$

$$\tilde{V}_B(V_B) = V_B - \frac{e^{-\beta\delta}}{1 - e^{-\beta\delta}} \left( h(\theta)(1-\gamma) \int_{L}^{\tilde{V}_B(V_B)} \left( \tilde{V}_B(V_B) - t \right) dF_S(t) - c_B \right), \forall V_B \in [L, U]. \tag{4.12}$$

Results in the following proposition are derived from the analysis of these characterization equations. In it we establish some properties of the dynamic value functions that will later simplify analysis. As will be discussed, the result also helps clarify the effect of attractiveness (meaning having low cost as a seller or having high valuation as a buyer) when dynamics play a role in matching.

**Proposition 1** (Monotonicity and Convexity/Concavity of Dynamic Value Functions). *The dynamic value function $\tilde{V}_S(\cdot)$ ($\tilde{V}_B(\cdot)$) is monotonically increasing (decreasing) and convex (con-*

*cave) in valuation types $V_S$ ($V_B$). In particular, for any $V_S, V_B \in [L, U]$,*

*(i)* $\tilde{V}_S'(V_S) = \left(1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma \overline{F}_B(\tilde{V}_S(V_S))\right)^{-1} \in (0,1)$,

$\tilde{V}_B'(V_B) = \left(1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} h(\theta)(1-\gamma) F_S(\tilde{V}_B(V_B))\right)^{-1} \in (0,1)$;

*(ii)* $\tilde{V}_S''(V_S) \geq 0$, $\tilde{V_B}''(V_B) \leq 0$.

That is, the dynamic value functions preserve the order of nominal valuations, and the dynamic valuations change more quickly with the nominal valuations when sellers' costs are high or when buyers' values are low. The proof of Proposition 1 can be found in the appendix. These results simplify the characterization of agents' strategies in the steady state equilibrium, which will be discussed in the next section.

The monotonicity and convexity/concavity properties of the dynamic value functions provide insights into the outcome of dynamic interactions of heterogeneous sellers and buyers. (1) Attractiveness sustains under dynamic interactions. This conclusion is not so direct because though the low cost sellers and high value buyers tend to contribute more to the surplus of any pairing, they also have higher opportunity costs to compensate in bargaining and leave less room of gain to the other party. However, (2) the marginal effect of attractiveness on the expected future payoff is diminishing. Improvement in nominal valuation has more impact on expected future payoff when an agent is less attractive. (3) Heterogeneity is modulated when dynamics play a role. In particular, in comparison with their static nominal counterparts, the dynamic valuations have the following features: a) the dynamic valuations are more modest, since for any $v \in \mathcal{A}_S$ ($v \in \mathcal{A}_B$), $\tilde{V}_S(v) > v$ ($\tilde{V}_B(v) < v$); b) the dynamic valuations differ less, since $d\tilde{V}_S(V_S)/dV_S, d\tilde{V}_B(V_B)/dV_B \in (0,1)$; (c) and last, attractiveness has diminishing marginal effect under dynamic interactions, c.f., (2) above. Dynamics shade the differences between heterogeneous agents as they shift agents' valuations by opportunity costs. Such cost increases with attractiveness, and thus reduces the difference in it.

### 4.3.2. Decentralized Decision Making

In this section we study how sellers and buyers make self-interested decisions in the steady state equilibrium. We base the analysis on the characterization of their dynamic continuation values $W_S(\cdot), W_B(\cdot)$, and correspondingly their dynamic value functions $\tilde{V}_S(\cdot), \tilde{V}_B(\cdot)$, from the previous section.

As discussed near the end of §4.2, in the steady state equilibrium free from coordination problems, strategies of sellers and buyers are suppressed and consist of only time-invariant entry/exit decisions. At the beginning of a period, sellers and buyers expect $W_S(\cdot)$ and $W_B(\cdot)$ payoffs from market participation, respectively, which take into account all future opportunities including those in the current period. These dynamic continuation values are determined recursively by Bellman equations in the previous section. By individual rationality, sellers and buyers should become and remain active in the market if and only if $W_S(\cdot) \geq 0$ (Recall that we have assumed indifferent agents join). That is, the decentralized strategies of agents in the steady state equilibrium are characterized by the active sets of types $\mathcal{A}_S, \mathcal{A}_B$.

The monotonicity results established in Proposition 1 can help us simplify the characterization of sets $\mathcal{A}_S, \mathcal{A}_B$. We first need to translate these analytical properties of the dynamic value functions into those of the dynamic continuation values. By their relationship in (4.4) and (4.5), for any $V_S, V_B \in [L, U]$,

$$W_S'(V_S) = \frac{de^{\beta\delta}(\tilde{V}_S(V_S) - V_S)}{dV_S} = e^{\beta\delta}\left(\frac{d\tilde{V}_S}{dV_S} - 1\right) < 0; \text{ analogously, } W_B'(V_B) > 0. \quad (4.13)$$

That is, monotonicity transfers from the dynamic value functions to the dynamic continuation values. Attractive agents (low cost sellers and high value buyers) expect higher future payoffs, which is intuitive since they would expected to transact sooner and also would reap more

benefit from any transaction.

$W_S(\cdot), W_B(\cdot)$ change signs at most once in $[L, U]$. We first claim that $W_S(U), W_B(L) < 0$. If not, take the selling side as example,

$$\tilde{V}_S(U) = U + e^{-\beta\delta} W_S(U) \geq U \geq V_B - e^{-\beta\delta} W_B(V_B) = \tilde{V}_B(V_B), \quad \forall V_B \in \mathcal{A}_B.$$

There is no buyer type that can produce positive surplus with type $U$ sellers, while they still need to pay $c_S > 0$ in each period if they participate. Hence their expected future payoff of market participation should be negative. We then rule out the null equilibrium where $W_S(L) < 0$ or $W_B(U) < 0$, in which case at least one side of the market collectively decide not to join and no trading activity occurs in the marketplace. The rest of the paper focuses on the characterization of nontrivial steady state equilibrium. Given the above two sets of boundary conditions, since $W_S(\cdot)$ is strictly decreasing and $W_B(\cdot)$ is strictly increasing, there must exist a pair of unique 'indifferent' seller type and buyer type, which we denote as $\overline{V}_S, \underline{V}_B$, such that $W_S(\overline{V}_S) = W_B(\underline{V}_B) = 0$, $\overline{V}_S \in [L, U), \underline{V}_B \in (L, U]$. Therefore, because of monotonicity,

$$\mathcal{A}_S = [L, \overline{V}_S], \quad \mathcal{A}_B = [\underline{V}_B, U], \tag{4.14}$$

are intervals contained in $[L, U]$ with thresholds $\overline{V}_S, \underline{V}_B$, which will then be determined in the characterization of the steady state equilibrium.

### 4.3.3. Flow Balances

The third dimension of the steady state equilibrium characterization is flow balancing. As shown in the previous two sections, the composition of the stocks of active sellers and buyers affect the payoffs and decision makings of the agents, and subsequently the market dynamics. In turn, agents' strategies and the resulting entry/exit and trade actions in the equilibrium

should reversely stabilize the composition of active sellers and buyers on the two sides of the market so that it can remain in steady state and the market can equilibrate.

The inflow into the market in each period consist of those who decide to join among the new arrivals of sellers and buyers. Recall from §4.2 that these newly arrived agents are exogenously given valuation types that are i.i.d. draws from distributions $G_S, G_B : [L, U] \mapsto [0, 1]$, and arrive at rate $\lambda$, and $\alpha\lambda$, respectively, on the selling side and the buying side. According to result in §4.3.2, the new arrivals who decide to join are sellers with types below $\overline{V}_S$ and buyers with types above $\underline{V}_B$.

The outflow of the market in each period results from transactions that take place after meetings of sellers and buyers who have surpluses between their dynamic valuations. We have assumed in Section 4.2 that agents do not leave the market for other reasons, but apparently such an assumption can be easily released by formulating additional exogenous outflows.

While arrivals (inflows) are indexed by agents' nominal types, transactions (outflows) are determined by dynamic valuations. To bridge the gap, we need to translate the arrivals of different nominal types, $V_S, V_B$, into those of dynamic valuation types, $\tilde{V}_S, \tilde{V}_B$, based on their characterization in Section 4.3.1. Denote the arrival distributions of sellers and buyers in terms of their dynamic valuations as $\tilde{G}_S(\cdot), \tilde{G}_B(\cdot)$.

The remaining results in this paper are predicated on the following simplifying assumption:

**Assumption 10.** *$G_S(\cdot), G_B(\cdot)$ are uniformly distributed over $[L, U]$.*

In this case, $\tilde{G}_S(\cdot), \tilde{G}_B(\cdot)$ can be easily determined using results from Section 4.3.1. In particular, take the selling side as example, by strict monotonicity of the dynamic value functions $\tilde{V}_S(\cdot), \tilde{V}_B(\cdot)$, they are invertible, and

$$\tilde{G}_S(x) = \mathbb{P}\left(\tilde{V}_S(V_S) \leq x\right) = \mathbb{P}\left(V_S \leq \tilde{V}_S^{-1}(x)\right) = G_S(\tilde{V}_S^{-1}(x)).$$

In addition, given the derivative of the dynamic value function provided in Proposition 1, the derivative of its inverse $\tilde{V}_S^{-1}(\cdot)$ should be

$$\frac{d\tilde{V}_S^{-1}(x)}{dx} = \left(\frac{d\tilde{V}_S(\tilde{V}_S^{-1}(x))}{d\tilde{V}_S^{-1}(x)}\right)^{-1} = 1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma \overline{F}_B(x).$$

Furthermore, because of Assumption 10, $G_S'(x) = 1/(U-L)$. Hence, by chain rule,

$$\tilde{G}_S'(x) = \frac{1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma \overline{F}_B(x)}{U-L}; \text{ analogously, } \tilde{G}_B'(x) = \frac{1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} h(\theta)(1-\gamma) F_S(x)}{U-L}. \tag{4.15}$$

Recall that seller types lower than $\overline{V}_S$ and buyer types higher than $\underline{V}_B$ would decide to join. And the thresholds $\overline{V}_S, \underline{V}_B$ are types that are indifferent between market participation and the outside option, so for them $\tilde{V}_S(\overline{V}_S) = \overline{V}_S, \tilde{V}_B(\underline{V}_B) = \underline{V}_B$. As a result, the transformed arrival distributions should satisfy boundary conditions $\tilde{G}_S(\overline{V}_S) = 1, \tilde{G}_B(\underline{V}_B) = 0$. Together with the derivatives in (4.15), they determine the inflows into the market in each period in terms of agents' dynamic valuations.

To ensure that the composition of the inventory on both sides of the market remain stationary, flows into and out of each type need to be balanced in every period, so that not only the total volumes, $T_S, T_B$, but also the type distributions, $F_S(\cdot), F_B(\cdot)$, of active sellers and buyers are in steady state. That is, the steady state equilibrium requires flow balance for any subset of seller or buyer types. Take the selling side as example, one equivalent condition is that for any $\tilde{V}_S \in [\tilde{V}_S(L), \overline{V}_S]$, i.e., any dynamic valuation that corresponds to an active seller type, flows are balanced for the set of sellers with types within $[\tilde{V}_S(L), \tilde{V}_S]$. In each period, the inflow into the set is arrival of sellers with types lower than $\tilde{V}_S$ and thus is at rate $\lambda \tilde{G}_S(\tilde{V}_S)$. On the other hand, for each seller in this set with type $t < \tilde{V}_S$, there is a probability of $q(\theta)\overline{F}_B(t)$ to transact in the current period. The outflow of the set is their potential

transactions following random meeting and matching with the other side, and is expected to take place at rate $T_S \int_{\tilde{V}_S(L)}^{\tilde{V}_S} q(\theta)\overline{F}_B(t)dF_S(t)$. Hence, including the buying side in parallel, flow balance conditions can be written as: for any $\tilde{V}_S \in [\tilde{V}_S(L), \overline{V}_S], \tilde{V}_B \in [\underline{V}_B, \tilde{V}_B(U)]$,

$$\lambda \tilde{G}_S(\tilde{V}_S) = T_S \int_{\tilde{V}_S(L)}^{\tilde{V}_S} q(\theta)\overline{F}_B(t)dF_S(t), \tag{4.16}$$

$$\lambda \tilde{G}_B(\tilde{V}_B) = T_B \int_{\underline{V}_B}^{\tilde{V}_B} h(\theta)F_S(s)dF_B(s). \tag{4.17}$$

Characterizing the equilibrium inventory distributions $F_S(\cdot), F_B(\cdot)$ is key to the study of the steady state equilibrium. To facilitate the analysis, we make the following technical assumption:

**Assumption 11.** *$F_S(\cdot), F_B(\cdot)$ has probability density functions, which we denote as $f_S(\cdot), f_B(\cdot)$.*

That is, in the steady state equilibrium, distributions over the dynamic valuations of active sellers and buyers are assumed to be atomless.

With Assumption 11, together with the previous discussion on the transformed arrival distributions, we can take derivative on both sides of equations (4.16) and (4.17), and obtain

$$\lambda \cdot \frac{1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma \overline{F}_B(\tilde{V}_S)}{U - L} = T_S q(\theta)\overline{F}_B(\tilde{V}_S) f_S(\tilde{V}_S), \tag{4.18}$$

for any $\tilde{V}_S \in [\tilde{V}_S(L), \overline{V}_S]$, and

$$\alpha\lambda \cdot \frac{1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} h(\theta)(1-\gamma) F_S(\tilde{V}_B)}{U - L} = T_B h(\theta) F_S(\tilde{V}_B) f_B(\tilde{V}_B), \tag{4.19}$$

for any $\tilde{V}_B \in [\underline{V}_B, \tilde{V}_B(U)]$. The above system of equations warrants flow balances and thus characterizes the equilibrium inventory distributions.

To study its solution, we segment the dynamic valuations of the active population. We

first claim that $\tilde{V}_S(L) \leq \underline{V}_B, \overline{V}_S \leq \tilde{V}_B(U)$. The reason is that the threshold types, if not being able to trade with the most competitive type from the other side of the market, would expect a strictly negative payoff from market participation, which contradicts with their definition. We segment the dynamic valuations of the active population into two categories. First we call sellers with types lower than $\underline{V}_B$ and buyers with types higher than $\overline{V}_S$ the *non-overlapping* population. These agents trade as soon as they meet. The distributions over the non-overlapping types are easy to derive. Because meeting is random, they should be proportional to the arrivals. In particular, for the non-overlapping sellers with $\tilde{V}_S \in [\tilde{V}_S(L), \underline{V}_B]$ and buyers with $\tilde{V}_B \in [\overline{V}_S, \tilde{V}_B(U)]$, $F_S(\tilde{V}_B) = \overline{F}_B(\tilde{V}_S) = 1$, so for the non-overlapping types,

$$f_S(\tilde{V}_S) = \frac{\lambda\left(1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma\right)}{(U-L)T_S q(\theta)}, \quad f_B(\tilde{V}_B) = \frac{\alpha\lambda\left(1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} h(\theta)(1-\gamma)\right)}{(U-L)T_B h(\theta)}. \tag{4.20}$$

Second, we call sellers and buyers of types above buyers' threshold $\underline{V}_B$ and below sellers' threshold $\overline{V}_S$, if any, the *overlapping* population. In addition, we call steady state equilibria in which the set of overlapping population is nonempty as *non full trade* equilibria. In a *full trade* equilibria, all types are non-overlapping and agents trade as soon as they meet with anyone from the other side of the market. The probability to trade in each period is then degenerate and should equal their chance of meeting, $q(\theta)$, for the selling side, or $h(\theta)$, for the buying side. Full trade equilibria can be simply characterized by combining (4.20) with results in the previous two sections.

The remainder of the paper concerns characterization of non full trade equilibria, i.e., equilibria in which $\underline{V}_B < \overline{V}_S$ and there exists overlapping population. For them, given that $f_S(\cdot) = F_S'(\cdot)$, $f_B(\cdot) = -\overline{F}_B'(\cdot)$, the system of characterizing equations of the equilibrium inventory distributions is actually a system of ordinary differential equations (ODEs), and

can be rewritten as

$$\begin{cases} f_S(t) = \dfrac{\lambda}{U-L}\left(\dfrac{1}{T_S q(\theta)\overline{F}_B(t)} + \dfrac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\dfrac{\gamma}{T_S}\right), \\ f_B(t) = \dfrac{\alpha\lambda}{U-L}\left(\dfrac{1}{T_B h(\theta) F_S(t)} + \dfrac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\dfrac{1-\gamma}{T_B}\right), \end{cases} \tag{4.21}$$

with $t \in [\underline{V}_B, \overline{V}_S]$ and boundary conditions $F_S(\overline{V}_S) = \overline{F}_B(\underline{V}_B) = 1$.

The following proposition provides a solution to this system of ODEs, whose proof can be found in the appendix.

**Proposition 2** (Characterization of Equilibrium Inventory Distributions). *Distributions $F_S(\cdot), \overline{F}_B(\cdot)$ satisfying*

$$C_2 - \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} \cdot \frac{\alpha\lambda(1-\gamma)}{(U-L)T_B} \cdot t = \int_{\overline{F}_B(\bar{V}_S)}^{\overline{F}_B(t)} \frac{W\left(C_1 u^{-\frac{1}{\alpha}} e^{-\frac{1}{\alpha}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma u}\right)}{1 + W\left(C_1 u^{-\frac{1}{\alpha}} e^{-\frac{1}{\alpha}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma u}\right)} du, \tag{4.22}$$

$$F_S(t) = \frac{1}{\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} h(\theta)(1-\gamma)} W\left(C_1 \overline{F}_B(t)^{-\frac{1}{\alpha}} e^{-\frac{1}{\alpha}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma \overline{F}_B(t)}\right), \tag{4.23}$$

*for $t \in [\underline{V}_B, \overline{V}_S]$ with constants $C_1, C_2$, where $W(\cdot)$ is the Lambert W function,* [2] *solves the system of ODEs in* (4.21).

Throughout §4.3.1-4.3.3, we have recursively determined the monotonic and convex/concave mapping from nominal types to dynamic valuations, have characterized that the equilibrium strategy profile features participating thresholds on both sides of the market, and have written flow balancing conditions as a system of ordinary differential equations in terms of market participants' dynamic valuations. Based on these results we propose the following definition

[2] Lambert W function is the inverse of the function $f(W) = We^W$.

of equilibrium for the dynamic matching market:

**Definition 4** (Mean-Field Steady State Equilibrium). *A steady state equilibrium*

$$(\overline{V}_S, \tilde{V}_S, F_S, \underline{V}_B, \tilde{V}_B, F_B, T_S, \theta) \in [L,U] \times \mathbb{C}[L,U] \times \mathbb{C}[L,U] \times [L,U] \times \mathbb{C}[L,U] \times \mathbb{C}[L,U]$$

*is a pair of sets of participating threshold, dynamic value function, and distribution over dynamic valuations that satisfies*

*(i)* ***Dynamic Valuation:*** *For any $V_S \in [L, \overline{V}_S], V_B \in [\underline{V}_B, U]$, the corresponding dynamic valuations $\tilde{V}_S(V_S), \tilde{V}_B(V_B)$ solve, respectively,*

$$\tilde{V}_S(V_S) = V_S + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} \left( q(\theta)\gamma \int_{\tilde{V}_S(V_S)}^{U} \left( s - \tilde{V}_S(V_S) \right) dF_B(s) - c_S \right), \tag{4.24}$$

$$\tilde{V}_B(V_B) = V_B - \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} \left( h(\theta)(1-\gamma) \int_{L}^{\tilde{V}_B(V_B)} \left( \tilde{V}_B(V_B) - t \right) dF_S(t) - c_B \right). \tag{4.25}$$

*(ii)* ***Marginal Indifference:*** *Type $\overline{V}_S$ sellers and type $\underline{V}_B$ buyers are indifferent in participating and have zero continuation values*

$$\tilde{V}_S(\overline{V}_S) = \overline{V}_S, \quad \tilde{V}_B(\underline{V}_B) = \underline{V}_B. \tag{4.26}$$

*(iii)* ***Flow Balance:*** *For any $t \in [\overline{V}_S, \underline{V}_B]$, the distribution functions $F_S(V_S), F_B(V_B)$ and their densities solve the system of ordinary differential equations*

$$\begin{cases} f_S(t) = \dfrac{\lambda}{U-L} \left( \dfrac{1}{T_S q(\theta) \overline{F}_B(t)} + \dfrac{e^{-\beta\delta}}{1-e^{-\beta\delta}} \dfrac{\gamma}{T_S} \right), \\ f_B(t) = \dfrac{\alpha\lambda}{U-L} \left( \dfrac{1}{T_B h(\theta) F_S(t)} + \dfrac{e^{-\beta\delta}}{1-e^{-\beta\delta}} \dfrac{1-\gamma}{T_B} \right), \end{cases} \tag{4.27}$$

*with boundary conditions* $F_S(\overline{V}_S) = 1 - F_B(\underline{V}_B) = 1$. *For the non-overlapping sellers with* $\tilde{V}_S \in [\tilde{V}_S(L), \underline{V}_B]$ *and buyers with* $\tilde{V}_B \in [\overline{V}_S, \tilde{V}_B(U)]$,

$$f_S(\tilde{V}_S) = \frac{\lambda\left(1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma\right)}{(U-L)T_S q(\theta)}, \quad f_B(\tilde{V}_B) = \frac{\alpha\lambda\left(1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} h(\theta)(1-\gamma)\right)}{(U-L)T_B h(\theta)}. \tag{4.28}$$

*(iv)* ***Buyer/Seller Ratio:*** $T_B = \theta T_S$.

