Computational Models of Algorithmic Trading in Financial Markets

by

Elaine Wah

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in The University of Michigan
2016

Doctoral Committee:

Professor Michael P. Wellman, Chair
Assistant Professor Jacob Abernethy
Professor Michael S. Barr
Professor Uday Rajan

We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.

— T. S. Eliot, *Four Quartets*

To my parents and my twin sister

ACKNOWLEDGMENTS

First and foremost, this dissertation would not have been possible without the guidance and support of my advisor, Michael Wellman. As a collaborator and advisor he has pushed me to hold my work to the highest standards via honest, constructive feedback; as a researcher and role model he has consistently exemplified integrity; and as a mentor he has been unfailingly generous with his time and advice. I would be remiss if I did not also acknowledge his profound impact on my personal and professional development over the course of my graduate career. It has been a singular privilege to work with Mike, and I am immensely grateful for his patience, encouragement, and generosity.

I am deeply indebted to my dissertation committee of Jake Abernethy, Michael Barr, and Uday Rajan, who have offered much insight and guidance throughout this process. Their feedback has been invaluable, and this thesis is unquestionably the better for it. I am also grateful to the CSE administrative staff, in particular Dawn Freysinger, Kimberly Mann, and Rita Rendell, for their support and assistance throughout my time at Michigan.

I have had many incredible colleagues and collaborators outside the University of Michigan, most notably Sébastien Lahaie and Dave Pennock at Microsoft Research New York City, and Amy Edwards and Austin Gerig at the U.S. Securities and Exchange Commission. Through these individuals I have worked on a variety of diverse, stimulating projects, and I am grateful for their advice and mentorship.

I have also been very fortunate to have had the opportunity to work with (and learn from) a peerless group of labmates. Of these past and present members of the Strategic Reasoning Group, I am particularly grateful to Erik Brinkman and Travis Martin for

their continued friendship through endeavors both academic and extracurricular, and to Ben Cassell, Quang Duong, Bartley Tablante, Bryce Wiedenbeck, and Mason Wright for selflessly offering me much help and advice. I have also greatly appreciated working with many bright, motivated, and enthusiastic undergraduate students while at Michigan, most notably Dylan Hurd and Zhiyi (Scarlett) Zhang.

Much gratitude goes to Hannah Imlay and Theresa Kim for their generosity, kindness, and friendship over the years. I would also like to thank the students and alumni who have made my time at Michigan so very memorable: Nilmini Abeyratne, Aarthi Balachander, Rob Goeddel, Lauren Hinkle, James Kirk, Zach Musgrave, Sanae Rosen, Gaurav Singhal, and the ladies of ECSEL, among many others.

Lastly, I owe a significant debt of gratitude to my parents Benjamin and Christine and my twin sister Catherine; their unwavering support has been invaluable throughout my graduate studies. As my wombmate and best friend, my sister has been there for me since day one, and her compassion and thoughtfulness know no bounds. Navigating the peaks and valleys of the doctoral journey with her by my side has been truly been a gift. From my father, I learned the fundamentals of good research, as well as the importance of perseverance, diligence, and a strong work ethic. I am deeply appreciative of his steadfast faith and confidence in my abilities. Our fruitful collaboration together, albeit brief, remains my fondest memory of any research undertaking. Finally, I am profoundly grateful for the tireless support and encouragement of my mother, without whose sacrifices I would not be where I am today. From her I have learned to face all obstacles with good humor and optimism, and she continually inspires me with her wisdom, grace, and strength.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Computational Models of Algorithmic Trading in Financial Markets

by

Elaine Wah

Chair: Michael P. Wellman

Today's trading landscape is a fragmented and complex system of interconnected electronic markets in which algorithmic traders are responsible for the majority of trading activity. Questions about the effects of algorithmic trading naturally lend themselves to a computational approach, given the nature of the algorithms involved and the electronic systems in place for processing and matching orders. To better understand the economic implications of algorithmic trading, I construct computational agent-based models of scenarios with investors interacting with various algorithmic traders. I employ the simulation-based methodology of empirical game-theoretic analysis to characterize trader behavior in equilibrium under different market conditions.

I evaluate the impact of algorithmic trading and market structure within three different scenarios. First, I examine the impact of a market maker on trading gains in a variety of environments. A market maker facilitates trade and supplies liquidity by simultaneously maintaining offers to buy and sell. I find that market making strongly tends to increase total welfare and the market maker is itself profitable. Market making may or may not benefit investors, however, depending on market thickness, investor impatience, and the number of trading opportunities. Second, I investigate the interplay between market frag-

mentation and latency arbitrage, a type of algorithmic trading strategy in which traders exercise superior speed in order to exploit price disparities between exchanges. I show that the presence of a latency arbitrageur degrades allocative efficiency in continuous markets. Periodic clearing at regular intervals, as in a frequent call market, not only eliminates the opportunity for latency arbitrage but also significantly improves welfare. Lastly, I study whether frequent call markets could potentially coexist alongside the continuous trading mechanisms employed by virtually all modern exchanges. I examine the strategic behavior of fast and slow traders who submit orders to either a frequent call market or a continuous double auction. I model this as a game of market choice, and I find strong evidence of a predator-prey relationship between fast and slow traders: the fast traders prefer to be with slower agents regardless of market, and slow traders ultimately seek the protection of the frequent call market.

CHAPTER I

Introduction

The predominantly electronic infrastructure of the U.S. stock market has come under intense scrutiny in recent years, during which several major technology-related disruptions have roiled the markets. In August 2013, for example, an overflow of market quotes caused a three-hour halt in trading at Nasdaq (De La Merced, 2013) and, in a separate incident, Goldman Sachs unintentionally flooded U.S. exchanges with a large number of erroneous stock-option orders (Gammeltoft and Griffin, 2013). Nasdaq's computer systems were similarly overwhelmed during the Facebook IPO on May 18, 2012, when a surge in order cancellations and updates delayed the opening of the shares for trading (Mehta, 2012). These events are reminiscent of the tumultuous trading activity caused by technological problems at Knight Capital in August 2012 (Popper, 2012) and the so-called "Flash Crash" of May 6, 2010, during which the Dow Jones Industrial Average exhibited its largest single-day decline (approximately 1,000 points) (Bowley, 2010).

These episodes of market turbulence are symptomatic of today's trading landscape, a fragmented and complex system of interconnected electronic markets that compete with each other for order flow. There are over 40 trading venues for stocks in the U.S. alone (O'Hara and Ye, 2011). The majority of activity on these markets comes from *algorithmic trading*, which employs computational and mathematical tools to automate the process of making trading decisions in financial markets. Algorithmic trading has been the subject

of much discussion and research, particularly regarding its benefits and drawbacks (Government Office for Science, London, 2012). Controversies about algorithmic trading in today's financial markets reached a critical point with the publication of *Flash Boys* by Michael Lewis, which discusses the ways in which many algorithmic traders strive for speed advantages in the pursuit of profit. *Flash Boys* tells the story of IEX, a trading venue designed specifically in response to such activity (Lewis, 2014).

Trading practices that exploit latency advantages in market access and execution in order to enhance profits are collectively called *high-frequency trading* (HFT), and are estimated to account for over half of daily trading volume (Cardella et al., 2014). *Latency* refers to the time needed to receive, process, and act upon new information. There is no formal regulatory definition of HFT, and the term itself encompasses a broad array of strategies—including but not limited to *latency arbitrage*, in which HFTs use their speed advantages to exploit price disparities in the same or correlated securities. General attributes of HFT include high daily trading volume, extremely short holding periods (on the order of microseconds), and liquidation rather than carrying significant open positions overnight (Wheatley, 2010). Proponents of high-speed traders posit that HFT activity reduces trading costs for market participants. Others argue that these traders harm investors and that practices to reduce latency contribute to a wasteful *latency arms race*, in which HFTs compete to access and respond to information faster than their competitors (Goldstein et al., 2014).

High-frequency traders gain latency advantages through various means. One method is co-location, in which HFT firms pay a premium to place their computers in the same data center that houses an exchange's servers. Many HFT firms also pay for direct data feeds in order to receive market data and market-moving information faster than non-HF investors. However, firms may spend millions of dollars to build a new, faster communication line only to be made obsolete by technology improvements that shave off additional milliseconds. One example of this rapid antiquation is Spread Networks' fiber optic cable, which was deprecated less than two years after its completion by the introduction of a network

reliant on microwave beams through air (Adler, 2012). According to estimates by the Tabb Group, firms spent approximately $1.5 billion in 2013 on technology to reduce latency (Patterson, 2014).

The ubiquity of algorithmic trading and the perpetuation of the latency arms race have been facilitated in no small part by the complex and fragmented nature of current markets. Clearly, a more comprehensive understanding of the dynamics between algorithmic trading and market structure—as well as their effects on market participants—is essential for ensuring the efficiency and integrity of U.S. financial markets.

Previous work on the effects of algorithmic trading and market structure has relied primarily on either analytical models or examination of historical order and transaction data. Historical market data alone is insufficient as it cannot be used to answer counterfactual questions about the impact of modifying strategies or market rules. Analytical models, on the other hand, can capture essential aspects of market structure, but would require stifling complexity to specify the interactions between multiple entities or the precise timing of event occurrences (such as the propagation of information between markets and participants)—at which point a closed-form solution or any other reasoning would be rendered infeasible or otherwise unhelpful.

Questions about the interplay between algorithmic trading and market structure naturally lend themselves to a computational approach. Indeed, these questions are inherently computational due to the very nature of the trading algorithms involved and the electronic systems in place for processing and matching orders. Algorithmic traders are prime examples of *autonomous agents*: they are highly responsive to changes in their environment, they often learn from historical performance, and they direct their activity towards achieving an objective. In the case of algorithmic traders, their goal is to maximize profit. The study of the economic implications of these intelligent systems is therefore extremely well-suited for the methodologies and approaches honed within the field of computer science: paradigms such as agent-based modeling and simulation provide the ability to specify agent objectives

individually rather than in aggregate, and game theory offers a framework for evaluating strategic behavior in a multiagent environment.

To better understand the impact of algorithmic trading and market structure on market outcomes, I construct computational agent-based models of market scenarios in which investors interact with various forms of algorithmic traders, and I characterize trader behavior in equilibrium under different market conditions. What follows is an overview of this dissertation.

In Chapter II, I present the class of financial market models I use. I focus on two types of markets: the *continuous double auction*, in which orders are matched as they arrive, and the *frequent call market*, in which orders are matched at periodic, fixed intervals. I also present the background traders (representing investors in the market) who populate my models, and I define the market performance characteristics I measure.

In Chapter III, I describe the simulation-based approach I employ to study the interactions of competing trading algorithms in different market environments. I employ *discrete-event simulation*, a paradigm that facilitates the exploration of interactions between traders by treating each change in system state at a given time as an event, with all events maintained in a queue ordered by time of occurrence. I also present the methodology of *empirical game-theoretic analysis*, which I use to compute equilibria in various market scenarios.

In Chapter IV, I present a study on equilibrium outcomes given the presence of a *market maker*, or MM. A market maker facilitates trade by simultaneously maintaining offers to buy and sell, and it is generally considered to perform a valuable function in continuous markets. However, I find that the impact of market making on welfare depends on the market environment. In this work, I compare settings both with and without MM in a variety of market environments. I model a single security traded in a continuous double auction populated by multiple background traders, and I characterize the strategic play in equilibrium. I find that presence of the market maker strongly tends to increase total welfare across a variety of environments. Market making may or may not be beneficial to

background investors, depending on market thickness, investor impatience, and the number of trading opportunities.

In Chapter V, I investigate the interplay between market fragmentation and latency arbitrage, a certain type of HFT strategy in which traders exercise superior speed to exploit price disparities between exchanges. I present a two-market model that captures market fragmentation, current U.S. securities regulations, and market clearing rules. In my model, latency arbitrage opportunities arise due to order routing based on outdated information as reflected in a global price quote that is updated with some delay. My results show that the presence of a latency arbitrageur can significantly degrade overall gains from trade. Switching to a centralized frequent call market not only eliminates the opportunity for latency arbitrage but also significantly increases welfare. Fragmentation can provide some benefit to welfare, but this effect depends on factors related to market conditions, such as the number of trading opportunities.

In Chapter VI, I investigate the potential for widespread adoption of frequent call markets as a market design solution to the latency arms race. That is, will traders prefer to submit limit orders to a frequent call market over a continuous double auction market, and if so, under what conditions? I model this question as a game of strategic market choice with fast and slow traders. Traders select a market type (continuous or discrete, as in the frequent call market) as part of their strategy. I find strong evidence of a predator-prey relationship between fast and slow traders: the fast traders prefer to be with slower agents regardless of market, and slow traders ultimately seek the protection of the frequent call market. My results demonstrate that frequent call markets are potentially a viable alternative to the continuous markets that dominate today's trading landscape.

By characterizing trader and market performance in these three case studies, my dissertation offers a closer look at the interplay between algorithmic trading and market structure in different scenarios. This thesis also presents a framework for modeling and analyzing algorithmic trading in financial markets: I construct agent-based models of markets popu-

lated by various market participants, and I employ a computational methodology coupling simulation with game-theoretic analysis in order to compare outcomes in equilibrium.

CHAPTER II

Financial Market Models

My thesis explores the impact of market structure modifications on traders and the markets themselves, with the goal of analyzing the impact of algorithmic trading and designing market rules that mitigate any detrimental effects of such trading. Comparing models of various market configurations facilitates the identification of weaknesses in current markets—which may be regulatory or more fundamentally structural—and the study of how certain advantaged traders may exploit these vulnerabilities to extract profits from other market participants.

I focus on two types of markets in this thesis. The *continuous double auction* (CDA), in which orders are matched as they arrive, is used in virtually all stock markets today. This is in contrast to a periodic or *frequent call market*, in which orders are matched to trade at regular, fixed intervals (on the order of tenths of a second). I describe these two types of markets in Section 2.1.

My market models are populated by *background traders*, who represent investors in the market. This is in contrast to market participants who exclusively pursue trading profit (Chapters IV and V). I describe the valuation model of background traders in Section 2.2, and I discuss the class of background-trader strategies in Section 2.3.

As with any simulation model, my results are valid only to the extent my assumptions capture the essence of real-world markets. My financial market models generally rely on

simple characterizations of trader behavior, and they consider a limited range of regulatory mechanisms and responses, as I focus much of this thesis on the relationship between algorithmic trading and market clearing rules. While some modeling choices (e.g., those specifying the trader valuation model or agent arrivals into a market) are somewhat arbitrary, they are largely based on prior studies in the literature. In general, I seek to ensure my models capture the structural details and trader behaviors of interest, without adding unnecessary complexity.

2.1 Market Clearing Mechanisms

The continuous double auction is a simple and standard two-sided market that forms the basis for most financial and commodities markets (Friedman, 1993). Agents submit bids, or *limit orders*, specifying the maximum price at which they would be willing to buy a unit of the security, or the minimum price at which they would be willing to sell (hence, the CDA is often referred to as a limit-order market in the finance literature). CDAs are continuous in the sense that when a new order matches an existing incumbent order in the order book, the market clears immediately and the trade is executed at the price of the incumbent order—which is then removed from the book. Orders may be submitted at any time, and a buy order matches and transacts with a sell order when the limits of both parties can be mutually satisfied.

An alternative to continuous trading is a *frequent call market* or frequent batch auction, in which order matching is performed only at discrete, periodic intervals (e.g., on the order of tenths of a second). A discrete-time market facilitates more efficient trading by aggregating supply and demand and matching orders to trade at a uniform price (Biais et al., 2005; Gode and Sunder, 1997; Wah and Wellman, 2013). As in the CDA, traders in the frequent call market can arrive and submit orders at any time. The submitted limit orders remain in the order book until executed or canceled. In a frequent call market, orders are accumulated over a series of fixed-length clearing intervals. Orders are processed in batch

via a uniform-price auction: at the end of each interval, the market computes the aggregate supply and demand functions based on current outstanding orders. No trade occurs if supply and demand do not intersect. If supply and demand intersect, the market clears at a uniform price that best matches the aggregated buy and sell orders, i.e., where supply equals demand. Buy orders strictly greater than the computed price, as well as sell orders strictly less than this price, will execute and subsequently be removed from the order book. If supply and demand intersect horizontally or at a single point, there exists a unique clearing price for the given interval. Orders that do not trade in the current period will remain outstanding and carry over to the next clearing interval.

A frequent call market effectively eliminates the latency advantages of HFTs by hiding all submitted orders within each clearing interval, as in a sealed-bid auction. The removal of time priority within each batch period helps ensure that standing offers cannot be readily picked off by incoming orders, thereby transforming the competition on speed into a competition on price. This ensures that there is no significant advantage to receiving and responding to information faster than other traders, because all orders within a clearing interval are processed and matched at the same time. Periodic clears every second or so would be imperceptible to most investors but would prevent the exploitation of small speed advantages, thus curbing HFT participation in the latency arms race.

In my implementation of these market models, prices are fine-grained but discrete, taking values at integer multiples of the *tick size* $p_{ts} = 1$. Agents arrive at designated times, and submit limit orders to their associated market(s). Each market continually publishes a price quote consisting of two parts, the BID and the ASK. Other bids in the order book are not visible to traders. CDA price quotes reflect the best current outstanding orders, while the frequent call market quotes reflect the best outstanding orders immediately following the most recent market clear. Specifically, for the CDA, BID_t is the price of the highest buy offer at time t and ASK_t is the price of the lowest offer to sell. For the frequent call market, BID_t corresponds to the highest outstanding buy offer after the clear at the most

recent clear time c, such that $BID_{t_1} = BID_{t_2}$ for any $c \leq t_1 < t_2 \leq t$. Similarly, ASK_t is the lowest outstanding offer to sell, such that $ASK_{t_1} = ASK_{t_2}$ for any $c \leq t_1 < t_2 \leq t$. The difference between the two quote components is called the BID-ASK *spread*. An invariant for both the CDA and the call market is that $BID < ASK$. Otherwise, the orders would have matched and been removed from the order book—either immediately in the case of the CDA or upon the clear in the frequent call market.

2.2 Valuation Model

Each background trader possesses a individual valuation for the security in question, which is comprised of private and common components. The common component is defined as follows. I denote by r_t the common *fundamental value* for the security at time t. The fundamental time series is generated by a mean-reverting stochastic process:

$$r_t = \max\left\{0,\ \kappa\bar{r} + (1-\kappa)\, r_{t-1} + u_t\right\}.$$

Parameter $\kappa \in [0, 1]$ specifies the degree to which the fundamental reverts back to the mean $\bar{r}$, and parameter $u_t \sim \mathcal{N}\left(0, \sigma_s^2\right)$ is a random shock at time t.

The private component for agent i is a vector Θ_i representing differences in the agent's private benefits of trading given its net position, similar to the model of Goettler et al. (2009). This private valuation vector reflects individual preferences in the marginal value of the security (e.g., due to risk aversion, outside portfolio holdings of related securities, or immediate liquidity needs), as well as preferences regarding urgency to trade. The vector is of size $2q_{\max}$, where $q_{\max} = 10$ is the maximum number of units the agent can be long or short at any time, with

$$\Theta_i = \left(\theta_i^{-q_{\max}+1}, \ldots, \theta_i^0, \theta_i^{+1}, \ldots, \theta_i^{q_{\max}}\right).$$

Element θ_i^q is the incremental private benefit obtained from selling one unit of the security given current position q, where positive (negative) q indicates a long (short) position. Similarly, θ_i^{q+1} is the marginal private gain from buying an additional unit given current net position q.

I generate Θ_i from a set of $2q_{\max}$ values drawn independently from a Gaussian distribution. Let $\hat{\theta} \sim \mathcal{N}(0, \sigma_{PV}^2)$ denote one of these drawn values. To ensure that the valuation reflects diminishing marginal utility, that is, $\theta^{q'} \geq \theta^q$ for all $q' \leq q$, I sort the $\hat{\theta}$ and set the θ_i^q to respective values in the sorted list.

Background trader i's valuation v for the security at time t is based on its current position q_t and the value of the global fundamental at time T, the end of the trading horizon:

$$v_i(t) = r_T + \begin{cases} \theta_i^{q_t+1} & \text{if buying 1 unit} \\ \theta_i^{q_t} & \text{if selling 1 unit.} \end{cases}$$

For a single-quantity limit order transacting at time t and price p, a trader obtains surplus:

$$\begin{cases} v_i(t) - p & \text{for buy transactions, or} \\ p - v_i(t) & \text{for sell transactions.} \end{cases}$$

Since the price and fundamental terms cancel out in exchange, the total surplus achieved when agent B buys from agent S is $\theta_B^{q(B)+1} - \theta_S^{q(S)}$, where $q(i)$ denotes the pre-trade position of agent i.

2.3 Background-Trader Strategies

There is an extensive literature on autonomous bidding strategies for CDAs (Das et al., 2001; Friedman, 1993; Wellman, 2011). In this thesis, I consider trading strategies in the so-called *Zero Intelligence* (ZI) family (Gode and Sunder, 1993).

The background traders arrive at the market according to a Poisson process with rate λ_{BG}. On each arrival, they are assigned to buy or sell (with equal probability), and accordingly submit an order to buy or sell a single unit. (A trader is randomly reassigned to buy or to sell each time it arrives.) Background traders subsequently reenter the market, with time between entries distributed exponentially at the same rate λ_{BG}—in other words, each trader reenters the market according to an independent Poisson process, just like its arrival process. Background traders are notified of all transactions and current price quotes with zero delay, and may use this information in computing their bids. Agents may trade any number of times, as long as their net positions do not exceed $q_{\max}$ (either long or short).