To summarize, the mean-filed steady state equilibrium takes inputs of arrival imbalance $\alpha$, discount $\beta$, bargaining imbalance $\gamma$, period length $\delta$, arrival rate $\lambda$, participation costs $c_S, c_B$, and valuation range $L, U$ (market primitives), and yields outputs of market depth $T_S, T_B$, S/D imbalance $\theta$, distribution of waiting agents $F_S(\cdot), F_B(\cdot)$, participation thresholds $\overline{V}_S, \underline{V}_B$, and value functions $\tilde{V}_S(\cdot), \tilde{V}_B(\cdot)$ (depth, dispersion). The equilibrium integrates individual rationality, flow balance, boundary conditions, and endogenous dynamic value functions

## 4.4. Equilibrium Characterization by Linearization

Exact analysis of the equilibrium in Definition 4, in which both the dynamic valuations $\tilde{V}_S, \tilde{V}_B$ and the inventory distributions $F_S, F_B$ are determined implicitly, c.f., (4.11) - (4.12) and Proposition 2, is theoretically intractable and computationally expensive.

In this section, we show that when the range of the valuations for buyers and sellers is small, bidding strategies are linear (§4.4.1). Motivated by this result we study a market under an assumption that buyers and sellers always employ linear strategies. This assumption leads to a more tractable model, and is motivated by markets with "almost homogeneous" goods and agents, e.g., market of apartments in the same neighborhood. In particular, for uniformly distributed valuation distributions, we obtain characterization of the market equilibrium, in which delay increases with sellers' costs and decreases with buyers' valuations as power law,

or exponentially in some cases (§4.4.2). We solve for the steady state equilibrium under linear strategies and answer the questions of who would join/exit the market, what are the determinants of the depth of market, who would wait for how long in the market, and what is the distribution of valuations for buyers and sellers in the market (§4.4.3). Later on, we shall also show via numerical tests that such an approximation is close to the true market equilibrium when valuation range is moderate (§4.5.2).

### 4.4.1. Linear Approximations of Dynamic Value Functions

We propose to approximate the recursively determined equilibrium dynamic value functions $\tilde{V}_S(\cdot), \tilde{V}_B(\cdot)$ by their first order Taylor expansions (linear) at the thresholds $\overline{V}_S, \underline{V}_B$ on the selling side and the buying side, respectively. We provide in this subsection a theoretical justification: this linear approximation is asymptotically accurate when the extent of heterogeneity represented as the width of the valuation range $U - L$ diminishes.

We denote the linearly approximated dynamic value functions, of types $V_S \in [L, \overline{V}_S], V_B \in [\underline{V}_B, U]$ on the selling side and the buying side, respectively, as follows

$$\tilde{V}_S^{(L)}(V_S) := \overline{V}_S + a_S(V_S - \overline{V}_S), \tag{4.29}$$

$$\tilde{V}_B^{(L)}(V_B) := \underline{V}_B + a_B(V_B - \underline{V}_B), \tag{4.30}$$

where $a_S$ is the left derivative of the function $\tilde{V}_S(\cdot)$ at $\overline{V}_S$, and similarly, $a_B$ is the right derivative of the function $\tilde{V}_B(\cdot)$ at $\underline{V}_B$. From Proposition 1, we have that

$$a_S = \left(1 + \frac{e^{-\beta\delta}}{1 - e^{-\beta\delta}} q(\theta)\gamma \overline{F}_B(\overline{V}_S)\right)^{-1}, \tag{4.31}$$

$$a_B = \left(1 + \frac{e^{-\beta\delta}}{1 - e^{-\beta\delta}} h(\theta)(1-\gamma) F_S(\underline{V}_B)\right)^{-1}. \tag{4.32}$$

Also, recall that the exact dynamic valuations $\tilde{V}_S, \tilde{V}_B$ can be characterized by equations (4.11) - (4.12).

The following theorem shows that when the valuation range diminishes, i.e., $U - L \to 0$, the difference between these two $\tilde{V}$ characterizations vanishes at a faster speed, and therefore the former can be used to approximate the latter when agents' valuations are similar.

**Theorem 6.** *(Linearization Justification) Under Assumption 11, if, in addition, the dynamic value functions $\tilde{V}_S, \tilde{V}_B$ are continuous, then*

$$\sup_{t\in[0,1]} \left| \frac{\tilde{V}_S(\overline{V}_S - t(\overline{V}_S - L)) - \tilde{V}_S^{(L)}(\overline{V}_S - t(\overline{V}_S - L))}{\overline{V}_S - L} \right| \to 0 \text{ as } U - L \to 0, \tag{4.33}$$

$$\sup_{t\in[0,1]} \left| \frac{\tilde{V}_B(\underline{V}_B + t(U - \underline{V}_B)) - \tilde{V}_B^{(L)}(\underline{V}_B + t(U - \underline{V}_B))}{U - \underline{V}_B} \right| \to 0 \text{ as } U - L \to 0. \tag{4.34}$$

**Proof**. Take the selling side as example, for an active seller with nominal valuation $V_S \in [L, \overline{V}_S]$, its type can be rewritten as

$$V_S = \overline{V}_S - t(\overline{V}_S - L) \in [L, \overline{V}_S] \subseteq [L, U], \quad \forall t \in [0,1].$$

According to Proposition 1, we have

$$\tilde{V}_S'(V_S) = \left(1 + \frac{e^{-\beta\delta}}{1 - e^{-\beta\delta}} q(\theta)\gamma \overline{F}_B(\tilde{V}_S(V_S))\right)^{-1} \in (0,1), \quad \tilde{V}_S''(V_S) > 0.$$

That is, the dynamic value function $\tilde{V}_S(\cdot)$ is increasing and convex in $V_S$. As a result, we

have the following inequality

$$\overline{V}_S - \tilde{V}_S(V_S) = \int_{V_S}^{\overline{V}_S} \tilde{V}_S'(x)dx < \int_{V_S}^{\overline{V}_S} \left(1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma \overline{F}_B(\overline{V}_S)\right)^{-1} dx = a_S(\overline{V}_S - V_S). \tag{4.35}$$

By rearranging the terms on the two sides of inequality (4.35), we can show that $\tilde{V}_S$ is greater than its linear approximation $\tilde{V}_S^{(L)}$ for active types. In particular, for any $V_S \in [L, \overline{V}_S]$, we have that

$$\tilde{V}_S(V_S) - \tilde{V}_S^{(L)}(V_S) = \tilde{V}_S(V_S) - (\overline{V}_S + a_S(V_S - \overline{V}_S)) > 0. \tag{4.36}$$

Then, we explicitly write out the left hand side of (4.36), i.e., the difference between the dynamic value function $\tilde{V}_S$ and its linear approximation $\tilde{V}_S^{(L)}$, arriving at

$$\begin{aligned} 0 &< \tilde{V}_S(V_S) - \tilde{V}_S^{(L)}(V_S) \\ &= V_S + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} \left( q(\theta)\gamma \int_{\tilde{V}_S(V_S)}^{U} \left(s - \tilde{V}_S(V_S)\right) dF_B(s) - c_S \right) - \overline{V}_S - a_S(V_S - \overline{V}_S) \\ &= (1 - a_S)(V_S - \overline{V}_S) + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} \left( q(\theta)\gamma \int_{\tilde{V}_S(V_S)}^{U} \left(s - \tilde{V}_S(V_S)\right) dF_B(s) - c_S \right), \end{aligned} \tag{4.37}$$

for any $V_S \in [L, \overline{V}_S]$. Because the marginal sellers of the threshold type $\overline{V}_S$ should be indifferent between market participation and the outside option of zero payoff and have zero continuation values, c.f., (4.26), we have that

$$c_S = q(\theta)\gamma \int_{\overline{V}_S}^{U} \left(s - \overline{V}_S\right) dF_B(s).$$

As a result, we can substitute the $c_S$ and obtain

$$
\begin{aligned}
&0 < \tilde{V}_S(V_S) - \tilde{V}_S^{(L)}(V_S) \\
=&(1-a_S)(V_S - \overline{V}_S) \\
&+ \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} \left( q(\theta)\gamma \int_{\tilde{V}_S(V_S)}^{U} \left(s - \tilde{V}_S(V_S)\right) dF_B(s) - q(\theta)\gamma \int_{\overline{V}_S}^{U} \left(s - \overline{V}_S\right) dF_B(s) \right) \\
=&(1-a_S)(V_S - \overline{V}_S) + \\
&\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma \left( \int_{\tilde{V}_S(V_S)}^{\overline{V}_S} \left(s - \tilde{V}_S(V_S)\right) dF_B(s) + \overline{F}_B(\overline{V}_S)(\overline{V}_S - \tilde{V}_S(\overline{V}_S)) \right).
\end{aligned}
$$

Note that we have $\tilde{V}_S(V_S) \leq \overline{V}_S$ from $V_S \leq \overline{V}_S$ and the monotonicity of $\tilde{V}_S(\cdot)$. We then progress to provide an upper bound of the right hand side,

$$
\begin{aligned}
&0 < \tilde{V}_S(V_S) - \tilde{V}_S^{(L)}(V_S) \\
&\leq (1-a_S)(V_S - \overline{V}_S) \\
&\quad + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma \left( (\overline{F}_B(\tilde{V}_S(V_S)) - \overline{F}_B(\overline{V}_S))(\overline{V}_S - \tilde{V}_S(V_S)) + \overline{F}_B(\overline{V}_S)(\overline{V}_S - \tilde{V}_S(\overline{V}_S)) \right) \\
&= (1-a_S)(V_S - \overline{V}_S) + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma \overline{F}_B(\tilde{V}_S(V_S))(\overline{V}_S - V_S) \\
&< (1-a_S)(V_S - \overline{V}_S) + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma \overline{F}_B(\tilde{V}_S(V_S)) a_S(\overline{V}_S - V_S).
\end{aligned}
$$

The last inequality is from (4.35). Now we have established the following bounds on $\tilde{V}_S(V_S) -$

$\tilde{V}_S^{(L)}(V_S)$:

$$\begin{aligned}
0 &< \tilde{V}_S(V_S) - \tilde{V}_S^{(L)}(V_S) \\
&< (\overline{V}_S - V_S)\left(a_S\left(1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma \overline{F}_B(\tilde{V}_S(V_S))\right) - 1\right) \\
&= (\overline{V}_S - V_S)\frac{\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma(\overline{F}_B(\tilde{V}_S(V_S)) - \overline{F}_B(\overline{V}_S))}{1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma \overline{F}_B(\overline{V}_S)} \\
&\leq (\overline{V}_S - L)\frac{\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma(\overline{F}_B(\tilde{V}_S(V_S)) - \overline{F}_B(\overline{V}_S))}{1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma \overline{F}_B(\overline{V}_S)}
\end{aligned}$$

Because of continuity of $F_B(\cdot), \tilde{V}_S(\cdot)$ and the boundedness of $q(\theta), \gamma, \overline{F}_B(\overline{V}_S)$,

$$\lim_{V_S \to \overline{V}_S} \overline{F}_B(\tilde{V}_S(V_S)) - \overline{F}_B(\overline{V}_S) = 0. \tag{4.38}$$

Therefore,

$$\lim_{U-L\to 0} \sup_{V_S = U - t(U-L), t\in[0,1]} \left| \frac{\tilde{V}_S(V_S) - \tilde{V}_S^{(L)}(V_S)}{\overline{V}_S - L} \right| \to 0$$

Analysis of the buying side is analogous. ■

### 4.4.2. Inventory Distributions: Derivation of the Functional Form

In this subsection, we derive the functional form of the equilibrium distribution of buyer and seller types in the waiting populations $F_S(\cdot), F_B(\cdot)$, assuming uniform arrival distributions and linear dynamic value functions.

Recall from §4.3.3 that we denote the arrival distributions of sellers and buyers in terms of their dynamic valuations as $\tilde{G}_S(\cdot), \tilde{G}_B(\cdot)$. Now we assume that the dynamic valuations of different types of agents are linear perturbations around the the cutoffs $\overline{V}_S$ and $\underline{V}_B$, as in (4.29) - (4.30). Then, denoting $g(\cdot) := G'(\cdot), g_S(\cdot) := \tilde{G}'_S(\cdot)$, $g_B(\cdot) := \tilde{G}'_B(\cdot)$, we have that for

any $\tilde{V}_S \in [\tilde{V}_S(L), \overline{V}_S], \tilde{V}_B \in [\underline{V}_B, \tilde{V}_B(U)]$,

$$g_S(\tilde{V}_S) = g\left(\overline{V}_S + \frac{\tilde{V}_S - \overline{V}}{a_S}\right) \cdot \frac{1}{a_S} = \frac{1}{a_S(U-L)} \tag{4.39}$$

$$g_B(\tilde{V}_B) = g\left(\underline{V}_B + \frac{\tilde{V}_B - \underline{V}_B}{a_B}\right) \cdot \frac{1}{a_B} = \frac{1}{a_B(U-L)}. \tag{4.40}$$

In this case, taking derivatives on both sides of the flow balance equations in (4.16) - (4.17), the system of ODEs that characterize the equilibrium distribution of the waiting populations now simplifies to:

$$\begin{cases} f_S(t) = \dfrac{\lambda}{a_S(U-L)T_S q(\theta)\overline{F}_B(t)}, \\ f_B(t) = \dfrac{\alpha\lambda}{a_B(U-L)T_B h(\theta)F_S(t)}. \end{cases} \tag{4.41}$$

As before, the distribution over non-overlapping population can be easily derived. In particular, for $\tilde{V}_S \in [\tilde{V}_S(L), \underline{V}_B]$, $\tilde{V}_B \in [\overline{V}_S, \tilde{V}_B(U)]$, we can derive from the flow balance conditions that

$$f_S(\tilde{V}_S) = \frac{\lambda}{(U-L)a_S T_S q(\theta)}, \quad f_B(\tilde{V}_B) = \frac{\alpha\lambda}{(U-L)a_B T_B h(\theta)}. \tag{4.42}$$

The distribution over the overlapping population, however, is more complicated and the result is provided in the following proposition.

**Proposition 3.** *(Functional Forms of $F_S(\cdot), F_B(\cdot)$) When $a_B \neq \alpha a_S$,*

$$\begin{cases} F_S(t) = C_1\left(\dfrac{a_B - \alpha a_S}{C_1 a_B}t + C_2\right)^{\frac{a_B}{a_B - \alpha a_S}}, \\ \overline{F}_B(t) = \dfrac{\lambda}{a_S(U-L)T_S q(\theta)}\left(\dfrac{a_B - \alpha a_S}{C_1 a_B}t + C_2\right)^{-\frac{\alpha a_S}{a_B - \alpha a_S}}, \end{cases} \tag{4.43}$$

*solves the system of ODEs in* (4.41)*; or when* $a_B = \alpha a_S$

$$\begin{cases} F_S(t) = C_2 e^{\frac{\lambda t}{C_1 a_S (U-L) T_S q(\theta)}}, \\ \overline{F}_B(t) = \dfrac{C_1}{C_2} e^{-\frac{\lambda t}{C_1 a_S (U-L) T_S q(\theta)}}, \end{cases} \tag{4.44}$$

*solves the system of ODEs in* (4.41)*.*

The proof of Proposition 3 can be found in the Appendix.

### 4.4.3. Equilibrium Characterization

In this section, we first update the definition of the mean-field steady state equilibrium under the linear strategies. We then discuss about how to characterize a non full trade equilibrium. In particular, for uniformly distributed valuation distributions, we obtain near closed-form characterization that can be leveraged to answer the questions of who would join/exit the market, what are the determinants of the depth of market, who would wait for how long in the market, and what is the distribution of valuations for buyers and sellers in the market.

**Definition 5** (Linearized Mean-Field Steady State Equilibrium)**.** *A linearized mean-field steady state equilibrium*

$$(T_S, T_B, \theta, F_S(\cdot), \overline{F}_B(\cdot), \overline{V}_S, \underline{V}_B, a_S, a_B),$$

*given a set of market primitives* $(U, L, c_S, c_B, \lambda, \alpha, \gamma, \beta, \delta)$*, is a vector satisfying the following conditions:*

*(1)* ***Flow Balance:***

$$\frac{\lambda}{a_S(U-L)} = T_S f_S(t) q(\theta) \overline{F}_B(t), \quad \forall t \in [\tilde{V}_S(L), \tilde{V}_B(U)], \tag{4.45}$$

$$\frac{\alpha\lambda}{a_B(U-L)} = T_B f_B(\tilde{V}_B)h(\theta)F_S(\tilde{V}_B), \quad \forall t \in [\tilde{V}_S(L), \tilde{V}_B(U)]. \tag{4.46}$$

*(2)* ***Marginal Indifference:***

$$q(\theta)\gamma\left(\int_{\overline{V}_S}^{\tilde{V}_B(U)} f_B(u)\cdot u du - \overline{V}_S \int_{\overline{V}_S}^{\tilde{V}_B(U)} f_B(u)du\right) = c_S, \tag{4.47}$$

$$h(\theta)(1-\gamma)\left(\underline{V}_B \int_{\tilde{V}_S(L)}^{\underline{V}_B} f_S(u)du - \int_{\tilde{V}_S(L)}^{\underline{V}_B} f_S(u)\cdot u du\right) = c_B. \tag{4.48}$$

*(3)* ***Boundary Conditions:***

$$F_S(\overline{V}_S) = 1, \quad \overline{F}_B(\underline{V}_B) = 1. \tag{4.49}$$

*(4)* ***Taylor Approximation Based at the Bounds:***

$$a_S = \frac{1}{1+\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}q(\theta)\gamma\overline{F}_B(\overline{V}_S)}, \tag{4.50}$$

$$a_B = \frac{1}{1+\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}h(\theta)(1-\gamma)F_S(\underline{V}_B)}. \tag{4.51}$$

In the following we discuss about how to characterize a non full trade equilibrium, i.e., an equilibrium with $\overline{V}_S > \underline{V}_B$. We start with a lemma characterizing the bounds and the corresponding boundary conditions.

**Lemma 6.** *(Characterization of the Bounds) Under Assumptions 10 - 11, if in the equilibrium*

*$\overline{V}_S > \underline{V}_B$, i.e., in non full trade equilibrium, then the bounds should take values*

$$\overline{V}_S = \frac{1}{a_B + a_S - a_B a_S} \times \left( (1 - a_B) \sqrt{\frac{2(U-L) a_S T_B c_B}{\lambda(1-\gamma)}} - \sqrt{\frac{2(U-L) a_B T_S c_S}{\alpha \lambda \gamma}} + a_S (1 - a_B) L + a_B U \right), \tag{4.52}$$

$$\underline{V}_B = \frac{1}{a_B + a_S - a_B a_S} \times \left( \sqrt{\frac{2(U-L) a_S T_B c_B}{\lambda(1-\gamma)}} - (1 - a_S) \sqrt{\frac{2(U-L) a_B T_S c_S}{\alpha \lambda \gamma}} + a_S L + a_B (1 - a_S) U \right). \tag{4.53}$$

*Furthermore, the following boundary conditions should be satisfied,*

$$F_S(\underline{V}_B) = \sqrt{\frac{2\lambda c_B}{(U-L) a_S T_S q(\theta) h(\theta) (1-\gamma)}}, \tag{4.54}$$

$$\overline{F}_B(\overline{V}_S) = \sqrt{\frac{2\alpha\lambda c_S}{(U-L) a_B T_B h(\theta) q(\theta) \gamma}}. \tag{4.55}$$

The proof of Lemma 6 can be found in the Appendix. Given this proposition, the marginal agent types $\overline{V}_S, \underline{V}_B$ on the selling and buying side, respectively, are provided as functions of the other characteristics of the equilibrium. Note that the functional form of the equilibrium inventory distributions has also been characterized in §4.4.2 by Proposition 3. The following proposition gives a characterization of the remaining parameters - $T_S, T_B, \theta, a_S$, and $a_B$ - of the equilibrium in Definition 5.

**Proposition 4.** *(Non Full Trade Equilibrium Characterization) Under Assumptions 10 - 11, if in the equilibrium $\overline{V}_S > \underline{V}_B$, i.e., in non full trade equilibrium, then the equilibrium*

$(T_S, T_B, \theta, F_S(\cdot), \overline{F}_B(\cdot), \overline{V}_S, \underline{V}_B, a_S, a_B)$ *should satisfy the following system of equations:*

$$T_S = \left(\frac{a_S}{1-a_S}\right)^2 \left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2 \frac{2\alpha\lambda c_S\gamma}{(U-L)a_B}, \tag{4.56}$$

$$T_B = \left(\frac{a_B}{1-a_B}\right)^2 \left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2 \frac{2\lambda c_B(1-\gamma)}{(U-L)a_S}, \tag{4.57}$$

$$\theta = T_B/T_S = \frac{\frac{a_B^3}{(1-a_B)^2} c_B(1-\gamma)}{\frac{a_S^3}{(1-a_S)^2}\alpha c_S\gamma}, \tag{4.58}$$

$$\left(\frac{2\alpha\lambda c_S}{a_B(U-L)T_B h(\theta)q(\theta)\gamma}\right)^{a_B} = \left(\frac{2\lambda c_B}{a_S(U-L)T_S q(\theta)h(\theta)(1-\gamma)}\right)^{\alpha a_S}, \tag{4.59}$$

*when* $a_B \neq \alpha a_S$,

$$\begin{aligned}
&\frac{a_B-\alpha a_S}{a_B+a_S-a_Ba_S}\left(2c_B\frac{a_B}{1-a_B}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} - 2c_Sa_S\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} + a_SL + a_B(1-a_S)U\right)\\
&+\frac{\alpha a_S}{a_B}\cdot 2c_S\frac{a_S}{1-a_S}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\\
=&\frac{a_B-\alpha a_S}{a_B+a_S-a_Ba_S}\left(2c_Ba_B\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} - 2c_S\frac{a_S}{1-a_S}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} + a_S(1-a_B)L + a_BU\right)\\
&+2c_B\frac{a_B}{1-a_B}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}},
\end{aligned} \tag{4.60}$$

*when* $a_B = \alpha a_S$,

$$T_S = \frac{(U-L)\lambda\gamma a_S\alpha^2}{2c_S(1+\alpha-\alpha a_S)^2\left(W\left(\frac{q\left(\frac{c_S(1-\gamma)}{c_B\gamma}\right)\gamma}{2c_S}\cdot\frac{\alpha a_S(U-L)}{1+\alpha-\alpha a_S}\cdot e^{\frac{1+\alpha}{1+\alpha-\alpha a_S}}\right)\right)^2}. \tag{4.61}$$

The proof of Proposition 4 can be found in the Appendix. As a result, we can solve for the characteristics $T_S, T_B, \theta, a_S$, and $a_B$ of the equilibrium in Definition 5 by solving the system of equations (4.56) - (4.61). In particular, the depth of market on either side is characterized

as $T_S, T_B$, while $\theta$ captures the demand-supply ratio in the equilibrium.

Afterwards, the marginal agent type $\overline{V}_S, \underline{V}_B$ on the selling and buying side, respectively, can be determined by Lemma 6 as in equations (4.52) - (4.53). These cutoff values will determine who will join/exit in the mean-field steady state equilibrium of the market: sellers with valuations higher than $\overline{V}_S$ and buyers with valuations lower than $\underline{V}_B$ do not join.

Finally, when the above parameters are solved for, taking the boundary conditions of the inventory distributions $F_S(\cdot), F_B(\cdot)$ characterized by Lemma 6 as in equations (4.54) - (C.26) into its functional form in Proposition 3, the constants in the ODE solution therein can then be determined and thus the steady state equilibrium would be fully characterized. The resulting inventory distributions $F_S(\cdot), F_B(\cdot)$, as seen in Proposition 3, is of power law, or in some cases, exponential forms. In contrast with the uniform exogenous arrivals, the equilibrium waiting population has more of the unattractive types of the agents and thus presents a less amenable market condition, due to the effect of dynamics.

Moreover, by applying Little's Law, the characterization of the equilibrium inventory distributions $F_S(\cdot), F_B(\cdot)$ also leads to insights on the time-money tradeoff in dynamic matching markets. In particular, under the assumptions of uniform arrivals and linear strategies, it has been discussed at the beginning of §4.4.2 that the arriving distributions in terms of the dynamic valuations $\tilde{G}_S, \tilde{G}_B$ are also uniform. Therefore, the delay suffered by the various types, as their mass, will follow a power law, or exponential form, on either side. Sellers that want to set higher price (and similarly, buyers that want to set lower bid) wait longer. For downstream analysis, this characterization of the function that maps one's price/bid to her delay in transaction would offer important insight on the tactical decisions of buyers and sellers in dynamic matching markets.

## 4.5. Symmetric Case

Assume the selling side and buying side of the market are symmetric. That is, we assume that in the market primitives: $\alpha = 1$, $c_S = c_B = c$, and $\gamma = 1/2$, and we try to characterize a non full trade equilibrium in this setup. In this case, we can obtain a near closed-form characterization of the market equilibrium as shown in the following theorem.

**Theorem 7.** *(Symmetric Case Non Full Trade Equilibrium Characterization) Under Assumptions 10 - 11, together with the symmetry assumptions $\alpha = 1, c_S = c_B = c$, and $\gamma = 1/2$, if the common cost of waiting satisfies*

$$c < \frac{(U-L)q(1)}{4\left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} + \frac{2}{q(1)}\right)}, \tag{4.62}$$

*and the marketwise discount rate satisfies*

$$e^{-\beta\delta} < \frac{e^6}{e^6 + q(1)}, \tag{4.63}$$

*then there exists a unique non full trade equilibrium $(T_S, T_B, \theta, F_S(\cdot), \overline{F}_B(\cdot), \overline{V}_S, \underline{V}_B, a_S, a_B)$ that can be characterized as follows:*

$$a_B = a_S := a, \tag{4.64}$$

$$T_B = T_S = \frac{a}{(1-a)^2}\left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2 \frac{\lambda c}{(U-L)}, \quad \theta = 1, \tag{4.65}$$

$$F_S(t) = C_2 e^{\frac{\lambda t}{C_1 a(U-L)T_S q(1)}}, \quad \overline{F}_B(t) = \frac{C_1}{C_2} e^{-\frac{\lambda t}{C_1 a(U-L)T_S q(1)}}, \tag{4.66}$$

*where* $C_1 = \frac{2}{q(1)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}$, $C_2 = e^{-\frac{\frac{-2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}+(1-a)L+U}{(2-a)C_1cq(1)\left(\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2}}$,

$$\overline{V}_S = \left(-2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} + (1-a)L + U\right)/(2-a), \tag{4.67}$$

$$\underline{V}_B = \left(2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} + L + (1-a)U\right)/(2-a), \tag{4.68}$$

*with* $a \in (0,1)$ *solving the equation*

$$e^{2-\frac{(1-a)(U-L)}{2c\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}} = \left(\frac{2(1-a)}{aq(1)\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}\right)^{2-a}. \tag{4.69}$$

The proof of Theorem 7 can be found in the Appendix. Once the linear coefficient $a$ is determined from equation (4.69), all the other characteristics of the market equilibrium can be obtained by equations (4.64) - (4.68). That is, based on $a$, all the other entities can be determined explicitly. Theorem 7 connects the important measures in the market outcome to the market primitives more directly. The dynamic valuations $\tilde{V}_S, \tilde{V}_B$ and the inventory distributions $F_S, F_B$ can now be solved for without endogeneity or recursion.

### 4.5.1. Comparative Statics

Theorem 7 presents a near closed-form result and allows illustration on the effects of various market primitives - for example, meeting technology ($q(1)$), heterogeneity ($U - L$), search cost ($c$), and interest rate ($\beta$) - on the depth of market, price dispersion, delay in transaction, and join/exit behavior in the market outcome.

Specifically, we obtain a list of comparative statics results in the following proposition, by an analysis of equations (4.64) - (4.69). Its proof can be found in the Appendix.

**Proposition 5** (Comparative Statics). *Under the assumptions of Theorem 7, in the unique non full trade equilibrium shown to exist there, the following comparative statics results hold:*

*(i) If $q(1)$ increases: $a$ decreases, $T$ decreases, $F_S(\cdot), \overline{F}_B$ are steeper (have larger exponents), $\overline{V}_S$ increases, and $\underline{V}_B$ decreases.*

*(ii) If $U - L$ decreases: $a$ decreases, $F_S(\cdot), \overline{F}_B(\cdot)$ are steeper (have larger exponents), $\overline{V}_S$ decreases, $\underline{V}_B$ increases, and $F_S(\cdot), \overline{F}_B$ increases for active types.*

*(iii) If $c$ increases: $a$ decreases, $F_S(\cdot), \overline{F}_B$ are flatter (have smaller exponents), $\overline{V}_S$ decreases, and $\overline{V}_B$ increases.*

*(iv) If $e^{-\beta\delta}$ increases: $a$ decreases.*

Results in Proposition 5 provides the following insights into the operation of dynamic matching markets:

First, improved meeting technology allows more participation and faster trades but with lower benefits. Specifically, if the meeting technology becomes better, the probability to meet in a period - $q(1)$ in the symmetric case - should increase. In this case, the depth of the market decreases. Moreover, among the decreased liquidity, its distribution over valuation types $F_S, \overline{F}_B$ become steeper exponential functions, i.e., the selling distribution increases faster with cost and the buying distribution decreases faster with value. The proportion of unattractive types of agents in the equilibrium inventory distributions becomes relatively higher. At the same time, since $\overline{V}_S$ increases and $\underline{V}_B$ decreases, the bounds of participation become wider, admitting more agents. As for delay, since the total arrival increases while the market depth decreases, we expect delays to decrease. As for expected revenue from trade, since the distributions are steeper, they distribute more at the low value or high cost area, therefore their expected revenue decreases.

Second, decreased heterogeneity motivates faster trades but with lower benefits. Specifically, if the primitive valuation range $U - L$ decreases, the inventory distributions $F_S, \overline{F}_B$ become steeper exponential functions, i.e., the selling distribution increases faster with cost and the buying distribution decreases faster with value. The proportion of unattractive types of agents in the equilibrium inventory distributions becomes relatively higher. At the same time, since $\overline{V}_S$ decreases and $\underline{V}_B$ increases, the bounds for participation become tighter. As for delay, the probability to trade can be shown to increase for active types, and thus the expected delay generally decreases. As for revenue from trade, since the distributions are steeper, they distribute more at the low value or high cost area, therefore their expected revenue decreases.

Third, increased search cost deviates some unattractive sellers and buyers from market participation and thus benefits those who still join. Specifically, if participation becomes more expensive for each period, for example, the sellers suffer higher carrying cost or the buyers pay more rent, the common cost of delay $c$ increases. In this case, the inventory distributions $F_S, \overline{F}_B$ become flatter exponential functions, i.e., the selling distribution increases slower with cost and the buying distribution decreases slower with value. The proportion of attractive types of agents in the equilibrium inventory distributions becomes relatively higher. At the same time, since $\overline{V}_S$ decreases and $\overline{V}_B$ increases, the bounds for participation become tighter. As for delay, now that the bounds for participation becomes tighter and the inventory distributions become flatter, the overlapping agents should expect higher probability to trade as now they are closer to the bounds where probability to trade is 1 and the probability to trade decays slower at the same time.

Fourth, if the level of interest rate in the market decreases, i.e., when the discount coefficient $e^{-\beta\delta}$ becomes closer to 1, we can find that the symmetric linear coefficient $a$ decreases.

### 4.5.2. Numerical Tests

In this section, we numerically illustrate that the linearly approximated equilibrium is close to the true market equilibrium when valuation range is moderate. The objective is to compare the non full trade equilibria as characterized in Definition 4 under the symmetric case and its linear approximation in Theorem 7.

We consider a symmetric case ($\alpha = 1, c_S = c_B, \gamma = 1/2$) with an arrival rate of $\lambda = 10$ units and buyers per week on the selling and buying side, respectively. The meeting technology is such that $q(1) = 0.1$, i.e., one meeting can be formed per week for every ten active sellers or buyers. Their utilities are discounted at weekly rate $e^{-\beta\delta} = .98$. Furthermore, for each week of participation, active sellers and buyers pay a search cost of $c_S = c_B = c = 300$ dollars for their failure to transact. Note that the above selection of parameters satisfy the conditions on $c, \beta$ as specified in (4.62) - (4.63) in Theorem 7. Under this setup, we consider three different ranges of valuation $U - L$/median price $= \pm 20\%, \pm 10\%, \pm 5\%$ around a median price of \$500K, respectively.

We first compute the true market equilibrium in Definition 4 without the linearization technique. Under the symmetric case, we want to find a tuple

$$(\overline{V}_S, \tilde{V}_S, F_S, \underline{V}_B, \tilde{V}_B, F_B, T_S, \theta) \in [L, U] \times \mathbb{C}[L, U] \times \mathbb{C}[L, U] \times [L, U] \times \mathbb{C}[L, U] \times \mathbb{C}[L, U],$$

so that it satisfies the boundary conditions of the ODEs, the marginal indifference equations, and also flow balance conditions in Definition 4. [3] See footnote 3 for a detailed description of the computation scheme.

[3] Among several computation schemes, the following is chosen because it is feasible and often converges: 1.initialize $V_B, \theta, \overline{F}_B(\overline{V}_S)$ by the corresponding linearization result; 2.infer $V_S$ from overall flow balance; 3.derive $T_S, C_1, C_2, F_S(\underline{V}_B)$ according to the solution to the ODEs; 4.integrate via ODE solver for later calculation of expectations; 5.find $\mathbb{E}_S(\tilde{V}_S), \mathbb{E}_B(\tilde{V}_B), \tilde{V}_S(L), \tilde{V}_B(U)$ using the above integral; 6.solve for updated $V_B, \theta, \overline{F}_B(\overline{V}_S)$ by marginal indifference conditions and the remaining overall flow balance condition.

Computation of the true market equilibrium is hard. This computation complexity price is paid for the lack of analytical tractability. In comparison with the computation in the linearization case later on, not only the computation here takes more cycles, within each cycle the computation is more complex, involving not only solution of systems of nonlinear equations, but also lambert function and integrations. It takes much longer time: linearization computations take at most 1-2 seconds, while general computations take at least minutes. When solving for the true market equilibrium, we need to solve for the true market equilibrium by nailing down a system of ODEs by boundary conditions while parameters in the ODEs themselves, the boundaries, and the boundary conditions are all variable. In the linearization case, the main challenge in computation is the solution of a system of nonlinear equations. That challenge is more of a technical problem. But in here, we search in a much larger space, and we solve a system of ODEs with almost nothing in closed form.

Also, we solve for the linearized symmetric equilibrium in Theorem 7 by solving the system of equations (4.64) - (4.69). We recursively find the value of $a$ in equation (4.69). After that, the other entities in the tuple $(T_S, T_B, \theta, F_S(\cdot), \overline{F}_B(\cdot), \overline{V}_S, \underline{V}_B, a_S, a_B)$ can then be determined explicitly.

We focus on two aspects in the market outcome: 1) agent's participation behavior; as discussed in §4.2.2, it captures the strategies of the sellers and buyers in dynamic matching markets under our setting. 2) depth of market. Table 4.1 highlights the accuracy in these measures of the linearized symmetric equilibrium as an approximation to the true market equilibrium. We see that the difference between the true market equilibrium and its linear approximation is in general small when agents' valuations do not vary dramatically from the median price. For example, a variation of $\pm 20\%$ around the market median can be deemed practical when considering the house search among a few neighboring zip codes. That difference further diminishes when the valuation range gets smaller. And, in this specific

| $U - L$/median | $\Delta$participation rate | $\Delta\text{depth}^{(L)}$/depth |
|---|---|---|
| $\pm 20\%$ | 3.8% | 25% |
| $\pm 10\%$ | 0.3% | 2% |
| $\pm 5\%$ | 0.0% | 0% |

*weekly: median price=\$500K, $C_S = C_B$ =\$300, $\gamma = 0.5$, discount= 0.98, $q(1) = 0.1$, arrival rate=10.

**Table 4.1:** Numerical comparison of the true market equilibrium and the linearized equilibrium in the symmetric case. Tests with different primitive valuation ranges show that the linearly approximated results are close to true market equilibrium when valuation range is moderate.

numerical experiment, the two equilibria are identical when the agents' idiosyncratic valuation differ by at most 10% of the market median price.