Recall that each background trader has an individual valuation for the security comprised of private and common components, as described in the previous section. Based on this valuation, each background trader obtains a payoff at the end of the simulation period. This payoff is computed as the sum of the private value of the trader's holdings, the net cash flow from trading, and the liquidation proceeds of any accumulated inventory at the end-time fundamental value r_T (i.e., the common component of the valuation).

A ZI trader assesses its valuation $v_i(t)$ at the time of market entry t, using an estimate $\hat{r}_t$ of the terminal fundamental r_T. The estimate is based on the current fundamental, r_t, adjusted to account for mean reversion:

$$\hat{r}_t = \left(1 - (1-\kappa)^{T-t}\right)\bar{r} + (1-\kappa)^{T-t} r_t. \tag{2.1}$$

The ZI agent then submits a bid shaded from this estimate by a random offset—the degree of surplus it demands from the trade. The amount of shading is drawn uniformly from range $[R_{\min}, R_{\max}]$. Specifically, a ZI trader i arriving at time t with current position q

submits a limit order for a single unit of the security at price

$$p_i \sim \begin{cases} \mathcal{U}\left[\hat{r}_t + \theta_i^{q+1} - R_{\max}, \hat{r}_t + \theta_i^{q+1} - R_{\min}\right] & \text{if buying} \\ \mathcal{U}\left[\hat{r}_t + \theta_i^{q} + R_{\min}, \hat{r}_t + \theta_i^{q} + R_{\max}\right] & \text{if selling.} \end{cases}$$

I extend ZI by including a threshold parameter $\eta \in [0, 1]$, whereby if the agent could achieve a fraction η of its requested surplus at the current price quote, it would simply take that quote rather than posting a limit order to the book. Setting $\eta = 1$ is equivalent to the strategy without employing the threshold.

In my setting, background traders are permitted to reenter the market. Upon each entry, the trader withdraws its previous order (if not transacted yet) before executing the extended ZI strategy described above.

2.4 Performance Measures

In exploring the relationship between trader behavior and market structure, I am interested in the following performance characteristics:

Allocative efficiency Total surplus (welfare) is my key measure of market performance. Welfare indicates how well the market allocates trades according to underlying private valuations.

Liquidity Markets are liquid to the extent they maintain availability of opportunities to trade at prevailing prices. Two liquidity metrics are fast execution and tight BID-ASK spreads. I measure *execution time* by the difference in time between order submission and transaction for orders that eventually trade. Execution time is potentially important to investors for many reasons, including the risk of changes in valuation while an order is pending, the effect of transaction delay on other contingent decisions, and general time

preference. I also measure spread, which is the distance between prices quoted to buyers and sellers (Section 2.1).

Price discovery This reflects how well prices incorporate information. I measure price discovery using the root mean square deviation (RMSD) between the midquote price (midpoint of the BID-ASK spread) and fundamental value at every time step.

CHAPTER III

Computational Approach

To answer questions regarding the interplay between trader behavior and market structure, I employ a computational approach that comprises agent-based modeling, simulation, and equilibrium computation. Using real-world financial data is infeasible and inadequate for this type of analysis: not only is data containing the necessary level of detail at the requisite time scale simply not available, but historical data also does not permit the measurement of trading gains (as trader valuations are unknown) or any evaluation of the effects of modifying market rules.

I employ *agent-based modeling*, or ABM, to represent the interactions between traders in one or more markets. In ABM, autonomous agents interact dynamically based on algorithmic rules specified ahead of time. These rules govern each agent's actions and responses, but do not explicitly define or specify the interactions between traders; instead, emergent phenomena can be observed from collective agent behavior. I simulate interactions between agents in a variety of market environments to study the effect of market structure and trader strategies on market performance. Simulation modeling enables me to incorporate causal premises, specifically presumptions of how trading behavior is shaped by environmental conditions. I present my simulation system in Section 3.1.

Using trader performance assessed from simulation runs, I employ game-theoretic analysis to evaluate traders' strategic interactions with each other under a variety of market

settings. I focus on trader behavior in equilibrium, when all market participants are best responding to each other's strategies in order to optimize their own gains from trade. Equilibrium outcomes offer a basis for predicting an agent's actions when it is faced with strategic decisions. I explore various market scenarios and environments in order to characterize trader behavior in equilibrium under different market conditions. I describe the methodology of empirical game-theoretic analysis that I employ to compute equilibria in Section 3.2.

In order to mitigate the stochasticity in my simulations and reduce sampling error, I collect large numbers of observations for each environment setting and trader population of interest. I utilize the EGTAOnline infrastructure (Cassell and Wellman, 2013) to conduct and manage my experiments, and I run my simulations on the high-performance computing cluster at the University of Michigan.

3.1 Discrete-Event Simulation

The financial markets I study are stochastic, dynamic systems with discrete states that change in response to communication events. These events occur at high frequency, often on the order of microseconds. To faithfully model such systems in simulation, ensuring the unambiguous timing of agent and market interactions is paramount. This necessitates fine-grained modeling at the level of communication.

I therefore design my system based on principles of *discrete-event simulation* (DES), which affords the precise specification of temporal changes in system state. In the DES framework, a simulation run is modeled as a sequence of events (Banks et al., 2005). Each event is an instantaneous occurrence that marks a change to the system state at a given time, and events are maintained in a queue ordered by time of occurrence.

My DES system simulates the interactions among traders in a set of markets. An *event* in my system consists of a sequence of *activities* that are to be executed by various *entities* (e.g., traders, markets, and information processors). The events are ordered in a priority queue by event time and executed sequentially until the event queue is empty. Multiple

events may be scheduled for the same time step, in which case they are executed deterministically in the order in which they are enqueued. Each event's list of activities is sequenced by priority; activities with matching priorities are inserted in the order they arrive. Priorities are assigned based on activity type (e.g., bid submission, market clearing). This guarantees determinism in the sequential execution of activities and the correct operation of markets.

To illustrate event sequencing in my simulation system, I present an example of quote aggregation across two markets. U.S. securities regulations require exchanges to report their best buy and sell orders to an entity called the Security Information Processor, or SIP (Blume, 2007; Securities and Exchange Commission, 2005). Given order information from exchanges, the SIP continually publishes a public price quote called the "National Best Bid and Offer" (NBBO). This process of computing and disseminating the NBBO takes some finite time, say δ time steps.

To control the latency of the SIP within my simulation system, I specify three activities: `SendToSIP`, `ProcessQuote`, and `UpdateNBBO`. The `SendToSIP` activity is inserted when a market publishes a quote at time t; upon execution of this activity, the market sends its updated quote to the SIP entity and inserts a `ProcessQuote` and an `UpdateNBBO` activity, both to execute at time $t + \delta$. When `ProcessQuote` is executed, the SIP updates its information on the best quotes in the markets. It then computes and publishes an updated NBBO based on this information during the execution of the `UpdateNBBO` activity.

In this way, activities in my simulation system are sequenced to reflect the communication latencies arising as a consequence of market fragmentation. In Figure 3.1, market 1 clears and publishes an updated quote at time t_1. Market 2 publishes its new quote at time t_2. For $\delta > t_2 - t_1$, a `ProcessQuote` followed by an `UpdateNBBO` activity is executed sequentially at $t_1 + \delta$, as well as at time $t_2 + \delta$. The `UpdateNBBO` executing at $t_1 + \delta$ does not incorporate market 2's updated quote, as the `ProcessQuote` activity to add market 2's best quote is not executed until $t_2 + \delta$. This process serves to model the

behavior of the SIP with a delay of δ.

My financial market simulation system affords sufficient versatility to model a wide range of market environments, including variform populations of market participants, as well as different market structures (e.g., varying in the number of markets or types of market mechanisms employed). The simulator has been extended by other members of the Strategic Reasoning Group at the University of Michigan, and it is in current use in a number of other studies.

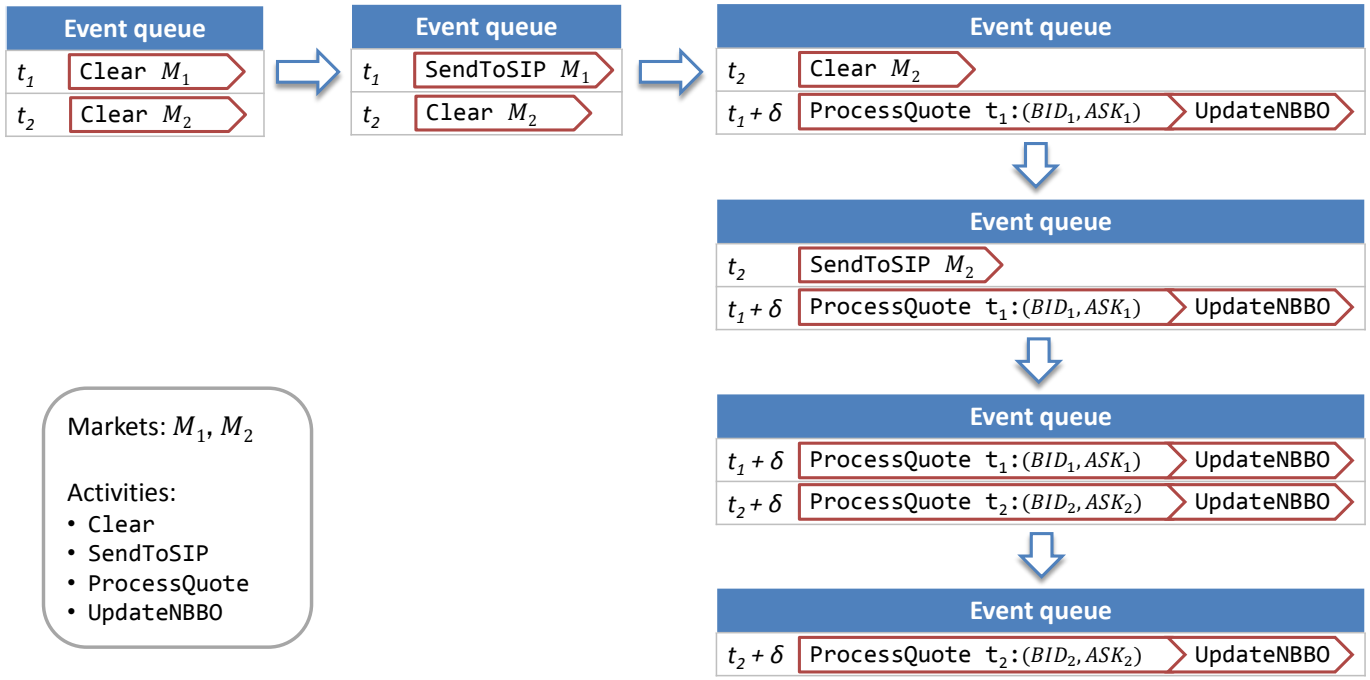


Figure 3.1: Discrete-event simulation system event queue during the dissemination and processing of updated market quotes for NBBO computation, given latency $\delta > t_2 - t_1$. There are two markets, M_1 and M_2. When the NBBO update activity executes at time $t_1 + \delta$, the SIP has just processed market 1's best quote (BID_1, ASK_1) at time t_1; this is therefore the most up-to-date information that could be reflected in the NBBO at time $t_1 + \delta$.