# Bibliography

A. Abdulkadiroğluand, P. Pathak, and A. Roth. The new york city high school match. *American Economic Review*, 95(2):364–367, 2005. 112

P. Afeche. Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing & Service Operations Management*, 15(3):423–443, 2013. 114

A. Alfonsi, A. Fruth, and A. Schied. Optimal execution strategies in limit order books with general shape functions. *Quantitative Finance*, 10(2):143–157, 2010. 14, 66

G. Allon and A. Federgruen. Competition in service industries. *Operations Research*, 55(1): 37–55, 2007. 15

R. Almgren. Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied Mathematical Finance*, 10(1):1–18, 2003. 65

R. Almgren and N. Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3: 5–40, 2001. 65, 113

R. Almgren, C. Thum, E. Hauptmann, and H. Li. Direct estimation of equity market impact. *Risk*, 18(7):58–62, 2005. 67

A. Atakan. Efficient dynamic matching with costly search. TÜSİAD-Koç University Economic Research Forum working paper series, 2010. 112

M. Barclay, T. Hendershott, and T. McCormick. Competition among trading venues: Information and trading on electronic communications networks. *Journal of Finance*, 58(6): 2637–2666, 2003. 13

A. Bassamboo, J. Harrison, and A. Zeevi. Design and control of a large call center: Asymptotic analysis of an lp-based method. *Operations Research*, 54(3):419–435, 2006. 10, 36

G. Becker. A theory of marriage: Part I. *Journal of Political Economy*, 81(4):813–846, 1973. 112

D. Bergemann and M. Said. Dynamic auctions. *Wiley Encyclopedia of Operations Research and Management Science*, 2:1511–1522, 2011. 113

D. Bertsimas and A. Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1(1):1–50, 1998. 65, 113

O. Besbes and C. Maglaras. Revenue optimization for a make-to-order queue in an uncertain market environment. *Operations Research*, 57(6):1438–1450, 2009. 36

H. Bessembinder. Quote-based competition and trade execution costs in NYSE listed stocks. *Journal of Financial Economics*, 70:385–422, 2003. 13

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. 13

K. Binmore, A. Rubinstein, and A. Wolinsky. The nash bargaining solution in economic modelling. *The RAND Journal of Economics*, 17(2):176–188, 1986. 120

J. Blanchet and X. Chen. Continuous-time modeling of bid-ask spread and price dynamics in limit order books. Working paper, 2013. 14, 64

A. Bogomolnaia and H. Moulin. Random matching under dichotomous preferences. *Econometrica*, 72(1):257–279, 2004. 112

J. 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. 13

J. Bouchaud, J. Farmer, and F. Lillo. How markets slowly digest changes in supply and demand. In *Handbook of Financial Markets: Dynamics and Evolution*, pages 57–156. Elsevier: Academic Press, 2008. 66

M. Bramson. State space collapse with applications to heavy-traffic limits for multiclass queueing networks. *Queueing Systems*, 30:89–148, 1998. 10, 36

S. Buti, B. Rindi, Y. Wen, and I. Werner. Tick size regulation, intermarket competition and sub-penny trading. Working paper, 2011. 13

G. Cachon and P. Harker. Competition and outsourcing with scale economies. *Management Science*, 48(10):1314–1333, 2002. 15

G. Chacko, J. Jurek, and E. Stafford. The price of immediacy. *Journal of Finance*, 63(3): 1253–1290, 2008. ISSN 1540-6261. 66

Y-J. Chen, C. Maglaras, and G. Vulcano. Design of an aggregated marketplace under congestion effects: Asymptotic analysis and equilibrium characterization. Working paper, 2010. 15

R. Cont and A. Kukanov. Optimal order placement in limit order markets. Working paper, 2013. 14, 65

R. Cont and A. De Larrard. Price dynamics in a markovian limit order market. *SIAM Journal of Financial Mathematics*, 4(1):1–25, 2013. 14, 64

R. Cont, S. Stoikov, and R. Talreja. A stochastic model for order book dynamics. *Operations Research*, 58(3):549–563, 2010. 14, 64, 113

R. Cont, A. Kukanov, and S. Stoikov. The price impact of order book events. *Journal of Financial Econometrics*, 12(1):47–88, 2014. 67

H. Degryse, F. de Jong, and V. van Kervel. The impact of dark trading and visible fragmentation on market quality. Working paper, 2011. 13

P. Diamond. Aggregate demand management in search equilibrium. *Journal of Political Economy*, 90(5):881–894, 1982. 112

A. Dufour and R. F. Engle. Time and the price impact of a trade. *Journal of Finance*, 55 (6):2467–2498, 2000. 13

T. Foucault and A. Menkveld. Competition for order flow and smart order routing systems. *Journal of Finance*, 63(1):119–158, 2008. 13

T. Foucault, O. Kadan, and E. Kandel. Limit order book as a market for liquidity. *Review of Financial Studies*, 18(4):1171–1217, 2005. 13

D. Gale. *Strategic foundations of general equilibrium: Dynamic matching and bargaining games*. Cambridge University Press, 2000. 112

D. Gale and L. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962. 112

O. Garnett and A. Mandelbaum. An introduction to skills-based routing and its operational complexities. *Teaching notes*, 2000. 114

J. Gatheral. No-dynamic-arbitrage and market impact. *Quantitative Finance*, 10(7):749–759, 2010. 13, 66

D. Genesove and L. Han. Search and matching in the housing market. *Journal of Urban Economics*, 72(1):31–45, 2012. 112

L. Glosten. Components of the bid/ask spread and the statistical properties of transaction prices. *Journal of Finance*, 42(5):1293–1307, 1987. 13

L. Glosten. Is the electronic order book inevitable? *Journal of Finance*, 49(4):1127–1161, 1994. 13

L. Glosten. Competition, design of exchanges and welfare. Working paper, 1998. 13

L. Glosten and P. Milgrom. Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14:71–100, 1985. 13

M. Griffiths, B. Smith, D. Turnbull, and R. White. The costs and the determinants of order aggressiveness. *Journal of Financial Economics*, 56(1):65–88, 2000. 13

X. Guo, A. De Larrard, and Z. Ruan. Optimal placement in a limit order book. Working paper, 2013. 14, 65

J. Hamilton. Marketplace fragmentation, competition, and the efficiency of the stock exchange. *Journal of Finance*, 34(1):171–187, 1979. 13

J. Harrison. Brownian models of queueing networks with heterogeneous customer populations. In W. Fleming and P. 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. 14

J. 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. 14

J. Harrison. Brownian models of open processing networks: Canonical representation of workload. *The Annals of Applied Probability*, 16(3):1703?1732, 2006. 14

J. Harrison and M. Lopez. Heavy traffic resource pooling in parallel- server systems. *Queueing Systems*, 33:339–368, 1999. 15

R. Hassin and M. Haviv. *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston, MA, 2003. 15, 114

F. Hatheway. Nasdaq access fee experiment. Nasdaq, Technical report, 2015a. 52

F. Hatheway. Nasdaq access fee experiment. Nasdaq, Technical report, 2015b. 52

B. Hollifield, R. A. Millerz, and P. Sandas. Empirical analysis of limit order markets. *Review of Economic Studies*, 71(4):1027–1063, 2004. 13

R. Holthausen, R. 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. 13

A. Hosios. On the efficiency of matching and related models of search and unemployment. *The Review of Economic Studies*, 57(2):279–298, 1990. 112

G. Huberman and W. Stanzl. Price manipulation and quasi-arbitrage. *Econometrica*, 74(4): 1247–1276, 2004. 13, 66

G. Huberman and W. Stanzl. Optimal liquidity trading. *Review of Finance*, 9:165–200, 2005. 65

B. Jovanovic and A. Menkveld. Middlemen in limit-order markets. Working paper, 2011. 13

J. Nash Jr. The bargaining problem. *Econometrica*, 18(2):155–162, 1950. 120

D. Keim and A. Madhavan. The cost of instituional equity trades. *Financial Analysts Journal*, 54(4):50–59, 1998. 13

T. Kurtz. Strong approximation theorems for density dependent Markov chains. *Stochastic Processes Appl.*, 6(3):223–240, 1977/78. ISSN 0304-4149. 35

A. Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335, 1985. 13

P. Lakner, J. Reed, and S. Stoikov. High frequency asymptotics for the limit order book. Working paper, 2013. 14, 64

P. Lakner, J. Reed, and F. Simatos. Scaling limit of a limit order book model via the regenerative characterization of lévy trees. Working paper, 2014. 14, 64

M. A Lariviere. A note on probability distributions with increasing generalized failure rates. *Operations Research*, 54(3):602–604, 2006. 32

P. Lederer and L. Li. Pricing, production, scheduling and delivery-time competition. *Operations Research*, 45(3):407–420, 1997. 15

D. Levhari and I. Luski. Duopoly prcing and waiting lines. *European Economic Review*, 11: 17–35, 1978. 15

L. Li and Y. Lee. Pricing and delivery-time performance in a competitive environment. *Management Science*, 40(5):633–646, 1994. 15

I. Luski. On partial equilibrium in a queueing system with two servers. *The Review of Economic Studies*, 43(3):519–525, 1976. 15

C. Maglaras and C. Moallemi. A multiclass queueing model of limit order book dynamics. Working paper, 2011. 14

C. Maglaras, C. Moallemi, and H. Zheng. Queueing dynamics and state space collapse in fragmented limit order book markets. *Columbia Business School Research Paper*, (14-13), 2014. 113

C. Maglaras, C. Moallemi, and H. Zheng. Optimal execution in a limit order book and an associated microstructure market impact model. *Available at SSRN*, 2015a. 113

C. Maglaras, J. Yao, and A. Zeevi. Optimal price and delay differentiation in queueing systems. *Forthcoming Management Science*, 2015b. 114

K. Malinova and A. Park. Liquidity, volume, and price behavior: The impact of order vs. quote based trading. Working paper, 2010. 13

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. 15, 69

H. Mendelson and S. Whang. Optimal incentive-compatible priority pricing for the m/m/1 queue. *Operations Research*, 38(5):870–883, 1990. 15, 114

C. Moallemi, M. Saglam, and M. Sotiropoulos. Short-term predictability and price impact. Working paper, 2014. 67

D. Mortensen. Property rights and efficiency in mating, racing, and related games. *The American Economic Review*, 72(5):968–979, 1982. 112

D. Mortensen and C. Pissarides. Job creation and job destruction in the theory of unemployment. *The review of economic studies*, 61(3):397–415, 1994. 112

A. Obizhaeva and J. Wang. Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*, 16(1):1–32, 2013. 14, 66, 113

M. O'Hara and M. Ye. Is market fragmentation harming market quality? *Journal of Financial Economics*, 100(3):459–474, June 2011. 13

M. Ostrovsky. Stability in supply chain networks. *The American Economic Review*, 98(3): 897–923, 2008. 114

C. Parlour. Price dynamics in limit order markets. *Review of Financial Studies*, 11(4): 789–816, 1998. 14

C. Parlour and D. Seppi. Limit order markets: A survey. In A. Boot and A. Thakor, editors, *Handbook of Financial Intermediation & Banking*, pages 63–96. Elsevier Science, 2008. 13

P. Pathak and J. Sethuraman. Lotteries in student assignment: An equivalence result. *Theoretical Economics*, 6(1):1–17, 2011. 112

P. Pearson. Takeways from the NASDAQ pilor program. ITG Technical Report, 2015. 52

B. Petrongolo and C. Pissarides. Looking into the black box: A survey of the matching function. *Journal of Economic literature*, 39(2):390–431, 2001. 117

E. Plambeck and A. R. Ward. Optimal control of a high-volume assemble-to-order system. *Mathematics of Operations Research*, 31(3):453–477, 2006. 15

V. Rashkovich and A. Verma. Trade cost: Handicapping on PAR. *Journal of Trading*, 7(4), 2012. 67

C. René, K. Edward, and W. Gideon. FCFS infinite bipartite matching of servers and customers. *Advances in Applied Probability*, 41(3):695–730, 2009. 114

R. Rogerson, R. Shimer, and R. Wright. Search-theoretic models of the labor market-a survey. Technical report, National Bureau of Economic Research, 2004. 112

I. Rosu. A dynamic model of the limit order book. *Review of Financial Studies*, 22(11): 4601–4641, 2009. 14, 66

A. Roth. Axiomatic models of bargaining. *Lecture Notes in Economics and Mathematical Systems No.170*, 1979. 120

A. Roth. The college admissions problem is not equivalent to the marriage problem. *Journal of economic Theory*, 36(2):277–288, 1985. 112

A. Rubinstein and A. Wolinsky. Equilibrium in a market with sequential bargaining. *Econometrica*, 53(5):1133–1150, 1985. 112

M. Satterthwaite and A. Shneyerov. Dynamic matching, two-sided incomplete information, and participation costs: Existence and convergence to perfect competition. *Econometrica*, 75(1):155–200, 2007. 112

S. Shi. Frictional assignment: I. Efficiency. *Journal of Economic Theory*, 98(2):232–260, 2001. 112

R. Shimer and L. Smith. Assortative matching and search. *Econometrica*, 68(2):343–369, 2000. 112

R. Shimer and L. Smith. Matching, search, and heterogeneity. *Advances in Macroeconomics*, 1(1):1010–1029, 2001a. 112

R. Shimer and L. Smith. Nonstationary search. *University of Chicago and University of Michigan mimeo*, 2001b. 112, 122

K. So. Price and time competition for service delivery. *Manufacturing & Service Operations Management*, 2(4):392–409, 2000. 15

G. Sofianos. Specialist gross trading revenues at the New York Stock Exchange. Working paper, 1995. 13

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. 14

M. Sotiropoulos. Execution strategies in equity markets. In *High-Frequency Trading: New Realities for Traders, Markets and Regulators*, pages 21–42. Risk Books, 2013. 89

S. Stoikov, M. Avellaneda, and J. Reed. Forecasting prices from level-I quotes in the presence of hidden liquidity. *Algorithmic Finance, Forthcoming*, 2011. 14, 64

A. L. Stolyar. Optimal routing in output-queued flexible server systems. *Probability in the Engineering and Informational Sciences*, 19(2):141 – 189, 2005. 15

V. van Kervel. Liquidity: What you see is what you get? Working paper, 2012. 14

W. Whitt. A multi-class fluid model for a contact center with skill-based routing. *AEU-International Journal of Electronics and Communications*, 60(2):95–102, 2006. 114

S. Zak. *Systems and Control*. Oxford University Press, 2003. 174

# Appendix A

# Appendix to Chapter 2

## A.1. Proofs: Equilibrium Characterization

**Proof of Lemma 1.** 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,

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

where the last inequality follows from (2.17) and Assumption 1(i). This contradicts the flow balance equation (2.14).

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 (2.17) and Assumption 1(i). This contradicts the flow balance equation (2.14). Thus, we must have $\mathcal{L}(\gamma_0) = 0$ and (2.20) holds.

Now, suppose exchange $i$ achieves the maximum in (2.20). Then, from the right side of (2.20), it follows that $\kappa_i = \beta_i(\tilde{r}_i - \tilde{r}_0) = \frac{W^*}{\mu\gamma_0}$. Further, for any exchange $j$, (2.20) implies that $\kappa_j = \beta_j(\tilde{r}_j - \tilde{r}_0) \leq \frac{W^*}{\mu\gamma_0} = \kappa_i$. For the converse, if

$$\kappa_i = \max_{j \neq 0} \ \kappa_j, \tag{A.1}$$

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.1). ■

**Proof of Theorem 2.** Suppose $(\pi^*, W^*)$ satisfies (2.22)–(2.23). We want to show that $(\pi^*, W^*)$ is an equilibrium, i.e., it must satisfy (2.13)–(2.14).

We first establish (2.13). In particular, we will establish that for any $\pi \in \mathcal{P}$ and all $\gamma$,

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

Equivalently,

$$\sum_{i=1}^{N} \pi_i(\gamma) \left(\gamma\left(\tilde{r}_i - \tilde{r}_0\right) - \frac{W^*}{\mu\beta_i}\right) \leq \sum_{i=1}^{N} \pi_i^*(\gamma) \left(\gamma\left(\tilde{r}_i - \tilde{r}_0\right) - \frac{W^*}{\mu\beta_i}\right). \tag{A.2}$$

If $\gamma \leq \gamma_0$ and $i \neq 0$, using (2.22) and Assumption 1(iii), we have that

$$\gamma\left(\tilde{r}_i - \tilde{r}_0\right) - \frac{W^*}{\mu\beta_i} = \frac{\gamma\beta_i\kappa_i - \gamma_0 \max_{j\neq 0} \kappa_j}{\beta_i} \leq \frac{\gamma_0\beta_i\kappa_i - \gamma_0 \max_{j\neq 0} \kappa_j}{\beta_i} \leq 0 \tag{A.3}$$

Since, by (2.23), $\pi_i^*(\gamma) = 0$ for $i \neq 0$, we have that (A.2) holds for all $\gamma < \gamma_0$. For $\gamma = \gamma_0$, note that equality holds in (A.3) iff $\kappa_i = \max_{j\neq 0} \kappa_j$, i.e., $i \in \mathcal{A}^*(\gamma_0)$. Thus, (A.2) also holds

for $\gamma = \gamma_0$. Finally, if $\gamma > \gamma_0$ and $i \neq 0$,

$$\gamma\left(\tilde{r}_i - \tilde{r}_0\right) - \frac{W^*}{\mu\beta_i} = \frac{\gamma\kappa_i - \gamma_0 \max_{j\neq 0} \kappa_j}{\beta_i} \geq \frac{\gamma\kappa_i - \gamma \max_{j\neq 0} \kappa_j}{\beta_i} \geq 0. \tag{A.4}$$

Thus, (A.2) continues to hold.

Next, we establish (2.14). By (2.23), $1 - \pi_0^*(\gamma) = 0$ 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 (2.17),

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

Thus, $(\pi^*, W^*)$ satisfies (2.14) as well and is an equilibrium.

Now suppose $(\pi^*, W^*)$ is an equilibrium. We would like to show that $(\pi^*, W^*)$ must satisfy (2.22)–(2.23), except possibly for $\gamma$ 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 $W^*$, (2.22) follows immediately.

Next, we verify (2.23). Define $\mathcal{M}$ to be the set of $\gamma \geq 0$ such that $\pi^*(\gamma)$ does not satisfy (2.23). Define $\bar{\pi} \in \mathcal{P}$ to be a set of routing decisions such that $(\bar{\pi}, W^*)$ satisfies (2.23), such a $\bar{p}i$ can easily be constructed by solving the optimization problem for $\mathcal{A}^*(\gamma)$ for each $\gamma \geq 0$.

Define

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

for $\gamma \geq 0$. Following the same arguments as in (A.3)–(A.4), it is easy to see that

$$\begin{aligned} &\Delta(\gamma) = 0 \qquad \text{if } \gamma \notin \mathcal{M}, \\ &\Delta(\gamma) < 0 \qquad \text{if } \gamma \in \mathcal{M} \text{ and } \gamma \neq \gamma_0. \end{aligned} \tag{A.5}$$

On the other hand, Since $\pi^*$ is optimal for the program (2.13), we have that

$$0 \leq \int_0^\infty \Delta(\gamma)\, dF(\gamma) = \int_{\mathcal{M}} \Delta(\gamma)\, dF(\gamma) = \int_{\mathcal{M}\cap[0,\gamma_0)} \Delta(\gamma)\, dF(\gamma) + \int_{\mathcal{M}\cap(\gamma_0,\infty)} \Delta(\gamma)\, dF(\gamma), \tag{A.6}$$

where, for the final equality, we use the fact that the point $\{\gamma_0\}$ has $F$-measure zero under Assumption 1(i). Together, (A.5)–(A.6) imply that $\mathcal{M}$ has $F$-measure 0. ■

## A.2. Proofs: Equilibrium Convergence

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

We first provide the proof of Theorem 3, which establishes uniqueness of the equilibrium queue length vector $Q^*$.

**Proof of Theorem 3.** Suppose $(\pi^{(1)}, Q^{(1)})$ and $(\pi^{(2)}, Q^{(2)})$ are both equilibria. Define $W^{(\ell)} \triangleq \beta^\top Q^{(\ell)}$, for $\ell \in \{1, 2\}$. By Theorem 1, both $(\pi^{(1)}, W^{(1)})$ and $(\pi^{(2)}, W^{(2)})$ satisfy (2.13)-(2.14). By Theorem 2, we have that

$$W^{(1)} = W^{(2)} = W^* \triangleq \gamma_0\mu \max_{i\neq 0}\ \kappa_i. \tag{A.7}$$

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 $\gamma$ 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 $\gamma$

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,

$$\gamma(\tilde{r}_i - \tilde{r}_j) = \frac{W^*}{\mu\beta_i} - \frac{W^*}{\mu\beta_j}. \tag{A.8}$$

The right hand side of (A.8) is independent of $\gamma$, 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 $\gamma$. 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 $\gamma$ of at most $F$-measure zero.