3.2 Empirical Game-Theoretic Analysis

Game theory offers a framework for analyzing strategic behavior within multiagent environments. A *game* is a model of interactions between two or more agents, or *players*, with outcomes dependent on the players' actions. Each player makes strategic decisions within the game, and a player's *strategy* specifies its actions. A *pure strategy* provides a complete, deterministic specification of the agent's actions. Players can also adopt a *mixed strategy*, which is a probability distribution over the strategy set (which is comprised of all pure strategies). A *strategy profile* specifies player-strategy assignments for all agents in the game. A *pure-strategy profile* consists of all players adopting pure strategies. In a *mixed-strategy profile*, at least one player adopts a mixed strategy. The *support* is defined as the set of strategies played with nonzero probability (hence, the size of the support of a pure-strategy profile is 1). The *outcome* of a game is defined as the set of strategies adopted by players, and a player's *payoff* is a number representing the desirability of an outcome.

The simulation system discussed in the previous section takes as input a strategy profile and generates as output a sample outcome, which depends on the combination of strategies played by traders. Traders can select from a set of strategies, where each strategy is a parameterization of a class of trading strategies, such as Zero Intelligence (Section 2.3). Traders may improve individual payoffs (i.e., surplus) by adopting certain strategies over others, but performance may vary depending on the environment and on what the other traders are doing.

Empirical-game theoretic analysis (EGTA) is a simulation-based process that facilitates strategy selection for agents by comparing the payoffs of different combinations of player and strategy assignments. Developed by Wellman (2006) and the Strategic Reasoning Group at the University of Michigan, EGTA entails systematic simulation of many strategy profiles, accumulating payoff observations, and inducing an empirical game model. To guide strategy choice, I apply the notion of *game-theoretic equilibrium*, which provides a rule for predicting agent behavior. More specifically, I focus on *Nash equilibrium*, in which

each player selects the strategy that maximizes its payoff, given the strategies of the other players. Given this equilibrium concept, solving the induced empirical game then gives optimal strategy assignments for the players.

In this thesis, I model a financial market as a *role-symmetric game*, in which players are partitioned into multiple roles, each with a specified strategy set. Payoffs in such a game are completely determined by role membership and by the number playing each strategy in each role. A *symmetric game*, in which payoffs depend only on the number of agents playing each strategy, can be modeled as a role-symmetric game with a single role. A *role-symmetric Nash equilibrium* (RSNE) assigns a strategy profile to every role, such that all agents in a role have the same mixed strategy, and no agent can benefit in expectation by unilaterally switching to a different strategy.

3.2.1 EGTA Process

The goal in this process is to identify Nash equilibria, and I focus my search on role-symmetric Nash equilibria, in pure or mixed strategies. To analyze a game, I apply EGTA in an iterative manner, interleaving exploration of the profile space with analysis of the empirical game model induced by average payoffs in simulation. I start by simulating all the role-symmetric pure-strategy profiles, where a single strategy is shared by all players in a role. Exploration then spreads through their neighbors, that is, those profiles related by single-agent deviations.

Observed payoffs from simulation runs of a given profile are added incrementally to the empirical game's payoff matrix. For this reason, the game is incomplete at any point during the EGTA process, as some profiles have been empirically evaluated whereas others have not. Each update to the empirical game's payoff matrix generates an intermediate game model. As payoffs from simulation are incorporated into the empirical game, I analyze each successive intermediate game model by computing (mixed) equilibria for each *complete subgame*. A *subgame* is the game obtained by restricting the set of strategies for each

role, and a complete subgame is defined as a subgame for which all profiles have been evaluated by simulation. The role-symmetric Nash equilibria of the complete subgames are *candidates* for equilibria in the full game. If I can identify a strategy in the full strategy set that beneficially deviates from the candidate, I say the candidate is *refuted*. A candidate profile is *confirmed* as an RSNE when all possible deviations have been evaluated, and none are beneficial. I confirm or refute each candidate by evaluating deviations to strategies outside their subgames. If a candidate is refuted, I construct a new subgame by adding the best response to its support, and proceed to explore the corresponding subgame.

I simulate additional profiles for a game until I have confirmed at least one RSNE, evaluated every pure-strategy symmetric profile (i.e., where the players in each role play a strategy with probability 1), and pursued with some degree of diligence every equilibrium candidate encountered. More specifically, I continue to refine the empirical game with additional simulations until the following conditions are met:

1. at least one equilibrium is confirmed,
2. all non-confirmed candidates are refuted (up to a threshold support size), and
3. for all refuted candidates (up to the threshold support size), we have explored subgames formed by adding the best response to the candidate's support.

When this process reaches quiescence, I consider the search to have satisfied the diligence requirement.

The procedure described above seeks to either confirm or refute the equilibrium candidates detected in my exploration of the strategy space. As I am not able to exhaustively search the entire profile space, however, additional qualitatively distinct equilibria are always possible. In addition, the equilibria I find are subject to refutation by other strategies outside the specified set. My search process described above attempts to evaluate all promising equilibrium candidates (e.g., by exploring subgames extending the support of a refuted candidate with the best response), but identifying these is not guaranteed.

3.2.2 Game Reduction

Even with a moderate number of players, the *game size* (number of possible strategy profiles) grows exponentially with the number of players and strategies, rendering analysis of the full game computationally infeasible. As such, I apply aggregation to approximate the many-player games as games with fewer players: I employ the technique of *deviation-preserving reduction* (DPR) developed by Wiedenbeck and Wellman (2012) to construct a reduced-game approximation of the full game.

DPR preserves the payoffs from single-player, unilateral deviations, and maintains in the reduced game the same proportion of opponents playing each strategy as in the full game. In a deviation-preserving reduced game, each player views itself as controlling one full-game agent and views the other-agent profile in the reduced game as an aggregation of all other players in the full game. Although the equilibrium approximations obtained via DPR are not guaranteed estimates, DPR has been shown to produce good approximations in other games (Wiedenbeck and Wellman, 2012).

DPR defines reduced-game payoffs in terms of payoffs in the full game as follows.[1] Consider first an N-player symmetric game, reduced to a k-player game, for $k < N$. The payoff for playing strategy s_1 in the reduced game, with other agents playing strategies $(s_2, \ldots, s_k)$, is given by the payoff of playing s_1 in the full N-player game when the other $N-1$ agents are evenly divided ($\frac{N-1}{k-1}$ each) among strategies $s_2, \ldots, s_k$.

Now consider an (N_A, N_B)-player role-symmetric game with two roles A and B, reduced to a (k_A, k_B)-player game for $k_A < N_A, k_B < N_B$. Given the other agents in role A play strategies $(a_2, \ldots, a_{k_A})$ and the agents in role B play strategies $(b_1, \ldots, b_{k_B})$, the payoff for an agent in role A playing strategy a_1 in the reduced game is given by the payoff of playing a_1 in the full (N_A, N_B)-player game when the other $N_A - 1$ traders in role A

[1] With the exception of one environment, in all the case studies presented in this thesis the number of players N in the full game and the number of reduced-game players k are selected to ensure that the DPR definitions result in integer numbers of players. See the original paper by Wiedenbeck and Wellman (2012) for the complete definition of the number of reduced-game players within each role when divisibility does not hold.

are divided evenly $(\frac{N_A-1}{k_A-1})$ among the strategies $a_2, \ldots, a_{k_A}$ and the other-role (i.e., role B) players are divided evenly $(\frac{N_B}{k_B})$ among their strategies $b_1, \ldots, b_{k_B}$. The payoff for a single agent in role B is analogous.

CHAPTER IV

Welfare Effects of Market Making in Continuous Double Auctions

In this chapter, I present a study investigating the effects of market making on market performance, focusing on allocative efficiency as well as gains from trade accrued by background traders. The results from this case study have been reported in other papers (Wah and Wellman, 2015; Wah et al., 2016). I employ the empirical simulation-based methods described in Chapter III to evaluate heuristic strategies for market makers as well as background investors in a variety of complex trading environments. I compare the surplus achieved by background traders in strategic equilibrium, with and without a market maker. My findings indicate that the presence of the market maker strongly tends to increase total welfare across a variety of environments. Market making may or may not be beneficial to background investors, depending on characteristics of the market environment. I find that the benefit tends to accrue in relatively thin markets, and situations where investors are impatient, due to limited trading opportunities. Comparison across environments reveals factors that influence the existence and magnitude of benefits provided by the market maker function.

4.1 Introduction

A *market maker* (MM) facilitates trade in a two-sided auction market by simultaneously maintaining offers to buy and sell. An ever-present MM supplies *liquidity* to the market. Liquidity refers to the availability of immediate trading opportunities at prices that reasonably reflect current market conditions. In compensation for liquidity provision, MMs profit from the *spread*, the difference between their buy and sell offers. MM activity is generally understood to stabilize prices and facilitate discovery of accurate prices in the market (Schwartz and Peng, 2013).

The exact role of market makers varies across market institutions. In a *pure dealer market*, multiple MMs competitively quote prices, and incoming market orders from investors trade at the best available MM price (Huang and Stoll, 1996). In a *pure limit-order market*, both investors and MMs submit orders with price limits, and whenever an incoming order matches an existing order, they trade at the incumbent order's limit price. This market mechanism is also called a *continuous double auction* (CDA), the name I use here. In a *specialist market*, there is a single MM designated to act as dealer, with an affirmative obligation to maintain fair and orderly markets (Saar, 2010). With the transition to electronic markets, pure limit-order markets are becoming predominant (Frey and Grammig, 2006; Glosten, 1994), thus this is the market mechanism I employ in my study.

Providing liquidity can generate profits from investors, but also runs the risk of *adverse selection*: when traders with newer or otherwise better information take advantage of the MM's standing offers. Much of the market making literature focuses on this tradeoff and its implications for MM strategies (Glosten and Milgrom, 1985; Kyle, 1985); other prior research has investigated the effects of MM on liquidity (e.g., as measured by price spreads) (Das and Magdon-Ismail, 2008) and price discovery (Leach and Madhavan, 1992). Although liquidity and price discovery are generally expected to be positive factors for market performance and therefore welfare, there has been a notable dearth of prior research modeling this directly. Of the existing work addressing welfare, the focus has been on the

need for affirmative MM obligation due to adverse selection (Bessembinder et al., 2011, 2015), the cost structure of market participation in supplying liquidity (Huang and Wang, 2010), and trading mechanisms to incentivize market making (Brusco and Jackson, 1999).