Combining these facts with the flow balance condition (2.11), 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. ■

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

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

$$\dot{Q}_i(t) = \lambda_i + \Lambda\chi_i(W(t)) - \mu_i(Q(t)), \quad i = 1, \cdots, N. \tag{A.9}$$

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

The fact that the equilibrium queue length vector exhibits state space collapse and leads

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

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

$$\begin{cases} \dot{W}(t) = \Lambda\beta^\top\chi(W(t)) - \mu(\beta_1+\beta_2)\cdot\mathbb{I}_{\{W(t)\neq 0\}} + \mu\dfrac{\beta_1\beta_2 S(t)}{W(t)}\cdot\mathbb{I}_{\{W(t)\neq 0\}}, \\ \dot{S(t)} = \Lambda\mathbf{1}^\top\chi(W(t)) - \mu\cdot\mathbb{I}_{\{S(t)\neq 0\}}. \end{cases} \tag{A.10}$$

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

### Overview of the Proof for $(W(t), S(t))$ Convergence

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

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

Accordingly, the proof will be organized around the following main steps, each of which corresponds to one of Lemmas 8-10 in the following subsection:

1. Lemma 8 (Local Stability). There exists $\varepsilon > 0$, such that if $(W(0), S(0))$ is in the set

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

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

2. Lemma 9 (Finite Inter-arrival Time). Starting from any initial point, a sample path either enters the local stability region $\mathcal{W}_{local}$ or hits the set $\mathcal{W}^+$ in finite time; in the latter case, starting from any point in $\mathcal{W}^+$ the sample path must, in finite time, either

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

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

3. Lemma 10 (Guaranteed Decay). There exists $\varphi > 0$, such that if $\tau_1 < \tau_2$ are times where

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

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

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

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

### Proving $(W(t), S(t))$ Convergence

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

**Lemma 7** (Bounded Trajectory). *There exists $\zeta \in (0, \mu\beta_2)$ and $\underline{W}, \overline{W} \in [0, +\infty)$ with $\underline{W} < \overline{W}$, such that given initial conditions $S(0) = \mathbf{1}^\top Q(0)$ and $W(0) = \beta^\top Q(0)$ where $Q(0) \in \mathbb{R}^2_+$, there exists finite time $T_b \in [0, +\infty)$ such that at any time $t > T_b$,*

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

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

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

$$\begin{aligned} \mathbf{1}^\top Q &\leq \frac{\beta_1}{\beta_2} Q_1 + Q_2 = \frac{1}{\beta_2}\left(\beta^\top Q\right), \\ \mathbf{1}^\top Q &\geq Q_1 + \frac{\beta_2}{\beta_1} Q_2 = \frac{1}{\beta_1}\left(\beta^\top Q\right). \end{aligned} \tag{A.11}$$

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

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

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

$$\begin{aligned} \dot{W} &= \Lambda\beta^\top\chi(W) - \mu(\beta_1 + \beta_2) + \mu\beta_1\beta_2\frac{S}{W} \\ &\leq -\zeta - \mu\beta_1 + \mu\beta_1\beta_2\frac{1}{\beta_2} = -\zeta. \end{aligned} \tag{A.12}$$

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

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

$$\begin{aligned} \dot{W} &= \Lambda\beta^\top\chi(W) - \mu(\beta_1 + \beta_2) + \mu\frac{\beta_1\beta_2 S}{W} \\ &\geq \zeta + \mu\beta_1\beta_2\frac{1}{\beta_1} = \zeta + \mu\beta_2. \end{aligned} \tag{A.13}$$

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

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

$$\Lambda\beta^\top \chi(W(t)) - \mu(\beta_1 + \beta_2) \leq \Lambda\beta^\top \chi(\underline{W}) - \mu(\beta_1 + \beta_2) \leq -\zeta. \tag{A.14}$$

■

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

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

$$\begin{aligned} \mathbf{1}^\top \chi(W) &= \chi_1(W) + \chi_2(W) \\ &= \mathsf{P}\left( \max_{i=1,2} \gamma\tilde{r}_i - \frac{W}{\mu\beta_i} > \gamma\tilde{r}_0 \right), \end{aligned} \tag{A.15}$$

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

$$\dot{S} = \dot{Q}_1 + \dot{Q}_2 = \Lambda\mathbf{1}^\top \chi(W^*) - \mu = 0. \tag{A.16}$$

Thus,

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

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

$$\bar{S}(W) \triangleq \frac{W}{\mu\beta_1\beta_2} \left( \mu(\beta_1 + \beta_2) - \Lambda\beta^\top \chi(W) \right). \tag{A.17}$$

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

terms of $\bar{S}(W)$ as

$$\begin{aligned}\dot{W} &= \Lambda\beta^\top\chi(W) - \mu(\beta_1+\beta_2) + \mu\frac{\beta_1\beta_2\bar{S}(W)}{W} + \mu\frac{\beta_1\beta_2(S-\bar{S}(W))}{W} \\ &= \mu\frac{\beta_1\beta_2(S-\bar{S}(W))}{W}.\end{aligned} \tag{A.18}$$

Thus,

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

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

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

As indicated by the following lemma, the system of ordinary differential equations in (A.10) is locally asymptotically stable. So there exists a local stability region around the equilibrium such that all points inside the region converge.

**Lemma 8** (Local Stability). *There exists $\varepsilon > 0$, such that if $(W(0), S(0))$ is in the set*

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

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

**Proof**. For $W > 0, S > 0$, the Jacobian matrix corresponding to the system of ordinary differential equations in (A.10) is

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

Denote $\lambda_1, \lambda_2$ as its two eigenvalues, then

$$\lambda_1 + \lambda_2 = \mathrm{tr}(J(W,S)) = \Lambda\beta^\top\frac{\partial\chi(W)}{\partial W} - \mu\beta_1\beta_2\frac{S}{W^2} < 0. \tag{A.19}$$

$$\lambda_1 \cdot \lambda_2 = \det(J(W,S)) = -\Lambda\mu\beta_1\beta_2\mathbf{1}^\top\frac{\partial\chi(W)}{\partial W} > 0. \tag{A.20}$$

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

Now we are ready to set out the argument for convergence. As laid out in the overview of Section A.2, the next step is to show that the trajectory returns to the set $\mathcal{W}^+$ and thus to region 1 with finite inter-arrival time as long as it does not enter the local stability region.

**Lemma 9** (Finite Interarrival Time). *There exists finite time $T_r \in (0, +\infty)$, such that for any $(W(0), S(0)) \in \mathbb{R}^2_+ \bigcap \mathcal{C} \bigcap \mathcal{B}$, there exists time $0 < t < T_r$ where $(W(t), S(t)) \in \mathcal{W}^+$ or $(W(t), S(t)) \in \mathcal{W}_{local}$.*

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

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

  Since $\bar{S}(W)$ is continuous and $\bar{S}(W^*) = S^*$, for $\varepsilon^- \in (0, \varepsilon)$, there exists a small $\delta_{\varepsilon^-} > 0$ such that for $W \in (W^*, W^* + \delta_{\varepsilon^-})$, $|\bar{S}(W) - S^*| < \varepsilon^-$. For $W \in (W^*, W^* + \min\{\delta_{\varepsilon^-}, \varepsilon\})$,

if $S - S^* < \varepsilon$, then the trajectory converges because of local stability. Otherwise, without entering the local stability region, for these $W$ values, $\dot{W} = \mu\beta_1\beta_2\frac{S-\bar{S}(W)}{W} > \mu\beta_1\beta_2\frac{\varepsilon-\varepsilon^-}{W^*+\varepsilon}$. So the trajectory will exceed $W^* + \min\{\delta_{\varepsilon^-}, \varepsilon\}$ within finite time.

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

$$S_\delta(S, W) \triangleq S - \bar{S}(W). \tag{A.21}$$

Since $W > \underline{W}$,

$$\bar{S}'(W) = \frac{1}{\mu\beta_1\beta_2}\left(\mu(\beta_1+\beta_2) - \Lambda\beta^\top\chi(W)\right) - \frac{W}{\mu\beta_1\beta_2}\Lambda\beta^\top\frac{\partial\chi(W)}{\partial W} > 0. \tag{A.22}$$

Then,

$$\dot{S}_\delta = \dot{S} - \bar{S}'(W)\cdot\dot{W} < \Lambda\mathbf{1}^\top\chi(W) - \Lambda\mathbf{1}^\top\chi(W^*), \quad W > W^* + \min\{\delta_{\varepsilon^-}, \varepsilon\}. \tag{A.23}$$

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

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

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

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

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

  $$\dot{S} = \Lambda\mathbf{1}^\top\chi(W) - \Lambda\mathbf{1}^\top\chi(W^*), \quad W > W^* + \min\{\delta_{\varepsilon^-}, \varepsilon\}. \tag{A.24}$$

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

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

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

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

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

  $$\dot{S}_\delta = \dot{S} - \bar{S}'(W) \cdot \dot{W} > \Lambda \mathbf{1}^\top \chi(W) - \Lambda \mathbf{1}^\top \chi(W^*), \quad W < W^* - \min\{\delta'_{\varepsilon^-}, \varepsilon\}. \quad \text{(A.25)}$$

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

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

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

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

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

  $$\dot{S} = \Lambda \mathbf{1}^\top \chi(W) - \Lambda \mathbf{1}^\top \chi(W^*), \quad W < W^* - \min\{\delta'_{\varepsilon^-}, \varepsilon\}. \quad \text{(A.26)}$$

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

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

■

The final step is to show that between two successive times when the trajectory returns to region 1, the $S$ coordinate gets closer to the equilibrium value $S^*$. Furthermore, the step

size is bounded away from zero as long as the trajectory does not enter the local stability region.

**Lemma 10** (Guaranteed Decay). *There exists $\varphi > 0$, such that for any $(W(0), S(0)) \in \mathcal{W}^+ \bigcap \mathbb{R}^2_+ \bigcap \mathcal{C} \bigcap \mathcal{B}$, If $t_1 > 0$ is a time with $(W(t_1), S(t_1)) \in \mathcal{W}^+/\mathcal{W}_{local}$, then $S(0) - S(t_1) > \varphi$.*

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

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

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

$$t_4(W(\tau)) \triangleq \inf\{s > t_3 : W(s) = W(t)\}, \quad \tau \in [0, t_3]. \tag{A.27}$$

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

$$\dot{W}(\tau) = \dot{W}(t_4(W(\tau))) \cdot t_4'(W(\tau)) \cdot \dot{W}(\tau), \quad \tau \in [0, t_3]. \tag{A.28}$$

$$t_4'(W(\tau)) = \frac{1}{\dot{W}_{t_4(W(\tau))}}, \quad \tau \in [0, t_3). \tag{A.29}$$

We define the following function,

$$F(\tau) \triangleq S(\tau) + S(t_4(W(\tau))) - 2\bar{S}(W)(\tau), \quad \tau \in [0, t_3]. \tag{A.30}$$

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

$F(\nu) \leq 0, \forall \nu \in [\tau, t_3)$, then,

$$\begin{aligned}
\dot{F}(\nu) &= \dot{S}(\nu) + \dot{S}(t_4(W_\nu)) \cdot t_4'(W(\nu)) \cdot \dot{W}(\nu) - 2\bar{S}'(W(\nu)) \cdot \dot{W}(\nu) \\
&= \dot{S}(W(\nu)) + \dot{S}(W(\nu)) \cdot \frac{1}{\dot{W}(t_4(W_\nu))} \cdot \dot{W}(\nu) - 2\bar{S}'(W(\nu)) \cdot \dot{W}(\nu) \\
&= \dot{S}(W_\tau) \cdot \dot{W}_\tau \cdot \left( \frac{W_\tau}{\mu\beta_1\beta_2(S_\tau - \bar{S}(W_\tau))} + \frac{W_\tau}{\mu\beta_1\beta_2(S_{t_4(W_\tau)} - \bar{S}(W_\tau))} \right) - \bar{S}'(W_\tau) \cdot \dot{W}_\tau \\
&= \frac{\dot{S}(W(\nu)) \cdot \dot{W}(\nu) \cdot W(\nu)}{\mu\beta_1\beta_2} \cdot \frac{F(\nu)}{(S(\nu) - \bar{S}(W(\nu))) \cdot (S(t_4(W(\nu))) - \bar{S}(W(\nu)))} \\
&\quad - \bar{S}'(W(\nu)) \cdot \dot{W}(\nu) < 0.
\end{aligned} \tag{A.31}$$

Then,

$$F(t_3) - F(\tau) = \int_\tau^{t_3} \dot{F}(\nu) d\nu < 0, \tag{A.32}$$

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

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

$$\dot{F}(t_{equal}) = -\bar{S}'(W(t_{equal})) \cdot \dot{W}(t_{equal}) < 0 \tag{A.33}$$

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

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

Recall from equation (A.31) that

$$\begin{aligned}
\dot{F}(\tau) = &\frac{\dot{S}(\tau) \cdot \dot{W}(\tau) \cdot W(\tau)}{\mu\beta_1\beta_2} \cdot \frac{F(\tau)}{(S(\tau) - \bar{S}(W(\tau))) \cdot (S(t_4(W(\tau))) - \bar{S}(W(\tau)))} \\
&- \bar{S}'(W(\tau)) \cdot \dot{W}(\tau).
\end{aligned} \tag{A.34}$$

Define

$$G(\tau) \triangleq \frac{\dot{S}(\tau) \cdot \dot{W}(\tau) \cdot W(\tau)}{\mu\beta_1\beta_2(S(\tau) - \bar{S}(W(\tau))) \cdot (S(t_4(W(\tau))) - \bar{S}(W(\tau)))}, \quad \tau \in [0, t_3]. \tag{A.35}$$

Then,

$$\dot{F}(\tau) = G(\tau) \cdot F(\tau) - \bar{S}'(W(\tau)) \cdot \dot{W}(\tau). \tag{A.36}$$

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

$$0 < \varepsilon_G < \frac{\beta_2\zeta(\varepsilon - \varepsilon^-)}{W^*(W^* + \varepsilon)} - \frac{\beta_2\Delta}{W^*} \tag{A.37}$$

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

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

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

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

$$\begin{aligned}
\dot{F}(\tau) &< \varepsilon_G F(\tau) - \frac{\zeta*}{\mu\beta_1\beta_2}\frac{\mu\beta_1\beta_2(\varepsilon - \varepsilon^-)}{W^* + \varepsilon} \\
&< \left(\frac{\beta_2\zeta(\varepsilon - \varepsilon^-)}{W^*(W^* + \varepsilon)} - \frac{\beta_2\Delta}{W^*}\right) \cdot \frac{W^*}{\beta_2} - \frac{\zeta(\varepsilon - \varepsilon^-)}{W^* + \varepsilon} \\
&= -\Delta.
\end{aligned} \tag{A.38}$$

Therefore,

$$F(0) > F(0) - F\left(\min\left\{t_G, \frac{(W^* + \min\{\delta, \varepsilon\})W^*}{\mu\beta_1\beta_2\varepsilon}\right\}\right) > \Delta \cdot \min\left\{t_G, \frac{(W^* + \min\{\delta, \varepsilon\})W^*}{\mu\beta_1\beta_2\varepsilon}\right\}. \tag{A.39}$$

We can make analogous claims on $(S^* - S(t_5)) - (S(t_1) - S^*)$, i.e., on the other half of the trajectory from region 5 back to region 1, and thus prove that in each cycle the trajectory gets closer to the equilibrium with a positive step size as long as it does not enter the local

stability region and therefore has to converge. ■

## A.3. Auxiliary empirical results

| | % of Variance Explained | | | % of Variance Explained | |
|---|---|---|---|---|---|
| | One Factor | Two Factors | | One Factor | Two Factors |
| Alcoa | 62% | 77% | JPMorgan | 68% | 82% |
| American Express | 68% | 80% | Kraft | 74% | 84% |
| Boeing | 52% | 66% | Coca-Cola | 71% | 82% |
| Bank of America | 73% | 84% | McDonalds | 64% | 76% |
| Caterpillar | 31% | 51% | 3M | 31% | 51% |
| Cisco | 76% | 87% | Merck | 76% | 86% |
| Chevron | 38% | 59% | Microsoft | 74% | 90% |
| DuPont | 59% | 74% | Pfizer | 76% | 84% |
| Disney | 74% | 83% | Procter & Gamble | 72% | 81% |
| General Electric | 80% | 91% | AT&T | 69% | 81% |
| Home Depot | 85% | 92% | Travelers | 75% | 85% |
| Hewlett-Packard | 71% | 84% | United Tech | 39% | 55% |
| IBM | 27% | 53% | Verizon | 76% | 87% |
| Intel | 74% | 86% | Wal-Mart | 77% | 85% |
| Johnson & Johnson | 71% | 82% | Exxon Mobil | 54% | 69% |

**Table A.1:** Results of PCA for queue lengths trajectories: how much variance in the data can the first two principle components explain.

# Appendix B

# Appendix to Chapter 3

## B.1. Proofs

**Proof of Lemma 2.** Without loss of generality, we consider the evolution of the buy limit order queues $Q^b(t) = (Q^b_1(t), \ldots, Q^b_N(t))$.

For an arbitrary initial condition $Q(0) \in \mathcal{Q}$, the fluid model ODEs in (3.1) are initialized at $Q^b(0) \in \mathbb{R}^N_+$, satisfying

$$Q^b_{b_0}(0) > 0; \quad Q^b_i(0) = 0 \text{ for all } b_0 < i \leq N.$$

Starting with best-bid $b_0$ at time $t = 0$, at least for small $t$, the fluid model ODEs in (3.1) can be specified as follows:

$$\begin{aligned} \forall 1 \leq i < b_0: &\quad \dot{Q}^b_i(t) = \lambda^b_i - \gamma Q^b_i(t), \\ i = b_0: &\quad \dot{Q}^b_{b_0}(t) = \lambda^b_{b_0} - \mu^s_{b_0} - \gamma Q^b_{b_0}(t), \\ \forall b_0 < i \leq N: &\quad \dot{Q}^b_i(t) = 0, \end{aligned} \tag{B.1}$$

which has unique solution

$$\begin{aligned} \forall 1 \le i < b_0: \quad & Q_i^b(t) = \frac{\lambda_i^b}{\gamma}\left(1 - e^{-\gamma t}\right) + Q_i^b(0)e^{-\gamma t}, \\ i = b_0: \quad & Q_{b_0}^b(t) = \frac{\lambda_{b_0}^b - \mu_{b_0}^s}{\gamma}\left(1 - e^{-\gamma t}\right) + Q_{b_0}^b(0)e^{-\gamma t}, \\ \forall b_0 < i \le N: \quad & Q_i^b(t) = 0. \end{aligned} \tag{B.2}$$

From (B.2), for $b_0 < i \le N$, $Q_i^b(t)$ will stay at 0. Moreover, since $\lambda_{b_0}^b > \mu_{b_0}^s$ from Assumption 4, $Q_{b_0}^b(t)$ will stay positive and never hit the border $Q_{b_0}^b(t) = 0$. Therefore, $b_t = b_0$ for all $t \ge 0$. Analogously, $a_t = a_0$ for all $t \ge 0$.

As a result, (B.1) holds for all $t \ge 0$. Subsequently, (B.2) is the unique solution to the fluid model ODEs in (3.1) for all $t \ge 0$.

Since $Q(0) \in \mathcal{Q}$, $b_t = b_0 < a_0 = a_t$ for all $t \ge 0$ . And we have shown that $Q_{b_0}^b(t) > 0$, and analogously $Q_{a_0}^s(t) > 0$, for all $t \ge 0$. Hence, $Q(t) \in \mathcal{Q}$ for all $t \ge 0$.

Finally, as $t \to \infty$, $e^{-\gamma t} \to 0$. From (B.2), we have $Q^b(t) \to q^{*,b}$, with $q^{*,b}$ as given in (ii). ■

**Proof of Lemma 3.** If $C_{a_0} \le Q_{a_0}^s(0^-)$, we have that $\{S^*(t), t \in [0, \tau]\} = \{S^*(0) = C_{a_0}\}$ and it satisfies the constraints in (3.11) - (3.12). Executing immediately with one block trade is feasible and thus is the optimal solution to the minimum time problem.

If $C_{a_0} > Q_{a_0}^s(0^-)$, we start with the feasibility of the proposed control trajectory. From (3.13),

$$S^*(0) = Q_{a_0}^s(0^-) - \varepsilon,$$

and then $Q_{a_0}^s(0) = \varepsilon$. From (3.14), $\dot{Q}_{a_0}^s(t) = 0$ for all $t \in (0, \tau)$, which guarantees the queue length stays at $Q_{a_0}^s(t) = \varepsilon$. Furthermore, $\dot{S}^*(t) = \kappa_{a_0}$ for the length of the execution interval,

which is determined as $\tau = \left(C_{a_0} - Q^s_{a_0}(0^-)\right)/\kappa_{a_0}$. As a result,

$$S^*(\tau^-) = S^*(0) + \int_0^\tau r^*(t)dt = C_{a_0} - \varepsilon.$$

Finally, from (3.15), we have that $S^*(\tau) - S^*(\tau^-) = \varepsilon = Q^s_{a_0}(\tau^-)$ and $S^*(\tau) = C_{a_0}$.