In this study, I investigate the effects of MM on market performance, focusing on allocative efficiency as well as gains from trade accrued by background investors. In the specific model I examine in this chapter, a single security is traded via CDA mechanism in a market environment comprising a single market maker and multiple background traders. Recall that my financial market models incorporate private and common valuation elements, with dynamic fundamental value and asymmetric information (Section 2.2). The background traders each possess an individual valuation for the security. They enter and reenter according to a stochastic arrival process, each time to offer to buy or sell a single unit of the security (Section 2.3). The single MM has no private value, and thus aims to profit by maintaining buy and sell offers with a positive price spread.

To compare outcomes both with and without market making, I search for strategy configurations where traders best-respond to the environment and other-agent behavior. As analytic game-theoretic solution of this rich dynamic model appears intractable, I employ empirical simulation-based methods (Section 3.2) to derive equilibria over a restricted strategy space. For background traders, I consider parameterized strategies based on Zero Intelligence agents (Section 2.3). For the MM, I consider heuristic strategies loosely based on that defined by Chakraborty and Kearns (2011). From extensive simulation over thousands of strategy profiles, I estimate game models for various instances of the target scenario.

Analysis of the empirical games provides strong support for overall welfare benefits of market making. I derive empirical equilibria with and without market making in 21 environments, finding that the mix of background-trader strategies in equilibrium varies depending on the presence and strategy choice of the MM. In all of my environments, the single market maker is profitable in equilibrium, and in all but three equilibrium comparisons, the presence of MM increases overall welfare (background-trader surplus combined

with MM profit).

Whether market making benefits background traders (i.e., increases welfare net of MM profits) is more ambiguous, however. A single market maker makes investors better off in the majority of environments tested, and tends to do so particularly in relatively thin markets. For impatient investors with relatively infrequent trading opportunities, the MM is more beneficial the shorter the trading horizon.

In the next section I explain by way of example the potential role of market makers in alleviating allocative inefficiencies. I describe relevant work in Section 4.3. Section 4.4 discusses the market environment, and Section 4.5 describes my MM strategies. I present my experiments and results in Sections 4.6 and 4.7, respectively. I conclude in Section 4.8.

4.2 Motivating Example

I illustrate the problem of allocative inefficiency in CDAs, and the influence of market makers, with the following simple example. Suppose a market with four background traders: two buyers and two sellers. The buyers have values b_1 and b_2, and seller values are s_1 and s_2, with $b_1 > s_1 > b_2 > s_2$. Let me further assume for this illustration that the traders submit orders at their valuations.

Suppose that the orders arrive at the market in the order shown in Figure 4.1. Then buyer 1 trades with seller 1, and buyer 2 with seller 2, achieving a total surplus of $(b_1 - s_1) + (b_2 - s_2)$. The socially optimal allocation, in contrast, would have buyer 1 trading with seller 2, for a total surplus of $b_1 - s_2$. The difference between the optimal and achieved surplus is $\Delta = s_1 - b_2 > 0$. This loss can be attributed to the vagaries of the sequencing of limit orders, combined with the greedy matching implemented by the CDA mechanism. I choose to depict in the figure a sequence that leads to a suboptimal allocation; however, this is not the only one. In fact, only one-third of the possible orderings of these bids (8 out of 24) would result in the optimal allocation, with the remaining two-thirds under-performing by Δ.

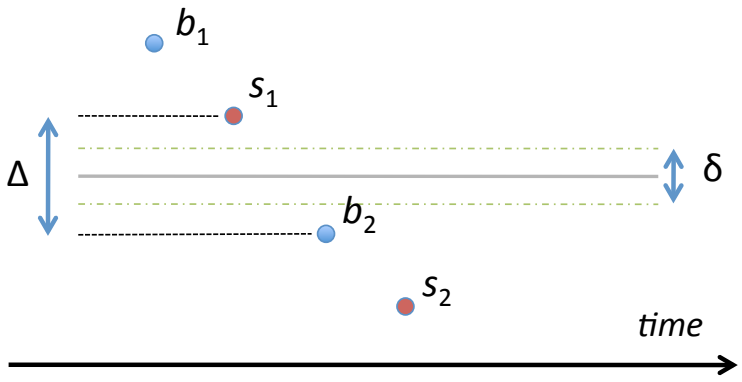


Figure 4.1: A sequence of CDA orders leading to a suboptimal allocation.

Now suppose there is a market maker who continually maintains buy and sell offers in the auction, with difference δ between them. As long as the MM's offer to buy is within the interval (s_2, s_1), and its offer to sell falls within (b_2, b_1), then for this sequence of order arrivals, buyer 1 and seller 2 will trade with the MM, and the allocation will be efficient. If the MM quotes lie within the narrower interval of *competitive equilibrium* prices[1] $[b_2, s_1]$, then the efficient allocation is achieved for *any* sequence. In such cases, the MM accrues a profit of δ, with the remaining surplus divided among background traders.

The MM promotes efficiency in this example by providing liquidity to the market. In the absence of MM, when buyer 1 arrives, it has nobody to trade with. Seller 1 fills the vacuum and makes a profitable trade with this buyer, but at a price quite removed from that which would match supply and demand aggregated over time. An MM with quotes approximating this long-run price, in contrast, allows arriving bidders to trade near prevailing prices. Equally important, it prevents bidders who should not trade based on their valuations from doing so.[2]

Even assuming that the MM improves overall efficiency, does it make the background traders better off? In the specific scenario of Figure 4.1, the background traders benefit (in aggregate) if $\delta < \Delta$. If instead I consider the same set of four bids, but submitted in

[1] A competitive equilibrium price balances supply and demand with price-taking bidders. Here the balance is with respect to cumulative orders over the time horizon.

[2] A modest amount of bid shading can also prevent inefficient trades, and indeed equilibrium shading strategies often lead to more efficient outcomes than truthful bidding in CDAs (Zhan and Friedman, 2007).

random order, then the background traders are clearly worse off in the third of instances where they would have achieved the efficient allocation without the MM's help. With random sequencing, the background traders benefit in expectation if and only if $\delta < \frac{2}{3}\Delta$.

More generally, it is clear that the question of whether MM presence is welfare-improving for background traders depends on specific details of the market setting. For background traders, the MM contribution may be sensitive to the distribution of valuations and bids, as well as their pattern of arrival over time. It also depends pivotally on the MM strategy—how well it tracks the prevailing market price and how large a spread the MM maintains between its buy and sell offers. In realistic environments, valuations include a combination of common and private elements and may evolve over time. Based on time and role, agents may have differential information about the common-value component. Thus for time-varying environments, I cannot assume the MM knows the underlying market equilibrium; it must instead act adaptively based on observations and statistical assumptions.

Moreover, individual traders may reenter the market to revise bids or reverse transactions, or to trade multiple units of the good. If such reentry were costless, market making would not be necessary to achieve allocative efficiency, as the traders could exchange among themselves to quiescence (Huang and Wang, 2010). As long as the traders do not indefinitely hold out for strict profits, the market would converge to an efficient allocation. In other words, *liquidity has economic value only to the extent that patience and market participation have costs or limits*.

With such complications, establishing general analytical conditions for the benefits of MM seems unlikely. Instead, I employ the empirical game-theoretic techniques presented in Section 3.2, which facilitate the search for strategically stable background-trader and MM strategies. My model includes all of the elements listed above, within an extensible framework, presented in Section 2.3, that could incorporate (in future work) additional relevant features of financial markets.

4.3 Related Work

Literature on market making lies predominantly within the field of *market microstructure*, which examines the process by which prices, information, and transactions are formed by detailed interactions of traders in a market mechanism (Biais et al., 2005; Madhavan, 2000; O'Hara, 1995). Early work focused on dealer markets, in which a monopolistic MM (the dealer) controls trading by acting as the middleman. Garman (1976) presents an explicit formulation of the market maker's optimization problem. O'Hara and Oldfield (1986) and Amihud and Mendelson (1980) concentrate on the impact of dealer inventory on spreads, while the seminal model of Glosten and Milgrom (1985) frames spread as arising from adverse selection. Others focus on the consequences of informed trading on MM (Chowhdry and Nanda, 1991; Das, 2008; Kyle, 1985), as well as the role of market makers as liquidity providers (Grossman and Miller, 1988; Seppi, 1997).

Much of the relevant theoretical literature, however, relies on simplifying assumptions of MM behavior and trader interactions (Biais et al., 2005). Empirical studies have provided insight on the effects of market makers in real-world markets (Frey and Grammig, 2006; Hasbrouck and Sofianos, 1993; Manaster and Mann, 1996; Menkveld, 2013; Sandås, 2001). Historical data alone, however, cannot elucidate the strategic choices faced by market participants. Agent-based modeling (ABM) and simulation of financial markets has proven conducive to exploring these questions (LeBaron, 2006); however, only a handful of ABM finance papers focus on market making (Chan and Shelton, 2001; Darley et al., 2000; Das, 2008).

Outside of microstructure, researchers have developed MM strategies for a variety of settings, including prediction markets (Abernethy et al., 2011; Chen and Pennock, 2007; Hanson, 2007), dealer-mediated markets (Das, 2005; Jumadinova and Dasgupta, 2010), CDAs (Feng et al., 2004), and environments where prices are generated exogenously (Abernethy and Kale, 2013). In this last category, Chakraborty and Kearns (2011) demonstrate the profitability of market making given a mean-reverting price series series. They propose

a simple MM algorithm to submit a ladder of prices; the market makers I investigate can be viewed as variations on this strategy.

None of these studies, however, address questions about allocative efficiency in the market. To my knowledge, the literature on welfare effects of MM behavior is quite limited, and existing studies are largely concerned with how adverse selection affects allocative efficiency. For example, Bessembinder et al. (2011) demonstrate that restricting spread widths improves allocative efficiency and encourages more traders to become informed. Their results suggest that MMs enhance efficiency primarily when information asymmetries are significant. Brusco and Jackson (1999) illustrate the inefficiencies of competitive markets in a two-period model in which the market maker position is designated via an auction. They also design a system of trading rules to reach an efficient allocation by identifying and incentivizing MM agents. Huang and Wang (2010) propose a model in which provision of liquidity is endogenous, finding that mandating participation tends to improve welfare, but that the welfare effects of lowering costs for liquidity provision per se are ambiguous. In a similar vein, Bessembinder et al. (2015) present a model in which a firm can sell an asset to an investor in an IPO, with the option of paying a designated market maker (DMM) in exchange for liquidity provision in a secondary market. When the secondary market is illiquid due to asymmetric information and uncertainty regarding the asset's fundamental value, social welfare can be improved if the firm enters into a DMM contract.

4.4 Market Environment

To investigate the effect of market making on allocative efficiency, I construct a simple model of a single security traded in a continuous double auction market. The market environment is populated by multiple background traders, representing investors, and (optionally) one market maker. Background traders each have an individual valuation for the security (described in Section 2.2), and they employ parameterizations of the Zero Intelligence (ZI) strategy described in Section 2.3. At any given time, the background investors

are restricted to a single order to buy or sell one unit, whereas the MM may maintain orders to buy and sell any number of units at various prices.