We prove the optimality of the proposed trajectory by contradiction. Under control trajectory $\{S^*(t), t \in [0, \tau]\}$, we have that $\tau = \left(C_{a_0} - Q^s_{a_0}(0^-)\right)/\kappa_{a_0}$. Suppose there exists another feasible trajectory that executes $C_{a_0}$ shares within time $\tau' < \tau$.

Within time $[0, \tau']$, the total amount of newly arriving sell limit orders into price level $p_{a_0}$ is $\lambda^s_{a_0}\tau'$. From the first constraint in (3.12), $Q^s_{a_0}(t) \geq \varepsilon$ for all $t \in [0, \tau')$. The total amount of departed sell limit orders from price level $p_{a_0}$ is greater than or equal to

$$\mu^b_{a_0}\tau' + \gamma\varepsilon\tau'.$$

From the constraints in (3.12), any feasible trajectory can only submit market orders at price level $p_{a_0}$. Accordingly, the completed number of shares $C_{a_0}$ is constrained by the available liquidity at price level $p_{a_0}$ in the interval $[0, \tau']$, and thus is upper bounded as follows,

$$C_{a_0} \leq Q^s_{a_0}(0^-) + \lambda^s_{a_0}\tau' - \mu^b_{a_0}\tau' - \gamma\varepsilon\tau'. \tag{B.3}$$

As a result, $\tau' \geq \left(C_{a_0} - Q^s_{a_0}(0^-)\right)/\kappa_{a_0} = \tau$, which contradicts with the fact that $\tau' < \tau$. ■

**Proof of Lemma 4.** If $C' \leq Q^s_{a_0}(0^-) + \kappa T$, we have that

$$C^*_{a_0} = C', \quad C^*_i = 0 \text{ for } i = a_0 + 1, \cdots, N.$$

It is easy to verify that $C^*_{a_0}, \cdots, C^*_N$ is feasible. Furthermore, the resulting total price satisfies

$$\sum_{i=a_0}^{N} C^*_i \cdot p_i = C' \cdot p_{a_0} \leq \sum_{i=a_0}^{N} C_i \cdot p_i,$$

for any feasible $C_{a_0}, \cdots, C_N$, as $p_i \geq p_{a_0}$ for $i = a_0, \cdots, N$.

If $C' > Q^s_{a_0}(0^-) + \kappa T$, we have that

$$C^*_{a_0} = Q^s_{a_0}(0^-) + \kappa T, \quad C^*_i = Q^s_i(0^-) \text{ for } i = a_0 + 1, \cdots, n^* - 1,$$

where $n^*$ is defined as $n^* := \min\left\{a_0 \le j \le N : \ \kappa T + \sum_{k=a_0}^{j} Q^s_k(0^-) \ge C'\right\}$, and

$$C^*_{n^*} = C' - \kappa T - \sum_{i=a_0}^{n^*-1} Q^s_i(0^-), \quad C^*_i = 0 \text{ for } i > n^*.$$

In this execution policy, price $p_{n^*}$ will be the highest price at which the trader should submit market orders. It is easy to verify that $C^*_{a_0}, \cdots, C^*_N$ is feasible.

Furthermore, we prove by contradiction that there does not exist an optimal solution with lower total price. Suppose $C_{a_0}, \cdots, C_N$ is such an optimal solution, in which $p_n$ is the highest price to be used by the trader, i.e.,

$$C_i \ge Q^s_i(0^-) \text{ for } i < n, \quad C_n > 0, \quad C_i = 0 \text{ for } i > n.$$

We first show that $n = n^*$. On one hand, if $n < n^*$, from the definition of $n^*$, we will have

$$\kappa T < C' - \sum_{i=a_0}^{n} Q^s_i(0^-) = \sum_{i=a_0}^{n} \left(C_i - Q^s_i(0^-)\right) \le \sum_{i=a_0}^{n} \left(C_i - Q^s_i(0^-)\right)^+,$$

which contradicts with the time constraint. Hence, $n \ge n^*$. On the other hand, if $n > n^*$, and at the same time $\sum_{i=a_0}^{n} l_i < T$, then there exists $\eta > 0$ that simultaneously satisfies

$$C_n - \eta > 0, \quad \sum_{i=a_0}^{n} l_i + \frac{\eta}{\kappa} \le T, \quad \text{and } \eta \cdot (p_n - p_{a_0}) > 0. \tag{B.4}$$

In contrast to the original policy, let the trader submit $\eta$ less market orders at price $p_n$, and continuously submit market orders for $\eta/\kappa$ time more at price $p_{a_0}$. The latter policy is still feasible yet has strictly lower price, which contradicts with the fact that $C_{a_0}, \cdots, C_N$ is an

optimal solution. Therefore, in this case we should have

$$\sum_{i=a_0}^{n} \kappa l_i = \sum_{i=a_0}^{n-1} \left(C_i - Q_i^s(0^-)\right) + \left(C_n - Q_n^s(0^-)\right)^+ = \kappa T,$$

Subsequently, since $n > n^*$, we have that

$$\sum_{i=a_0}^{n-1} C_i + \left(C_n - Q_n^s(0^-)\right)^+ = \kappa T + \sum_{i=a_0}^{n-1} Q_i^s(0^-) \geq \kappa T + \sum_{i=a_0}^{n^*} Q_i^s(0^-) \geq C'.$$

However, since $\left(C_n - Q_n^s(0^-)\right)^+ < C_n$, the left hand side of the above inequality is strictly less than $C'$, which results in contradiction. Therefore, $n = n^*$.

For the policy $C_{a_0}, \cdots, C_N$, when $n = n^*$, the resulting total price satisfies

$$\begin{aligned}
\sum_{i=a_0}^{N} C_i \cdot p_i &= \sum_{i=a_0}^{n^*-1} \left(Q_i^s(0^-) + \kappa l_i\right) p_i + \left(C' - \sum_{i=a_0}^{n^*-1} \left(Q_i^s(0^-) + \kappa l_i\right)\right) \cdot p_{n^*} \\
&= C' \cdot p_{n^*} - \sum_{i=a_0}^{n^*-1} Q_i^s(0^-) \cdot (p_{n^*} - p_i) - \kappa \sum_{i=a_0}^{n^*-1} l_i \cdot (p_{n^*} - p_i) \\
&\geq C' \cdot p_{n^*} - \sum_{i=a_0}^{n^*-1} Q_i^s(0^-) \cdot (p_{n^*} - p_i) - \kappa T \cdot (p_{n^*} - p_{a_0}) \\
&= \sum_{i=a_0}^{N} C_i^* \cdot p_i,
\end{aligned}$$

which contradicts with the fact that it is an optimal solution with lower total price than that of the solution $C_{a_0}^*, \cdots, C_N^*$. ■

**Proof of Lemma 5.** Recall that $Q^0(t)$ denotes the quantity of limit orders at the best-bid with higher priority than the trader's order. Its dynamics have been given in (3.7). Under the assumptions in Section 3.4, from Lemma 4, we have that $b_t = b_0$ for all $t \in [0, T]$. As a result, until it gets depleted, the dynamics of $Q^0(t)$ can be simplified to

$$\dot{Q}^0(t) = -\mu_{b_0}^s - \gamma Q^0(t).$$

This ODE has a unique solution for $t \geq 0$ given by

$$Q^0(t) = -\frac{\mu^s_{b_0}}{\gamma} \cdot \left(1 - e^{-\gamma t}\right) + Q^0(0) \cdot e^{-\gamma t}.$$

Thus, the draining time of $Q^0(0)$ is

$$t_{\text{drain}} = \frac{1}{\gamma} \log\left(1 + \frac{\gamma}{\mu^s_{b_0}} Q^0(0)\right).$$

If $T \leq t_{\text{drain}}$, no limit orders submitted by the trader can be executed before the higher priority limit orders get depleted. In this event, $S_L = 0$.

If $T > t_{\text{drain}}$, for $t \in (t_{\text{drain}}, T]$, we have that $Q^0(0) = 0$. Recall that $Q^L$ denote the number of shares left in the trader's limit order. Its dynamics have been given in (3.8). For $t \in (t_{\text{drain}}, T]$, $\dot{Q}_L(t) = \mu_{b_t}$ if $Q_L(t) > 0$. Therefore, the maximum size of limit order $S_L$ the trader can execute within time $t \in (t_{\text{drain}}, T]$ is $\mu^s_{b_t} \cdot (T - t_{\text{drain}})$.

Moreover, since $S_L \leq C$,

$$S_L = \min\left\{\mu^s_{b_t} \cdot (T - t_{\text{drain}})^+, C\right\}.$$

■

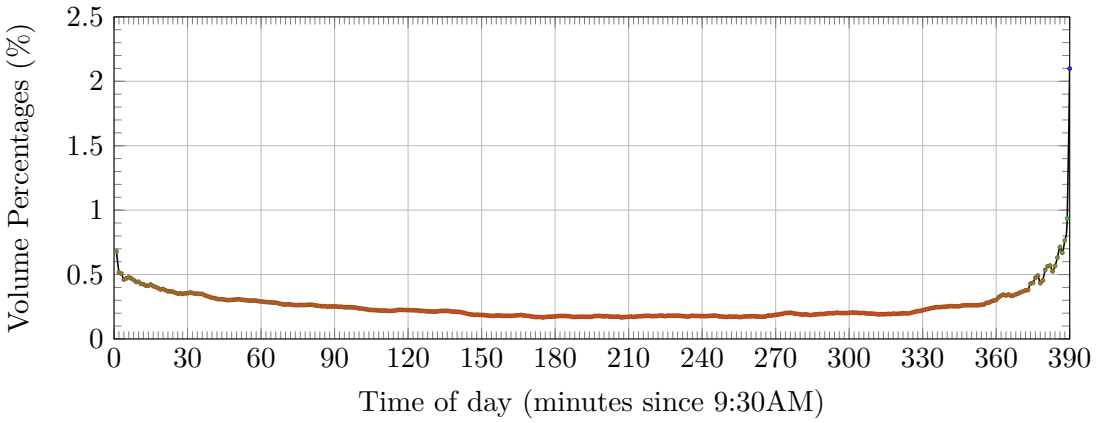



**Figure B.1:** S&P500 cross-sectional, smoothed intraday trading volume profile (min-by-min). Averaged across 5 consecutive trading days. A trading day in the US equities market starts at 9:30am and closes at 4:00pm, i.e., it has 390 minutes. This profile is indicative of "typical" days and it should be adjusted for special occasions such as option expirations, end of month, end of quarter, index rebalancing, Fed announcements, etc.; we do not include that level of granularity in our forecasts but instead apply the typical profile throughout the period of our sample and for all securities, including the ones that are not in the S&P500 and ETFs.

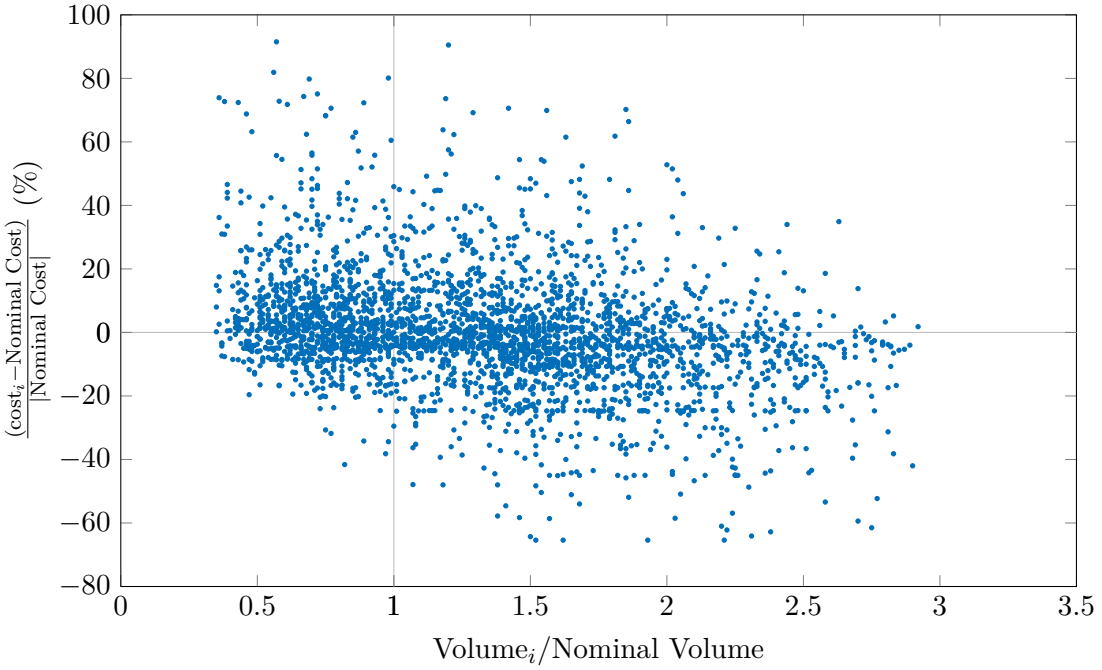



**Figure B.2:** Simulated costs as microstructure variables are varied. Order size = 3 × nominal queue length. Microstructure variables including queue lengths and market order arrival rates vary by a random multiplier in (1/3, 1) w.p. .5 and (1, 3) w.p. .5.

# Appendix C

# Appendix to Chapter 4

## C.1. Proofs

**Proof of Proposition 1.** Rewrite the characterization equations of the dynamic value functions in (4.11) (4.12) and denote the differences between the two sides as

$$E_S(V_S, \tilde{V}_S(V_S)) := -\tilde{V}_S(V_S) + V_S + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} \left( q(\theta)\gamma \int_{\tilde{V}_S(V_S)}^{U} \left( s - \tilde{V}_S(V_S) \right) dF_B(s) - c_S \right),$$

$$\begin{aligned} E_B(V_B, \tilde{V}_B(V_B)) := & -\tilde{V}_B(V_B) + V_B \\ & - \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} \left( h(\theta)(1-\gamma) \int_{L}^{\tilde{V}_B(V_B)} \left( \tilde{V}_B(V_B) - t \right) dF_S(t) - c_B \right). \end{aligned}$$

For the sellers, by integration by parts,

$$\begin{aligned} E_S(V_S, \tilde{V}_S(V_S)) = & -\tilde{V}_S(V_S) + V_S \\ & + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} \left( q(\theta)\gamma \left( (U - \tilde{V}_S(V_S)) F_B(U) - 0 - \int_{\tilde{V}_S(V_S)}^{U} F_B(s) ds \right) - c_S \right) \\ = & -\tilde{V}_S(V_S) + V_S \\ & + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} \left( q(\theta)\gamma \left( U - \tilde{V}_S(V_S) - \int_{\tilde{V}_S(V_S)}^{U} F_B(s) ds \right) - c_S \right). \end{aligned}$$

By chain rule, when conditions of implicit function theorem are satisfied, the derivative of the dynamic value functions can be derived from their characterization equations todo:check

implicit function theorem conditions

$$\begin{aligned}\frac{d\tilde{V}_S(V_S)}{dV_S} &= -\frac{\partial E_S/\partial V_S}{\partial E_S/\partial \tilde{V}_S(V_S)} \\ &= -\frac{1}{-1+\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}q(\theta)\gamma\left(-1-(-F_B(\tilde{V}_S(V_S)))\right)} \\ &= \frac{1}{1+\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}q(\theta)\gamma\overline{F}_B(\tilde{V}_S(V_S))} \in (0,1).\end{aligned} \tag{C.1}$$

Analogously for the buying side,

$$\frac{d\tilde{V}_B(V_B)}{dV_B} = -\frac{\partial E_B/\partial V_B}{\partial E_B/\partial \tilde{V}_B(V_B)} = \frac{1}{1+\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}h(\theta)(1-\gamma)F_S(\tilde{V}_B(V_B))} \in (0,1). \tag{C.2}$$

Now given the first derivatives of dynamic value functions $d\tilde{V}_S(V_S)/dV_S, d\tilde{V}_B(V_B)/dV_B$ and their signs, we examine the second derivatives,

$$\frac{d^2\tilde{V}_S(V_S)}{dV_S^2} = -\frac{\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}q(\theta)\gamma\frac{d\overline{F}_B(\tilde{V}_S(V_S))}{d\tilde{V}_S(V_S)}\frac{d\tilde{V}_S(V_S)}{dV_S}}{\left(1+\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}q(\theta)\gamma\overline{F}_B(\tilde{V}_S(V_S))\right)^2} \geq 0, \tag{C.3}$$

as $d\overline{F}_B(\tilde{V}_S(V_S))/d\tilde{V}_S(V_S) \leq 0$ and,

$$\frac{d^2\tilde{V}_B(V_B)}{dV_B^2} = -\frac{\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}h(\theta)(1-\gamma)\frac{dF_S(\tilde{V}_B(V_B))}{d\tilde{V}_B(V_B)}\frac{d\tilde{V}_B(V_B)}{dV_B}}{\left(1+\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}h(\theta)(1-\gamma)F_S(\tilde{V}_B(V_B))\right)^2} \leq 0. \tag{C.4}$$

as $dF_S(\tilde{V}_B(V_B))/d\tilde{V}_B(V_B) \geq 0$.

Therefore, $\tilde{V}_S(\cdot)$ is increasing and convex in $V_S$, and $\tilde{V}_B(\cdot)$ is increasing and concave in $V_B$. ■

**Proof of Proposition 2.** For simplicity of notation, let us denote $x(t) := F_S(t)$, $y(t) :=$

$\overline{F}_B(t)$. Rewrite the ODEs as

$$\begin{cases} x(t)' = \dfrac{\lambda}{U-L}\left(\dfrac{1}{T_S q(\theta) y(t)} + \dfrac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\dfrac{\gamma}{T_S}\right), \\ y(t)' = -\dfrac{\alpha\lambda}{U-L}\left(\dfrac{1}{T_B h(\theta) x(t)} + \dfrac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\dfrac{1-\gamma}{T_B}\right). \end{cases} \tag{C.5}$$

We prove that $(x(t), y(t))$ as given solves the system of ODEs in (C.5) by verification.

First, according to (4.23) and the definition of the Lambert function,

$$\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} h(\theta)(1-\gamma) x(t) e^{\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} h(\theta)(1-\gamma)x(t)} = C_1 y(t)^{-\frac{1}{\alpha}} e^{-\frac{1}{\alpha}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma y(t)}. \tag{C.6}$$

Divide $\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} h(\theta)(1-\gamma)$ and take logarithm on both sides of the above equation, we obtain

$$\begin{aligned} \log x(t) + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} h(\theta)(1-\gamma)x(t) = &\log\left(\frac{C_1}{\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} h(\theta)(1-\gamma)}\right) - \frac{1}{\alpha}\log y(t) \\ &- \frac{1}{\alpha}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma y(t). \end{aligned} \tag{C.7}$$

Take derivative on both sides of the above equation and reorganize, we can get

$$x(t)'\left(\frac{1}{x(t)} + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} h(\theta)(1-\gamma)\right) = y(t)'\left(-\frac{1}{\alpha y(t)} - \frac{1}{\alpha}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma\right). \tag{C.8}$$

Hence, by multiplying $\frac{-\alpha\lambda}{(U-L)T_B h(\theta)} = \frac{-\alpha\lambda}{(U-L)T_S q(\theta)}$ to the two sides of the above equation respectively,

$$x(t)' \cdot -\frac{\alpha\lambda}{U-L}\left(\frac{1}{T_B h(\theta) x(t)} + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\frac{1-\gamma}{T_B}\right) = y(t)' \cdot \frac{\lambda}{U-L}\left(\frac{1}{T_S q(\theta) y(t)} + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\frac{\gamma}{T_S}\right) \tag{C.9}$$

As a result, now we only need to show any one of the two ODEs is satisfied.

We want to show that the second ODE is satisfied. Take $x(t)$ as given in the proposition

into the second ODE, we can get a first order nonlinear ODE

$$y'(t) = -\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} \cdot \frac{\alpha\lambda(1-\gamma)}{(U-L)T_B} \cdot \left( \frac{1}{W\left( C_1 y(t)^{-\frac{1}{\alpha}} e^{-\frac{1}{\alpha}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma y(t)} \right)} + 1 \right). \tag{C.10}$$

Take $y(t)$ as given in the proposition into (C.10), we can check that (C.10) is also satisfied. Therefore, $x(t), y(t)$ as given in the proposition solves the system of ODEs in (C.5), and thus solves the system of ODEs in (4.21). ■

**Proof of Proposition 3.** We prove that the given functional forms of $F_S(\cdot), F_B(\cdot)$ solve the system of ODEs in (4.41) by verification.

First, when $a_B \neq a_S$,

$$\begin{aligned} f_S(t) = dF_S(t)/dt &= C_1 \frac{a_B}{a_B - \alpha a_S} \left( \frac{a_B - \alpha a_S}{C_1 a_B} t + C_2 \right)^{\frac{\alpha a_S}{a_B - \alpha a_S}} \frac{a_B - \alpha a_S}{C_1 a_B} \\ &= \left( \frac{a_B - \alpha a_S}{C_1 a_B} t + C_2 \right)^{\frac{\alpha a_S}{a_B - \alpha a_S}} = \frac{\lambda}{a_S (U-L) T_S q(\theta) \overline{F}_B(t)}. \end{aligned} \tag{C.11}$$

Thus the first ODE is satisfied. At the same time,

$$\begin{aligned} f_B(t) &= -d\overline{F}_B(t)/dt \\ &= -\frac{\lambda}{a_S(U-L)T_S q(\theta)} \frac{-\alpha a_S}{a_B - \alpha a_S} \left( \frac{a_B - \alpha a_S}{C_1 a_B} t + C_2 \right)^{-\frac{a_B}{a_B - \alpha a_S}} \frac{a_B - \alpha a_S}{C_1 a_B} \\ &= \frac{\alpha\lambda}{a_B (U-L) T_B h(\theta) C_1 \left( \frac{a_B - \alpha a_S}{C_1 a_B} t + C_2 \right)^{\frac{a_B}{a_B - \alpha a_S}}} \\ &= \frac{\alpha\lambda}{a_B (U-L) T_B h(\theta) F_S(t)}. \end{aligned} \tag{C.12}$$

Therefore the send ODE is satisfied as well.

Second, when $a_B = \alpha a_S$,

$$\begin{aligned} f_S(t) = dF_S(t)/dt &= C_2 e^{\frac{\lambda t}{C_1 a_S (U-L) T_S q(\theta)}} \frac{\lambda}{C_1 a_S (U-L) T_S q(\theta)} \\ &= \frac{\lambda}{a_S (U-L) T_S q(\theta)} \frac{C_2}{C_1} e^{\frac{\lambda t}{C_1 a_S (U-L) T_S q(\theta)}} = \frac{\lambda}{a_S (U-L) T_S q(\theta) \overline{F}_B(t)}. \end{aligned} \tag{C.13}$$

Thus the first ODE is satisfied. At the same time,

$$\begin{aligned} f_B(t) = -d\overline{F}_B(t)/dt &= \frac{C_1}{C_2} e^{-\frac{\lambda t}{C_1 a_S (U-L) T_S q(\theta)}} - \frac{\lambda}{C_1 a_S (U-L) T_S q(\theta)} \\ &= \frac{\lambda}{a_S (U-L) T_S q(\theta) F_S(t)} = \frac{\alpha \lambda}{a_B (U-L) T_B h(\theta) F_S(t)}. \end{aligned} \tag{C.14}$$

Therefore the send ODE is satisfied as well. ■

**Proof of Lemma 6.** First, directly derive $\tilde{V}_S(L)$ and $\tilde{V}_B(U)$ from the linear relations,

$$\tilde{V}_S(L) = (1 - a_S)\overline{V}_S + a_S L, \tag{C.15}$$

$$\tilde{V}_B(U) = (1 - a_B)\underline{V}_B + a_B U. \tag{C.16}$$

When $\overline{V}_S > \underline{V}_B$, for the marginal sellers, indifference yields

$$\begin{aligned} q(\theta)\gamma \left( \int_{\overline{V}_S}^{\tilde{V}_B(U)} f_B(u) \cdot u du - \overline{V}_S \int_{\overline{V}_S}^{\tilde{V}_B(U)} f_B(u) du \right) &= c_S \\ q(\theta)\gamma \frac{\alpha \lambda}{(U-L) a_B T_B h(\theta)} \int_{\overline{V}_S}^{\tilde{V}_B(U)} (u - \overline{V}_S) du &= c_S. \end{aligned} \tag{C.17}$$

Therefore

$$\tilde{V}_B(U) - \overline{V}_S = (1 - a_B)\underline{V}_B + a_B U - \overline{V}_S = \sqrt{\frac{2(U-L) a_B T_S c_S}{\alpha \lambda \gamma}}. \tag{C.18}$$

Similarly, for the marginal buyers,

$$\begin{aligned} h(\theta)(1-\gamma) \left( \underline{V}_B \int_{\tilde{V}_S(L)}^{\underline{V}_B} f_S(u) du - \int_{\tilde{V}_S(L)}^{\underline{V}_B} f_S(u) \cdot u du \right) &= c_B \\ h(\theta)(1-\gamma) \frac{\lambda}{(U-L) a_S T_S q(\theta)} \int_{\tilde{V}_S(L)}^{\underline{V}_B} (\underline{V}_B - u) du &= c_B. \end{aligned} \tag{C.19}$$

Therefore

$$\underline{V}_B - \tilde{V}_S(L) = \underline{V}_B - (1-a_S)\overline{V}_S - a_S L = \sqrt{\frac{2(U-L)a_S T_B c_B}{\lambda(1-\gamma)}}. \tag{C.20}$$

Combining equation (C.18) and equation (C.20), we get

$$\begin{aligned}\overline{V}_S =& \frac{1}{a_B + a_S - a_B a_S} \times \\ & \left((1-a_B)\sqrt{\frac{2(U-L)a_S T_B c_B}{\lambda(1-\gamma)}} - \sqrt{\frac{2(U-L)a_B T_S c_S}{\alpha\lambda\gamma}} + a_S(1-a_B)L + a_B U\right),\end{aligned} \tag{C.21}$$

$$\begin{aligned}\underline{V}_B =& \frac{1}{a_B + a_S - a_B a_S} \times \\ & \left(\sqrt{\frac{2(U-L)a_S T_B c_B}{\lambda(1-\gamma)}} - (1-a_S)\sqrt{\frac{2(U-L)a_B T_S c_S}{\alpha\lambda\gamma}} + a_S L + a_B(1-a_S)U\right).\end{aligned} \tag{C.22}$$

Furthermore, for agents $\tilde{V}_B > \overline{V}_S$ or $\tilde{V}_S < \underline{V}_B$, their densities are easy to get

$$f_S(\tilde{V}_S) = \frac{\lambda}{(U-L)a_S T_S q(\theta)}, \quad \forall \tilde{V}_S \in [\tilde{V}_S(L), \underline{V}_B], \tag{C.23}$$

$$f_B(\tilde{V}_B) = \frac{\alpha\lambda}{(U-L)a_B T_B h(\theta)}, \quad \forall \tilde{V}_B \in [\overline{V}_S, \tilde{V}_B(U)]. \tag{C.24}$$

Therefore,

$$F_S(\underline{V}_B) = \frac{\lambda}{(U-L)a_S T_S q(\theta)}(\underline{V}_B - \tilde{V}_S(L)) = \sqrt{\frac{2\lambda c_B}{(U-L)a_S T_S q(\theta) h(\theta)(1-\gamma)}}, \tag{C.25}$$

$$\overline{F}_B(\overline{V}_S) = \frac{\alpha\lambda}{(U-L)a_B T_B h(\theta)}(\tilde{V}_B(U) - \overline{V}_S) = \sqrt{\frac{2\alpha\lambda c_S}{(U-L)a_B T_B h(\theta) q(\theta)\gamma}}. \tag{C.26}$$

■

**Proof of Proposition 4.** From Lemma 6,

$$\overline{F}_B(\overline{V}_S) = \sqrt{\frac{2\alpha\lambda c_S}{(U-L)a_B T_B h(\theta) q(\theta)\gamma}}, \tag{C.27}$$

$$F_S(\underline{V}_B) = \sqrt{\frac{2\lambda c_B}{(U-L)a_S T_S q(\theta) h(\theta)(1-\gamma)}}. \tag{C.28}$$

Take these into the charactering equations of $a_B$ and $a_S$ (4.50) and (4.51), we can get

$$T_S = \left(\frac{a_S}{1-a_S}\right)^2 \left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2 \frac{2\alpha\lambda c_S\gamma}{(U-L)a_B}, \tag{C.29}$$

$$T_B = \left(\frac{a_B}{1-a_B}\right)^2 \left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2 \frac{2\lambda c_B(1-\gamma)}{(U-L)a_S}, \tag{C.30}$$

and therefore,

$$\theta = T_B/T_S = \frac{\frac{a_B^3}{(1-a_B)^2}c_B(1-\gamma)}{\frac{a_S^3}{(1-a_S)^2}\alpha c_S\gamma}. \tag{C.31}$$

When $a_B = \alpha a_S$, in this case from previous analysis,

$$F_S(t) = C_2 e^{\frac{\lambda t}{C_1 a_S(U-L)T_S q(\theta)}}, \tag{C.32}$$

$$\overline{F}_B(t) = \frac{C_1}{C_2} e^{-\frac{\lambda t}{C_1 a_S(U-L)T_S q(\theta)}}. \tag{C.33}$$

The bounds are now

$$\begin{aligned}\overline{V}_S =& \frac{1}{a_S(1+\alpha-\alpha a_S)}\times \\ & \left((1-\alpha a_S)\sqrt{\frac{2(U-L)a_S T_B c_B}{\lambda(1-\gamma)}} - \sqrt{\frac{2(U-L)a_S T_S c_S}{\lambda\gamma}} + a_S(1-\alpha a_S)L + \alpha a_S U\right),\end{aligned} \tag{C.34}$$

$$\begin{aligned}\underline{V}_B =& \frac{1}{a_S(1+\alpha-\alpha a_S)}\times \\ & \left(\sqrt{\frac{2(U-L)a_S T_B c_B}{\lambda(1-\gamma)}} - (1-a_S)\sqrt{\frac{2(U-L)a_S T_S c_S}{\lambda\gamma}} + a_S L + \alpha a_S(1-a_S)U\right).\end{aligned} \tag{C.35}$$

By overall flow balance,

$$\overline{V}_S + \alpha\underline{V}_B = \frac{1}{a_S}\left(\sqrt{\frac{2(U-L)a_S T_B c_B}{\lambda(1-\gamma)}} - \sqrt{\frac{2(U-L)a_S T_S c_S}{\lambda\gamma}} + a_S L + \alpha a_S U\right) = L + \alpha U, \tag{C.36}$$

$$\theta = \frac{T_B}{T_S} = \frac{c_S(1-\gamma)}{c_B\gamma}. \tag{C.37}$$

Therefore, the bounds can be further simplified to

$$\overline{V}_S = \frac{1}{1+\alpha-\alpha a_S}\left(-\alpha\sqrt{\frac{2(U-L)a_S T_S c_S}{\lambda\gamma}} + (1-\alpha a_S)L + \alpha U\right), \tag{C.38}$$

$$\underline{V}_B = \frac{1}{1+\alpha-\alpha a_S}\left(\sqrt{\frac{2(U-L)a_S T_S c_S}{\lambda\gamma}} + L + \alpha(1-a_S)U\right). \tag{C.39}$$

To solve for $C_1$,

$$\begin{aligned} F_S(\overline{V}_S)\cdot\overline{F}_B(\overline{V}_S) &= F_S(\underline{V}_B)\cdot\overline{F}_B(\underline{V}_B) = C_1 \\ &= \sqrt{\frac{2\lambda c_S}{(U-L)a_S T_S q(\theta)h(\theta)\theta\gamma}} = \sqrt{\frac{2\lambda c_B}{(U-L)a_S T_S q(\theta)h(\theta)(1-\gamma)}}. \end{aligned} \tag{C.40}$$

Therefore,

$$\left(\frac{2\alpha\lambda c_S}{a_B(U-L)T_B h(\theta)q(\theta)\gamma}\right)^{a_B} = \left(\frac{2\lambda c_B}{a_S(U-L)T_S q(\theta)h(\theta)(1-\gamma)}\right)^{\alpha a_S}. \tag{C.41}$$

To solve for $C_2$,

$$F_S(\overline{V}_S) = C_2 e^{\frac{\lambda\overline{V}_S}{C_1 a_S(U-L)T_S q(\theta)}} = 1, \tag{C.42}$$

$$C_2 = e^{-\frac{\lambda\overline{V}_S}{C_1 a_S(U-L)T_S q(\theta)}} = e^{-\frac{\overline{V}_S}{\sqrt{\frac{2(U-L)a_S T_S c_S}{\lambda\gamma}}}} = e^{\frac{\alpha}{1+\alpha-\alpha a_S} - \frac{(1-\alpha a_S)L+\alpha U}{1+\alpha-\alpha a_S}\cdot\frac{1}{\sqrt{\frac{2(U-L)a_S T_S c_S}{\lambda\gamma}}}}. \tag{C.43}$$

To solve for $T_S, T_B$,

$$F_S(\underline{V}_B) = \sqrt{\frac{2\lambda c_B}{(U-L)a_S T_S q(\theta)h(\theta)(1-\gamma)}} = C_2 e^{\frac{\lambda\underline{V}_B}{C_1 a_S(U-L)T_S q(\theta)}}, \tag{C.44}$$

therefore,

$$T_S = \frac{(U-L)\lambda\gamma a_S\alpha^2}{2c_S(1+\alpha-\alpha a_S)^2\left(W\left(\frac{q\left(\frac{c_S(1-\gamma)}{c_B\gamma}\right)^{\gamma}}{2c_S}\cdot\frac{\alpha a_S(U-L)}{1+\alpha-\alpha a_S}\cdot e^{\frac{1+\alpha}{1+\alpha-\alpha a_S}}\right)\right)^2}. \tag{C.45}$$

When $a_B \neq \alpha a_S$, from previous analysis

$$F_S(t) = C_1 \left( \frac{a_B - \alpha a_S}{C_1 a_B} t + C_2 \right)^{\frac{a_B}{a_B - \alpha a_S}}, \tag{C.46}$$

$$\overline{F}_B(t) = \frac{\lambda}{a_S(U-L)T_S q(\theta)} \left( \frac{a_B - \alpha a_S}{C_1 a_B} t + C_2 \right)^{-\frac{\alpha a_S}{a_B - \alpha a_S}}. \tag{C.47}$$

First, to solve for $C_1$,

$$F_S(\overline{V}_S)^{\frac{\alpha a_S}{a_B}} \cdot \overline{F}_B(\overline{V}_S) = \frac{C_1^{\frac{\alpha a_S}{a_B}} \lambda}{(U-L)a_S T_S q(\theta)} = \sqrt{\frac{2\alpha\lambda c_S}{(U-L)a_B T_B h(\theta) q(\theta)\gamma}}, \tag{C.48}$$

$$C_1 = \left( \frac{\alpha a_S}{a_B} \sqrt{\frac{2(U-L)a_B T_S c_S}{\alpha\lambda\gamma}} \right)^{\frac{a_B}{\alpha a_S}}. \tag{C.49}$$

Then, to solve for $C_2$,

$$\begin{aligned} F_S(\overline{V}_S) \cdot \overline{F}_B(\overline{V}_S) &= \frac{C_1 \lambda}{(U-L)a_S T_S q(\theta)} \left( \frac{a_B - \alpha a_S}{C_1 a_B} \overline{V}_S + C_2 \right) \\ &= \sqrt{\frac{2\alpha\lambda c_S}{(U-L)a_B T_B h(\theta) q(\theta)\gamma}}, \end{aligned} \tag{C.50}$$

and

$$C_2 = \frac{1}{C_1} \left( \frac{\alpha a_S}{a_B} \sqrt{\frac{2(U-L)a_B T_S c_S}{\alpha\lambda\gamma}} - \frac{a_B - \alpha a_S}{a_B} \overline{V}_S \right). \tag{C.51}$$

For any given $\theta$, to solve for $T_S, T_B$,

$$\overline{F}_B(\underline{V}_B)^{\frac{a_B}{\alpha a_S}} \cdot F_S(\underline{V}_B) = C_1 \cdot \left( \frac{\lambda}{(U-L)a_S T_S q(\theta)} \right)^{\frac{a_B}{\alpha a_S}} = \sqrt{\frac{2\lambda c_B}{(U-L)a_S T_S q(\theta) h(\theta)(1-\gamma)}}, \tag{C.52}$$

$$\left( \frac{2\alpha\lambda c_S}{(U-L)a_B T_B h(\theta) q(\theta)\gamma} \right)^{a_B} = \left( \frac{2\lambda c_B}{(U-L)a_S T_S q(\theta) h(\theta)(1-\gamma)} \right)^{\alpha a_S}. \tag{C.53}$$

Thus given $\theta$,

$$T_S = \frac{2\lambda}{(U-L)h(\theta)q(\theta)} \cdot \left( \frac{\alpha c_S}{a_B \gamma \theta} \right)^{\frac{a_B}{a_B - \alpha a_S}} \cdot \left( \frac{a_S(1-\gamma)}{c_B} \right)^{\frac{\alpha a_S}{a_B - \alpha a_S}}, \quad T_B = \theta T_S. \tag{C.54}$$

From the above equation we can also derive the following

$$q(\theta) = \frac{\left(\frac{\alpha a_S c_S(1-\gamma)}{a_B c_B \gamma \theta}\right)^{\frac{\alpha a_S}{2(a_B-\alpha a_S)}}}{\frac{a_S}{1-a_S}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\gamma} = \frac{\left(\frac{c_S}{c_B}\frac{\alpha a_S}{a_B}\frac{\frac{a_S}{1-a_S}}{\frac{a_B}{1-a_B}}\right)^{\frac{\alpha a_S}{a_B-\alpha a_S}}}{\frac{a_S}{1-a_S}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\gamma}. \tag{C.55}$$

In parallel,

$$h(\theta) = \frac{\left(\frac{c_S}{c_B}\frac{\alpha a_S}{a_B}\frac{\frac{a_S}{1-a_S}}{\frac{a_B}{1-a_B}}\right)^{\frac{\alpha a_B}{a_B-\alpha a_S}}}{\frac{a_B}{1-a_B}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}(1-\gamma)}. \tag{C.56}$$

Combine (C.31) and (C.55), given a functional form of $q(\theta)$, we can get one equation of $a_B, a_S$.