4.5 Market Maker Strategies

Much of the prior work on MM strategies treats the market maker as a dealer (Das, 2005; Glosten and Milgrom, 1985), which must take one side of each trade. In my model, however, all trades execute through the CDA order book, therefore the MM submits limit orders just as background traders do. I consider a family of MM strategies that submit at time t a *ladder* of single-quantity buy and sell orders, comprised of K rungs spaced ξ ticks apart:

$$\begin{cases} [S_t, S_t + \xi, S_t + 2\xi, \ldots, S_t + K\xi] & \text{for sell orders} \\ [B_t - K\xi, \ldots, B_t - 2\xi, B_t - \xi, B_t] & \text{for buy orders} \end{cases}$$

with $S_t > B_t$ and $K, \xi > 0$. The MM arrives at time 0 and reenters the market according to a Poisson process with rate λ_{MM}. On reentry at time t, the MM observes the current fundamental r_t, which it may use in determining its ladder of buy and sell orders. It cancels any standing orders remaining from its previous ladder when submitting a new ladder.

Like the background traders, the MM liquidates its inventory at the end of the trading horizon. The liquidation price is the global fundamental value r_T. The MM's total profit is defined by the sum of trading cash flow plus liquidation proceeds.

To avoid crossing the current BID-ASK quote, the MM truncates its ladder. Specifically, if $BID_t > S_t$ (or similarly, $B_t > ASK_t$), the agent cuts the ladder off at the rung that is at or above (below) the current BID (ASK) price. The truncated ladder is:

$$\begin{cases} [S_t + (K - x)\xi, \ldots, S_t + K\xi] & \text{if } BID_t > S_t \\ [B_t - K\xi, \ldots, B_t - (K - x)\xi] & \text{if } B_t > ASK_t, \end{cases}$$

where $x > 0$ specifies the rung immediately above BID (for sell orders) or below ASK

(for buy orders). That is, x satisfies the condition $S_t+(K-x-1)\xi < BID_t < S_t+(K-x)\xi$ for sell orders in the ladder, and $B_t - (K-x)\xi < ASK_t < B_t - (K-x-1)\xi$ for buy orders.

The MM uses its observation of the current fundamental r_t to inform its ladder construction. Specifically, the MM strategies I implement compute an estimate $\hat{r}_t$ of the terminal fundamental r_T via (2.1), and center the ladder around this estimate. The spread ω is set by a strategy parameter. The central ladder prices are:

$$S_t = \hat{r}_t + \frac{1}{2}\omega,\ B_t = \hat{r}_t - \frac{1}{2}\omega.$$

4.6 Experiments

Generally speaking, I am most interested in the effect of market making in equilibrium, when all agents are doing their best to generate profit. For the present study, I employ the discrete-event market simulation system described in Section 3.1. I consider only a restricted set of available strategy choices, defined by selected parameterized versions of the background-trader and MM strategies. I generate data for various combinations of the strategies introduced in Sections 2.3 and 4.5, each sampled over many runs (at least 20,000 per profile, often many more) to account for stochastic effects (valuation schedules, trajectories of the market fundamental, agent arrival patterns). I determine equilibria among these strategies through empirical game-theoretic analysis (described in Section 3.2). I then take these equilibria as the basis for evaluating MM welfare effects.

The experiments conducted for the present study supersede those reported in AAMAS 2015 Conference Proceedings (Wah and Wellman, 2015). The present results incorporate an expanded strategy set (Table 4.1) and subtle changes to the background-trader arrival process. I also more thoroughly sample the profile space, covering more profiles and with more simulations per profile. The results are qualitatively consistent with the previous findings, though with a more ambiguous relationship between trading horizon and MM

impact on surplus gains. The expanded strategy set includes background traders who persist in strict shading indefinitely, which affords greater scope for MM benefit even over long trading horizons.

4.6.1 Environment Settings

I evaluate the performance of background traders and the MM within 21 parametrically distinct environments. For each environment, I analyze two empirical games that differ in whether the MM is present. In all settings, there are $N \in \{25, 66\}$ background traders. Each simulation run lasts T time steps, for $T \in \{1, 4, 12, 24\} \times 10^3$. If present, the MM in each environment enters the market at the start of the simulation and reenters with rate $\lambda_{MM} = 0.005$, or on average once every 200 time steps. The global fundamental has a mean value $\bar{r} = 10^5$ and mean-reversion parameter $\kappa = 0.05$. The variance for the private value vector is $\sigma^2_{PV} = 5 \times 10^6$.

The environments differ in number of background traders (N), background-trader reentry rate (λ_{BG}), fundamental shock variance (σ^2_s), and time horizon (T). The configurations of parameter settings for $N \in \{25, 66\}$ background traders are as follows.

A $\lambda_{BG} = 0.0005$, $\sigma^2_s = 1 \times 10^6$

B $\lambda_{BG} = 0.005$, $\sigma^2_s = 1 \times 10^6$

C $\lambda_{BG} = 0.005$, $\sigma^2_s = 5 \times 10^5$

I describe each environment by its configuration label, followed by time horizon (in thousands). For example, B12 is the environment labeled **B** above with $T = 12000$.

4.6.2 EGTA Process

Recall that I model a financial market as a role-symmetric game, in which players are partitioned into roles, each with a specified strategy set (Section 3.2). The two roles in my model are background trader (25 or 66 players) and market maker (one player). As

game size grows exponentially with the number of players and strategies, I apply deviation-preserving reduction (described in Section 3.2.2) to approximate the many-player game as a game with fewer players. I choose values for N in this study to facilitate DPR by ensuring that the required aggregations come out as integers: the approximation of an $(N, 1)$-size game (i.e., N background traders and 1 MM) by a $(k, 1)$-player reduced game works best when k divides N and $k - 1$ divides $N - 1$. Specifically, I use simulation data from the $(66, 1)$-agent environments to estimate reduced $(6, 1)$-player games, where six players represent the 66 background traders in the simulated environment. I similarly estimate $(5, 1)$-player games from the $(25, 1)$-agent cases.

I iteratively apply EGTA to guide my exploration of the strategy space. In this study, I successfully found at least one and at most four non-trivial RSNEs for each game evaluated, with support sizes up to four for background traders and up to two for MMs. The process accumulates a dataset of profile simulation results, which I use to estimate payoff values for strategy profiles in the game.

For all the games I model, there exists a trivial pure RSNE in which all agents play a "NOOP" strategy that refrains from bidding. This exists because if none of the other agents (background traders or MM) submit limit orders, then there is nobody to trade with and there will be no transactions regardless of the strategy the subject agent employs. In my discussion below, I ignore this degenerate equilibrium, which obviously has payoff zero for all agents.

4.6.3 Social Optimum

To provide a benchmark for efficiency, I calculate the social optimum based on the trader population and valuation distribution used in my environments (i.e., $N \in \{25, 66\}$ background traders with parameters $q_{\max} = 10$ and $\sigma^2_{PV} = 5 \times 10^6$). I determine the optimum for a particular draw of N valuation vectors by treating each as a demand curve and finding a uniform competitive equilibrium price. This is conveniently implemented in my

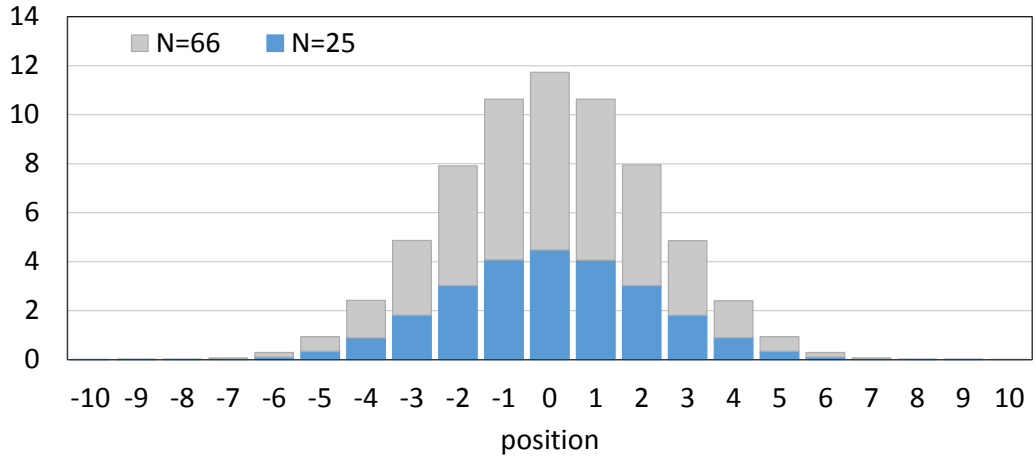


Figure 4.2: Histograms of the net position (i.e., number of units traded) of N background traders in the socially optimal allocations. The distributions (shown superimposed) are compiled from 20,000 samples.

simulation environment, where valuation vector Θ_i is represented by a background trader i, who submits $q_{\max}$ single-unit sell orders at prices $\bar{r} + \theta_i^s$, $s \in \{-q_{\max} + 1, \ldots, 0\}$, and $q_{\max}$ single-unit buy orders at prices $\bar{r} + \theta_i^b$, $b \in \{+1, \ldots, q_{\max}\}$. A call market computes a uniform clearing price to match supply and demand, which defines the optimal allocation for the sample. From 20,000 samples, I find a mean social welfare of 44155 and 16306 for 66 and 25 background traders, respectively. Figure 4.2 presents histograms of trades per background trader in the social optima.

4.7 Results

4.7.1 Game without Market Making

The empirical games without MM cover 14 background-trader strategies: 13 versions of ZI (see Table 4.1), and (implicitly) the no-trade strategy NOOP. I identified 1–3 ZI equilibria for each of my 21 environments as listed in Tables 4.2 and 4.3 (see Tables A.1 and A.2 for complete specifications of the equilibria found). For each equilibrium, I estimated background-trader surplus by sampling 2,500 profiles according to the equilibrium mixture, running 25–100 simulations per sampled profile (at least 62,500 full-game simu-

Table 4.1: ZI strategy combinations included in empirical game-theoretic analysis of games with and without a market maker.

$R_{\min}$	$R_{\max}$	η
0	65	0.8
0	125	0.8
0	125	1
0	250	0.8
0	250	1
0	500	1
0	1000	0.8
0	1000	1
0	1500	0.6
0	2500	1
250	500	1
500	1000	0.4
1000	2000	0.4

Table 4.2: Symmetric equilibria for games without market makers, $N = 66$, calculated from the 6-player DPR approximation. Each row of the table describes one equilibrium found and its average values for total surplus and two strategy parameters: R_{mid} (the midpoint of ZI range $[R_{\min}, R_{\max}]$) and threshold η. Values presented are the average over strategies in the profile, weighted by mixture probabilities. Surplus values are means from thousands of simulations of the full game, where strategies are randomly sampled from the equilibrium mixed-strategy profile.

Env	Surplus	R_{mid}	η
A1	3712	750	0.4
A1	4439	374	0.980
A4	16578	340	0.977
A4	16551	353	1
A12	33741	267	0.955
B1	29150	441	0.894
B4	40392	411	0.961
B12	40102	494	0.810
C1	30803	250	1
C1	29726	375	1
C4	41130	500	0.8
C4	39901	390	0.976
C12	41410	446	0.923

lations in total) and then recording the aggregate surplus.

Table 4.3: Symmetric equilibria for games without market makers, $N = 25$, calculated from the 5-player DPR approximation. Data presented is as for Table 4.2.

Env	Surplus	R_{mid}	η
A1	1041	750	0.404
A1	1351	371	1
A4	5616	350	0.986
A12	11670	335	1
A24	13697	375	1
A24	15162	218	0.949
A24	15543	117	0.826
B1	8752	750	0.4
B4	14041	517	0.773
B12	14256	553	0.715
B24	14478	556	0.710
C1	10378	375	1
C4	14225	476	0.838
C12	14617	441	0.894
C24	14618	490	0.816

4.7.2 Game with Market Making

My games with MM include the 14 background-trader strategies from the no-MM treatment above, plus 5–7 strategies for the MM role. The MMs employed in my game analysis are as described in Section 4.5, with $K = 100$ rungs spaced $\xi \in \{25, 50, 100\}$ units apart. Each MM strategy type employs a fixed spread $\omega \in \{64, 128, 256, 512, 1024\}$. Rung size ξ is 50, plus one variant with $\xi = 25$ for $\omega = 256$ and another variant with $\xi = 100$ for $\omega = 512$. The set of all MM strategies employed is in Table 4.4. Note that in some environments, only a subset of five of these strategies is used. The equilibria found are presented in Tables 4.5 and 4.6 (see Tables A.3 and A.4 for complete specifications of the RSNE found). Background-trader surplus and MM profit are estimated for each equilibrium based on the sampling method described for the no-MM game above.

Table 4.4: MM strategy parameter combinations explored. Strategies specified by the first two rows were omitted for some environments.

K	ξ	ω
100	25	256
100	50	64
100	50	128
100	50	256
100	50	512
100	50	1024
100	100	512

Table 4.5: Role-symmetric equilibria for games with one market maker, $N = 66$, calculated from the $(6, 1)$-player DPR approximation. Each row of the table describes one equilibrium found and its average values for background-trader surplus, MM profit, and four strategy parameters: R_{mid} (the midpoint of ZI range $[R_{\min}, R_{\max}]$), threshold η, MM spread ω, and rung size ξ. Values presented are the average over strategies in the profile, weighted by mixture probabilities.

Env	Surplus	Profit	R_{mid}	η	ω	ξ
A1	4545	461	238	0.933	512	61
A1	4553	485	205	1	512	100
A1	4465	382	298	0.943	512	50
A4	16503	1890	139	0.834	256	50
A12	32984	3196	109	0.948	256	50
A12	31920	3231	116	0.949	256	50
B1	29238	21	433	0.907	931	59
B4	40041	196	431	0.942	512	100
B12	40575	113	400	0.974	512	100
B12	42304	820	492	0.806	491	50
C1	29507	302	375	1	512	50
C4	39669	878	421	0.954	256	50
C4	41416	1715	240	0.984	256	50
C12	41658	2003	500	0.864	256	50
C12	40836	1233	431	0.911	256	50
C12	42037	1572	455	0.976	256	50

4.7.3 Comparison of Market Performance

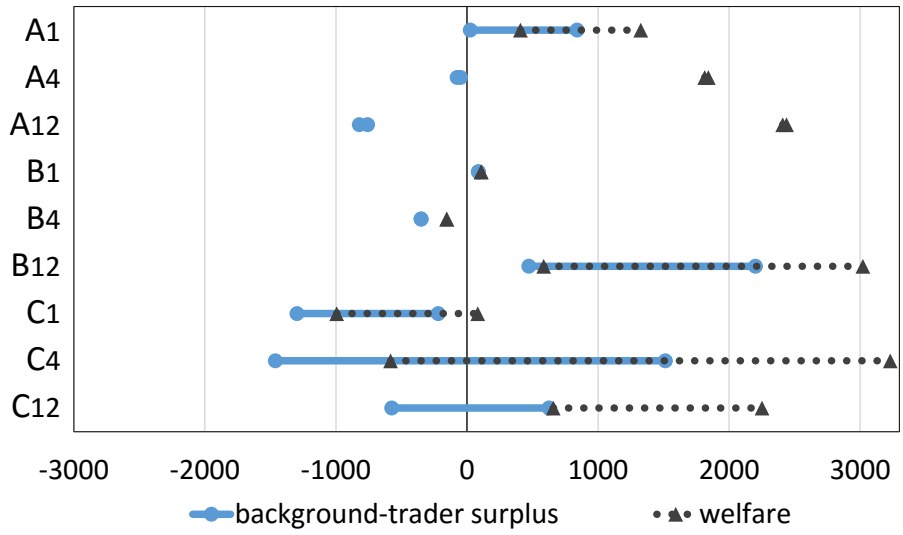
My findings with regard to the central question in this chapter are presented in Figure 4.3. For each environment, I compare equilibrium outcomes, with and without an MM,

Table 4.6: Role-symmetric equilibria for games with one market maker, $N = 25$, calculated from the $(5, 1)$-player DPR approximation. Data presented is as for Table 4.5.

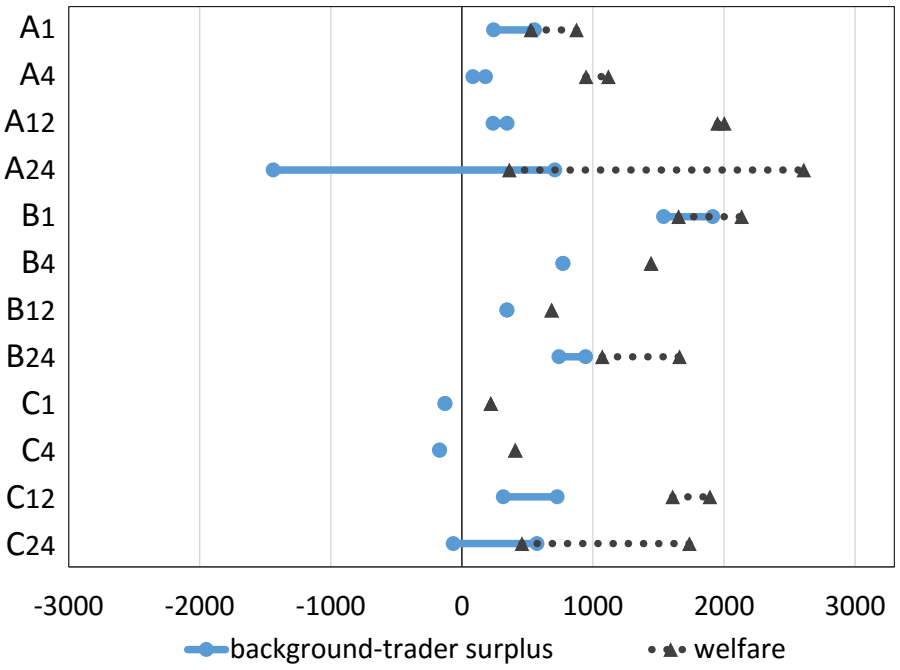
Env	Surplus	Profit	R_{mid}	η	ω	ξ
A1	1593	285	164	0.863	512	50
A1	1596	320	115	0.940	512	100
A4	5711	852	188	0.896	512	100
A4	5773	961	125	0.956	512	100
A4	5701	874	186	0.916	512	50
A4	5797	934	117	1	512	100
A12	11975	1696	88	0.883	512	25
A12	11907	1728	111	0.803	512	25
A12	12014	1606	103	1	512	47
A24	14103	1802	151	0.822	430	50
A24	14406	1899	57	0.8	256	50
B1	10666	220	329	1	512	100
B1	10292	113	446	0.886	512	100
B4	14813	671	478	0.818	294	50
B12	14560	341	444	0.889	512	67
B24	15423	716	552	0.825	256	50
B24	15219	330	424	0.994	512	100
C1	10251	348	375	1	512	100
C4	14055	577	365	1	512	50
C12	15121	1343	382	0.863	256	25
C12	14933	1292	254	0.931	512	50
C12	15344	1166	225	0.996	256	50
C24	15191	1163	502	0.814	256	50
C24	14552	524	388	0.979	512	50

on two measures: social welfare and background-trader surplus. Since there are often multiple equilibria, the differences are presented as ranges, delimiting the most and least favorable comparisons.

In the scenarios with 66 background traders (Figure 4.3(a)), the change in overall welfare is generally positive, with only three environments (B4, C1, and C4) providing small exceptions. The change in background-trader surplus, in contrast, varies widely across environments, with multiple examples of both positive and negative changes. The effect is strongly negative in the A environments with longer trading horizons, which may be explained by the significant information advantage of the MM over background traders due



(a) $N = 66$



(b) $N = 25$

Figure 4.3: The effect of presence of a single MM on background-trader surplus and social welfare in equilibrium, across all environments. Differences are presented as ranges, reflecting the multiplicity of equilibria found in some environments. The left point of each range is the minimum gain (in some cases a loss), that is, the lowest value observed with an equilibrium with MM minus the highest value observed in any equilibrium without MM. The right point is the maximum improvement observed: the difference between the highest value with a MM and the lowest without MM.

to their disparate reentry rates ($\lambda_{MM} = 0.005$ versus $\lambda_r = 0.0005$). For environments B4, B12, C4, and C12, the total social welfare without MM is over 90% of the socially efficient outcome of 44155. That is, the ZI background traders in these environments extract a high fraction of the potential surplus in the market on their own. Intuitively, given sufficient time for reentry (as governed by horizon T and reentry rate λ_r), agents with private values on the right side of competitive prices will eventually trade, and any inefficient trades can effectively be reversed. When the background traders have sufficient time to reach efficient outcomes, the MM may provide little benefit to overall welfare, and its profits tend to come out of background-trader surplus. Accordingly, I observe that the MM degrades investor surplus for some or all equilibria in three of these four cases (see Figure 4.4 for surplus comparison across environments with $N = 66$).

The trading horizon T reflects whatever might limit an investor's patience (liquidity needs, portfolio hedging, cost of monitoring, etc.). By curbing agents' ability to find efficient trades, the time constraint limits their ability to extract all potential surplus solely by trading with each other. This problem is exacerbated in a thin market, where agents encounter fewer potential counterparties per unit time. Both factors increase the likelihood that agents trade inefficiently, as they lack sufficient time and opportunity to reverse poor transactions. In such scenarios, the MM can boost not only overall welfare but also background-trader surplus by facilitating trade among impatient investors arriving at different times. In my study, for markets populated by 25 background traders (Figure 4.3(b)), the market maker improves welfare in all twelve test environments. It improves background-trader surplus unambiguously in eight, and with a range mostly on the positive side in one more. Two more cases exhibit small negative effects, and one (environment A24, a long horizon with slow traders) exhibits a large predominantly negative effect.