To solve for $a_B, a_S$, the other equation of them comes from

$$F_S(\underline{V}_B) \cdot \overline{F}_B(\underline{V}_B) = \sqrt{\frac{2\lambda c_B}{(U-L)a_S T_S q(\theta) h(\theta)(1-\gamma)}}, \tag{C.57}$$

which results in the following equation

$$\begin{aligned}
&\frac{a_B - \alpha a_S}{a_B + a_S - a_B a_S}\left(2c_B\frac{a_B}{1-a_B}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} - 2c_S a_S\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} + a_S L + a_B(1-a_S)U\right)\\
&+\frac{\alpha a_S}{a_B}\cdot 2c_S\frac{a_S}{1-a_S}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\\
=&\frac{a_B - \alpha a_S}{a_B + a_S - a_B a_S}\left(2c_B a_B\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} - 2c_S\frac{a_S}{1-a_S}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} + a_S(1-a_B)L + a_B U\right)\\
&+2c_B\frac{a_B}{1-a_B}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}.
\end{aligned} \tag{C.58}$$

■

**Proof of Theorem 7.** First, we want to show necessity. We start with showing that in the symmetric case $a_B = a_S$ has to be true. According to Lemma 6, by marginal indifference conditions,

$$F_S(\underline{V}_B) = \sqrt{\frac{2\lambda c}{(U-L)a_S T_S q(\theta) h(\theta)(1-\gamma)}} = \frac{2}{\sqrt{q(\theta)h(\theta)}\frac{a_S}{1-a_S}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}\sqrt{\frac{a_B}{a_S}}, \tag{C.59}$$

By the characterizing equation of $a_B$,

$$a_B = \frac{1}{1 + \frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} q(\theta)\gamma F_S(\underline{V}_B)} = \frac{1}{1 + \frac{\sqrt{\theta}}{\frac{a_S}{1-a_S}}\sqrt{\frac{a_B}{a_S}}}, \tag{C.60}$$

therefore,

$$\theta = \frac{\frac{a_S^3}{(1-a_S)^2}}{\frac{a_B^3}{(1-a_B)^2}} = \frac{1}{\theta}, \quad \theta = 1, \quad a_B = a_S, \tag{C.61}$$

From now on denote $a := a_B = a_S$. In this case $a_B = \alpha a_S$,

$$F_S(x) = C_2 e^{\frac{\lambda x}{C_1 a(U-L) T_S q(1)}} = C_2 e^{\frac{x}{C_1 cq(1)\left(\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2}}, \tag{C.62}$$

$$\overline{F}_B(x) = \frac{C_1}{C_2} e^{-\frac{\lambda x}{C_1 a(U-L) T_S q(\theta)}} = \frac{C_1}{C_2} e^{-\frac{x}{C_1 cq(1)\left(\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2}}. \tag{C.63}$$

By marginal indifference conditions, the bounds are now

$$\overline{V}_S = \left(-2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} + (1-a)L + U\right)/(2-a), \tag{C.64}$$

$$\underline{V}_B = \left(2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} + L + (1-a)U\right)/(2-a). \tag{C.65}$$

Also because of marginal indifference,

$$F_S(\underline{V}_B) = \overline{F}_B(\overline{V}_S) = \frac{2}{q(1)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}. \tag{C.66}$$

By boundary conditions,

$$F_S(\overline{V}_S) = C_2 e^{\frac{\overline{V}_S}{C_1 cq(1)\left(\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2}} = 1, \quad C_2 = e^{-\frac{\overline{V}_S}{C_1 cq(1)\left(\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2}}, \tag{C.67}$$

$$\overline{F}_B(\overline{V}_S) = \frac{C_1}{C_2} e^{-\frac{\overline{V}_S}{C_1 cq(1)\left(\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2}} = C_1 = \frac{2}{q(1)\frac{a}{1-a}\frac{e^{\beta\delta}}{1-e^{-\beta\delta}}}, \tag{C.68}$$

$$F_S(\underline{V}_B) = C_2 e^{\frac{\underline{V}_B}{C_1 cq(1)\left(\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2}} = e^{\frac{\underline{V}_B - \overline{V}_S}{C_1 cq(1)\left(\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2}} = \frac{2}{q(1)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}. \tag{C.69}$$

So the value of $a$ would be determined by the following equation

$$e^{\frac{\underline{V}_B - \overline{V}_S}{2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}} = e^{\frac{4c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} - a(U-L)}{2c(2-a)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}} = \frac{2}{q(1)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}, \tag{C.70}$$

$$e^{2-\frac{(1-a)(U-L)}{2c\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}} = \left(\frac{2(1-a)}{aq(1)\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}\right)^{2-a}. \tag{C.71}$$

Second, we want to show existence and uniqueness. In equation (C.71), denote the LHS and RHS as follows

$$E_L(a) := e^{2-\frac{(1-a)(U-L)}{2c\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}, \quad E_R(a) := \left(\frac{2}{q(1)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}\right)^{2-a}, \tag{C.72}$$

and $E(a) := E_L(a) - E_R(a)$. $E_L(a)$ is monotonically increasing in $a$ with derivative

$$E_L(a)' = \frac{U-L}{2c\frac{e^{-\beta\delta}}{1-e^{\beta\delta}}} e^{2-\frac{U-L}{2c\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}+\frac{U-L}{2c\frac{e^{-\beta\delta}}{1-e^{\beta\delta}}}a} > 0. \tag{C.73}$$

When $e^{-\beta\delta} < \frac{e^6}{q(1)+e^6}$, $E_R(a)$ is monotonically decreasing in $a$ with derivative

$$\begin{aligned} E_R(a)' &= -e^{(2-a)\left(\log\frac{2}{q(1)\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}+\log(1-a)-\log a\right)} \\ &\quad \times \left(\log\frac{2}{q(1)\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}} + \log(1-a) - \log a + \frac{2-a}{a(1-a)}\right) \\ &= -\left(\frac{2}{q(1)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}\right)^{2-a}\left(\log\frac{2}{q(1)\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}} + \log(1-a) - \log a + \frac{2-a}{a(1-a)}\right) < 0, \end{aligned} \tag{C.74}$$

because the function of $a$ within the second bracket is convex and

$$\min_{a\in(0,1)}\left\{\log\frac{2}{q(1)\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}+\log(1-a)-\log a+\frac{2-a}{a(1-a)}\right\}=6-\log\left(q(1)\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)>0, \tag{C.75}$$

which is taken at $a=2/3$, because $e^{-\beta\delta}<\frac{e^6}{q(1)+e^6}$, $q(1)\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}<e^6$.

At $a=0$ and $a=1$ respectively,

$$E_L(0)=e^{2-\frac{U-L}{2c\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}<\lim_{a\to 0}E_R(0)=\infty,\quad E_L(1)=e^2>E_R(1)=0. \tag{C.76}$$

Therefore $a\in(0,1)$ satisfying $E(a)=0$ exists and is unique.

Third, we want to show sufficiency. We start with showing $\overline{V}_S>\underline{V}_B$. When the bounds are as given,

$$\overline{V}_S-\underline{V}_B=\frac{-4c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}+a(U-L)}{2-a}. \tag{C.77}$$

Since $e^{2-\frac{(1-a)(U-L)}{2c\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}=\left(\frac{2(1-a)}{aq(1)\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}\right)^{2-a}$, take orders of $1/(2-a)$ on both sides we can get

$$e^{\frac{4c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}-a(U-L)}{2c(2-a)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}=e^{\frac{\underline{V}_B-\overline{V}_S}{2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}=\frac{2}{q(1)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}. \tag{C.78}$$

When $c<\frac{(U-L)q(1)}{4\left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}+\frac{2}{q(1)}\right)}$, given that $\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}$ is monotonically increasing in $a$, denote the solution to the equation $\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}=\frac{2}{q(1)}$ as $\tilde{a}$, then $\tilde{a}=\frac{2}{q(1)\left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}+\frac{2}{q(1)}\right)}$ and

$$\begin{aligned}E(\tilde{a})&=e^{2-\frac{\tilde{a}(U-L)q(1)}{4c}}-1\\&=e^{2\left(1-\frac{U-L}{4c\left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}+\frac{2}{q(1)}\right)}\right)}-1<0.\end{aligned} \tag{C.79}$$

Since $E(\cdot)$ in monotonically increasing in $a$, the equilibrium $a$ value is to the right of $\tilde{a}$, i.e., $a>\tilde{a}$. Because $\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}$ is also monotonically increasing in $a$, we have $\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}>\frac{2}{q(1)}$,

therefore $e^{\frac{\underline{V}_B-\overline{V}_S}{2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}} = \frac{2}{q(1)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}} < 1$, and thus $\underline{V}_B < \overline{V}_S$.

Now we proceed to show that all the requirements in Definition 4 are satisfied. According to Proposition 3, flow balances are satisfied by $F_S(\cdot), \overline{F}_B(\cdot)$ as given. When the bounds are as given,

$$\begin{aligned}\tilde{V}_B(U) - \overline{V}_S &= \underline{V}_B - \tilde{V}_S(L)) \\ &= \underline{V}_B - \overline{V}_S - a(L - \overline{V}_S) \\ &= \underline{V}_B - (1-a)\overline{V}_S - aL \\ &= 2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}.\end{aligned} \tag{C.80}$$

Since $\overline{V}_S > \underline{V}_B$,

$$\begin{aligned}&\int_{\overline{V}_S}^{\tilde{V}_B(U)} f_B(u)\cdot u du - \overline{V}_S \int_{\overline{V}_S}^{\tilde{V}_B(U)} f_B(u)du = \frac{\lambda}{a(U-L)Tq(1)}\frac{(\tilde{V}_B(U)-\overline{V}_S)^2}{2} \\ &= \frac{\lambda}{a(u-L)\frac{a}{(1-a)^2}\left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2\frac{\lambda c}{(U-L)}q(1)}\frac{4c^2\left(\frac{a}{1-a}\right)^2\left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2}{2} \\ &= \frac{c}{q(1)\frac{1}{2}} = \frac{c}{q(\theta)\gamma},\end{aligned} \tag{C.81}$$

so the marginal indifference condition holds for the marginal seller. We can make an analogous argument for the marginal buyer. When the distributions are as given,

$$\begin{aligned}F_S(\overline{V}_S) &= C_2 e^{\frac{\lambda\overline{V}_S}{C_1 a(U-L)Tq(1)}} \\ &= e^{-\frac{-2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}+(1-a)L+U}{(2-a)C_1 cq(1)\left(\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)}} \cdot e^{\frac{\lambda\left(-2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}+(1-a)L+U\right)}{(2-a)C_1 a(U-L)\frac{a}{(1-a)^2}\left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2\frac{\lambda c}{(U-L)}q(1)}} = 1,\end{aligned} \tag{C.82}$$

so the boundary condition holds for the selling distribution. We can make an analogous argument for the buying distribution. Also when the bounds are as given,

$$\overline{F}_B(\overline{V}_S) = \frac{\lambda}{a(U-L)Tq(1)}(\tilde{V}_B(U) - \overline{V}_S) = \frac{2}{q(1)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}, \tag{C.83}$$

$$\frac{1}{1+\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}q(1)\frac{1}{2}\overline{F}_B(\overline{V}_S)} = \frac{1}{1+\frac{1-a}{a}} = a, \tag{C.84}$$

so the characterizing equation of $a_S$ holds. We can make an analogous argument for the characterizing equation of $a_B$.

■

**Proof of Proposition 5.** Recall that $E_L(a) = e^{2-\frac{(1-a)(U-L)}{2c\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}$ is monotonically increasing in $a$ with derivative

$$E_L(a)' = \frac{U-L}{2c\frac{e^{-\beta\delta}}{1-e^{\beta\delta}}} e^{2-\frac{U-L}{2c\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}+\frac{U-L}{2c\frac{e^{-\beta\delta}}{1-e^{\beta\delta}}}a} > 0. \tag{C.85}$$

When $e^{-\beta\delta} < \frac{e^6}{q(1)+e^6}$, $E_R(a) = \left(\frac{2(1-a)}{aq(1)\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}\right)^{2-a}$ is monotonically decreasing in $a$ with derivative

$$\begin{aligned} E_R(a)' &= -e^{(2-a)\left(\log\frac{2}{q(1)\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}+\log(1-a)-\log a\right)} \\ &\quad \times \left(\log\frac{2}{q(1)\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}} + \log(1-a) - \log a + \frac{2-a}{a(1-a)}\right) \\ &= -\left(\frac{2}{q(1)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}\right)^{2-a}\left(\log\frac{2}{q(1)\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}} + \log(1-a) - \log a + \frac{2-a}{a(1-a)}\right) < 0. \end{aligned} \tag{C.86}$$

And at $a = 0$ and $a = 1$ respectively,

$$E_L(0) = e^{2-\frac{U-L}{2c\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}} < \lim_{a\to 0} E_R(0) = \infty, \quad E_L(1) = e^2 > E_R(1) = 0. \tag{C.87}$$

(i) When $q(1)$ increases, at each value of $a \in (0,1)$, $E_R(a)$ decreases. As such, at the original $a$, $E(a) = E_L(a) - E_R(a) > 0$ now. Since $E(a)$ is monotonically increasing, $a$ should decrease to make $E(a) = 0$ again. At the same time, since $E_L(\cdot)$ is monotonically increasing, $E_L(a) = E_R(a)$ decreases.

As for market depth,

$$T = \frac{a}{(1-a)^2}\left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2 \frac{\lambda c}{(U-L)}, \tag{C.88}$$

which decreases as $a$ decreases.

As for the distributions,

$$\begin{aligned} F_S(t) =& C_2 e^{\frac{\lambda t}{C_1 a(U-L)Tq(1)}} = C_2 e^{\frac{\lambda t}{\frac{2}{q(1)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}} a(U-L)\frac{a}{(1-a)^2}\left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2 \frac{\lambda c}{(U-L)}q(1)}} \\ =& C_2 e^{\frac{t}{2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}, \end{aligned} \tag{C.89}$$

and

$$\overline{F}_B(t) = \frac{C_1}{C_2} e^{-\frac{t}{2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}, \tag{C.90}$$

where $\frac{1}{2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}$ should increase as $a$ decreases. Therefore, the distributions become steeper, i.e., the selling distribution increases faster with high cost while the buying distribution decreases faster with high value. The market has more 'incompetetive' population waiting.

As for the bounds, since $E_R(a)$ decreases while $a$ decreases, $\frac{2}{q(1)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}$ decreases. Thus,

$$e^{\frac{-\overline{V}_S+\underline{V}_B}{2c}} = \left(\frac{2}{q(1)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}\right)^{\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}, \tag{C.91}$$

which is easy to see to be decreasing. At the same time, by overall flow balance, in the symmetric case $\overline{V}_S - L = U - \underline{V}_B$, or $\underline{V}_B + \overline{V}_S = U + L$, thus

$$\overline{V}_S = (\overline{V}_S - \underline{V}_B + \overline{V}_S + \underline{V}_B)/2 \text{ increases}, \quad \underline{V}_S = (\underline{V}_B + \overline{V}_S - (\overline{V}_S - \underline{V}_B))/2 \text{ decreases}. \tag{C.92}$$

So the bounds become wider, admitting more agents.

As for expected delay, since the total arrival increases while the market depth decreases, we expect delays to decrease.

As for expected revenue from trade, since the distributions are steeper, they distribute more at the low value or high cost area, therefore their expected revenue decreases.

(ii) When $U - L$ decreases, at each value of $a \in (0, 1)$, $E_L(a) = e^{2-\frac{(1-a)(U-L)}{2c\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}$ increases. As such, at the original $a$, $E(a) = E_L(a) - E_R(a) > 0$ now. Since $E(a)$ is monotonically increasing, $a$ should decrease to make $E(a) = 0$ again. At the same time, since $E_R(a)$

is monotonically decreasing, $E_L(a) = E_R(a)$ increases when $a$ decreases.

As for the distributions,

$$\begin{aligned} F_S(t) =& C_2 e^{\frac{\lambda t}{C_1 a(U-L)Tq(1)}} = C_2 e^{\frac{\lambda t}{\frac{2}{q(1)}\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} a(U-L)\frac{a}{(1-a)^2}\left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2 \frac{\lambda c}{(U-L)}q(1)}} \\ =& C_2 e^{\frac{t}{2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}, \end{aligned} \tag{C.93}$$

and

$$\overline{F}_B(t) = \frac{C_1}{C_2} e^{-\frac{t}{2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}, \tag{C.94}$$

where $\frac{1}{2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}$ should increase as $a$ decreases. Therefore, the distributions become steeper, i.e., the selling distribution increases faster with high cost while the buying distribution decreases faster with high value. The market has more 'incompetetive' population waiting.

As for the bounds, since $E_L(a)$ increases, $(1-a)(U-L)$ decreases. Thus,

$$\begin{aligned} \overline{V}_S - \underline{V}_B &= \left(-4c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} + a(U-L)\right)/(2-a) \\ &= \frac{a}{(1-a)(2-a)}\left(-4c\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} + (1-a)(U-L)\right), \end{aligned} \tag{C.95}$$

which is easy to see to be decreasing when both $a$ and $U-L$ are decreasing. At the same time, by overall flow balance, in the symmetric case $\overline{V}_S - L = U - \underline{V}_B$, or $\underline{V}_B + \overline{V}_S = U + L$, thus

$$\overline{V}_S = (\overline{V}_S - \underline{V}_B + \overline{V}_S + \underline{V}_B)/2 \text{ decreases}, \quad \underline{V}_S = (\underline{V}_B + \overline{V}_S - (\overline{V}_S - \underline{V}_B))/2 \text{ increases}. \tag{C.96}$$

So the bounds become narrower.

As for the probability to trade, since $\overline{V}_S$ decreases and $\frac{1}{2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}$ increases, $C_2 =$

$e^{-\frac{\overline{V}_S}{2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}$ should increase. Therefore, for each $t \in [\underline{V}_B, \overline{V}_S]$,

$$F_S(t) = C_2 e^{\frac{t}{2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}, \tag{C.97}$$

should increase, while for the non overlapping agents it remains 1. Therefore, expected delay generally decreases.

As for expected revenue from trade, since the distributions are steeper, they distribute more at the low value or high cost area, therefore their expected revenue decreases.

(iii) When $c$ increases, at each value of $a \in (0,1)$, $E_L(a) = e^{2-\frac{(1-a)(U-L)}{2c\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}$ increases. As such, at the original $a$, $E(a) = E_L(a) - E_R(a) > 0$ now. Since $E(a)$ is monotonically increasing, $a$ should decrease to make $E(a) = 0$ again.

As for the distributions, since $a$ decreases, $\frac{2}{q(1)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}$ increases. Since $E_L(a) = E_R(a)$ and the base of $E_R(a)$ becomes closer to $e$, we expect the difference in the exponent to be decreasing as well, and thus $\frac{c}{U-L}\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}$ increases. Therefore,

$$\begin{aligned} F_S(t) =& C_2 e^{\frac{\lambda t}{C_1 a(U-L)Tq(1)}} = C_2 e^{\frac{\lambda t}{\frac{2}{q(1)\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}} a(U-L)\frac{a}{(1-a)^2}\left(\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}\right)^2\frac{\lambda c}{(U-L)}q(1)}} \\ =& C_2 e^{\frac{t}{2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}, \end{aligned} \tag{C.98}$$

and

$$\overline{F}_B(t) = \frac{C_1}{C_2} e^{-\frac{t}{2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}}, \tag{C.99}$$

where $\frac{1}{2c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}}}$ should decrease. Therefore, the distributions become flatter, i.e., the selling distribution increases slower with high cost while the buying distribution decreases slower with high value. The market has more 'competetive' population waiting.

As for the bounds,

$$\begin{aligned}\overline{V}_S - \underline{V}_B &= \left(-4c\frac{a}{1-a}\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} + a(U-L)\right)/(2-a) \\ &= \frac{a}{(1-a)(2-a)}\left(-4c\frac{e^{-\beta\delta}}{1-e^{-\beta\delta}} + (1-a)(U-L)\right),\end{aligned} \tag{C.100}$$

which is easy to see to be decreasing. At the same time, by overall flow balance, in the symmetric case $\overline{V}_S - L = U - \underline{V}_B$, or $\underline{V}_B + \overline{V}_S = U + L$, thus

$$\overline{V}_S = (\overline{V}_S - \underline{V}_B + \overline{V}_S + \underline{V}_B)/2 \text{ decreases}, \quad \underline{V}_S = (\underline{V}_B + \overline{V}_S - (\overline{V}_S - \underline{V}_B))/2 \text{ increases}. \tag{C.101}$$

So the bounds become narrower.

As for expected delay, now that the bounds becomes closer and the distributions become flatter, the overlapping agents should expect higher probability to trade as now they are closer to the bounds where probability to trade is 1 and the probability to trade decays slower at the same time.

(iv) When $e^{-\beta\delta}$ increases, at each value of $a \in (0,1)$, $E_L(a)$ increases and $E_R(a)$ decreases. As such, at the original $a$, $E(a) = E_L(a) - E_R(a) > 0$ now. Since $E(a)$ is monotonically increasing, $a$ should decrease to make $E(a) = 0$ again.

■