I observe that background traders are prone to shade less (i.e., midpoint R_{mid} of the ZI bid range is lower) when MM is present, particularly for $N = 25$ and the A environments with $N = 66$. These are also the environments where MM tends to improve background-

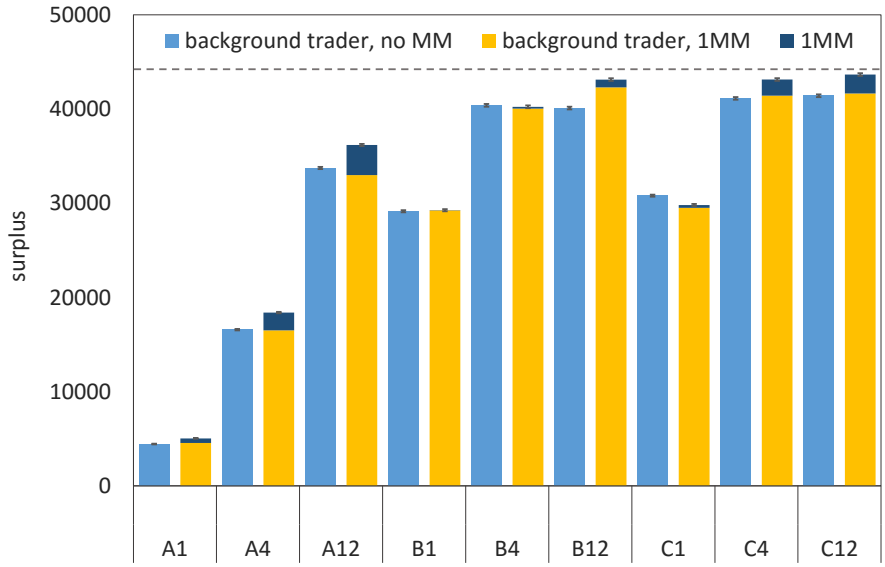


Figure 4.4: Comparison of background-trader surplus (with and without market making) and MM profit for $N = 66$. The dotted line is the optimal social welfare available (44155). Error bars indicate the 95% confidence interval for total welfare in the maximum-welfare role-symmetric Nash equilibrium in each environment, with and without MM. Each bar is compiled from 10,000 samples.

trader surplus (see Figure 4.5 for surplus comparison across environments with $N = 25$). This indicates that the MM facilitates optimal allocations: with MM present, background investors can demand less surplus per trade, yet still achieve greater payoff than without the market maker. I also find that MM spread ω tends to be larger for environments with shorter trading horizons, as would be expected when traders are more impatient.

Finally, I evaluate liquidity for the maximum-welfare RSNE (Figure 4.6), with and without MM, by sampling results from profiles at the RSNE proportions. I measure liquidity via the BID-ASK spread (narrower spreads reflect greater liquidity) and background-trader execution time (interval between order submission and transaction). In general, both spreads and execution times drop with MM, which is indicative of the liquidity-provisioning capacity of the MM. In the thinner markets, spreads without MM are significantly wider than in the thicker markets, as would be expected. The presence of the MM serves to significantly narrow spreads nearly down to the levels present in the more populous environments.

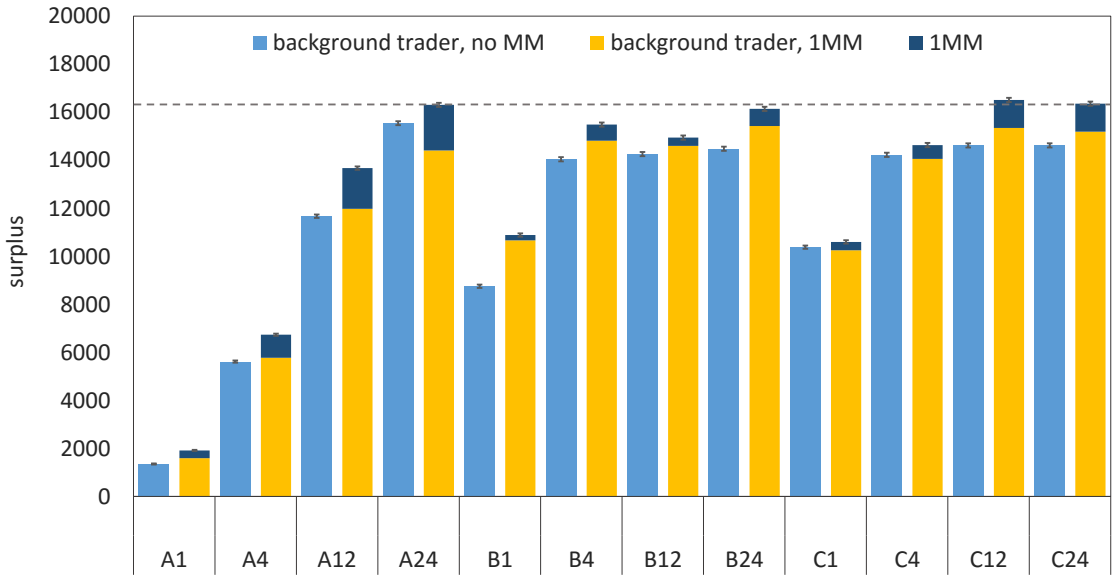
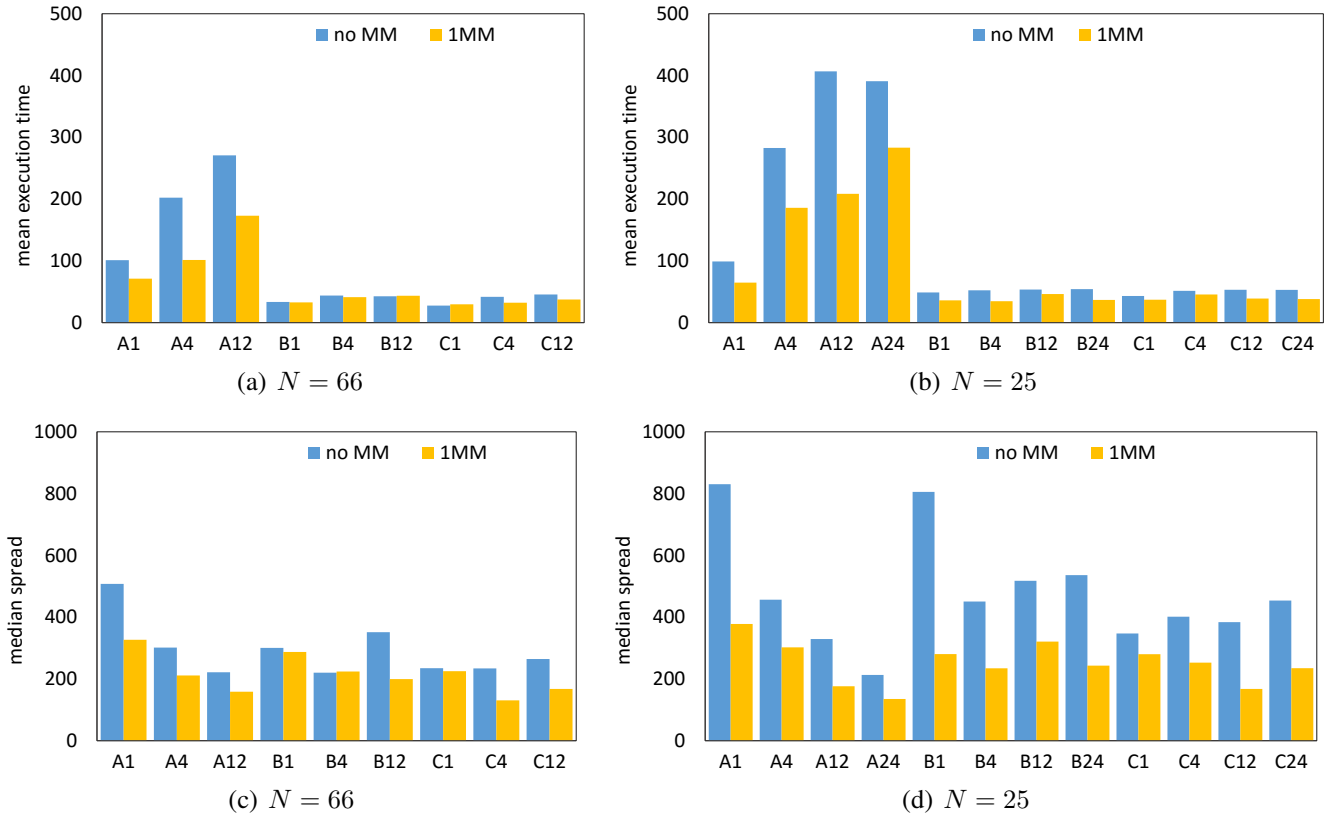


Figure 4.5: Comparison of background-trader surplus (with and without market making) and MM profit for $N = 25$. The dotted line is the optimal social welfare available (16306). Error bars indicate the 95% confidence interval for total welfare in the maximum-welfare role-symmetric Nash equilibrium in each environment, with and without MM. Each bar is compiled from 10,000 samples.



(a) $N = 66$ (b) $N = 25$

(c) $N = 66$ (d) $N = 25$

Figure 4.6: Comparison of background-trader execution time (Figures 4.6(a) and 4.6(b)) and median spread (Figures 4.6(c) and 4.6(d)) for the maximum-welfare RSNE in each environment, with and without MM. Mixed-strategy RSNE are approximated by profiles with trader population proportions corresponding to the strategy probabilities. Each bar is compiled from 10,000 samples.

4.7.4 Liquidity Measures as Proxies for Welfare

The fact that the liquidity proxy measures reported in the previous section improve with MM in environments where background-trader surplus does not, however, underscores that these measures are not adequate substitutes for direct evaluation of investor welfare. To address this question, I compare two spread measures to welfare in five pure-strategy profiles. The findings here are also reported in a paper to appear in the *Russell Sage Foundation Journal of the Social Sciences* (Wellman and Wah, 2016).

One way to estimate welfare is the spread measure (also called *quoted spread*) defined in Section 2.4. Recall that spread is measured as the difference between the BID and ASK quotes for a given point in time. Quotes may vary significantly over time, however, and in such cases, aggregating quoted spreads over all time steps may not be an accurate reflection of changes in welfare. An alternative measure is the *effective spread*, which focuses on spreads at the time of trade (Bessembinder, 2003; Madhavan et al., 2002).

To examine the correspondence between spreads and and welfare, I simulate 10,000 samples of five pure-strategy profiles for $N = 66$ and $N = 25$ under configuration B12. The strategies all belong to the ZI family, with the following ranges ($\eta = 1$ unless otherwise stated):

- B12a: ZI [0, 125] with $\eta = 0.8$
- B12b: ZI [0,250]
- B12c: ZI [0, 1000]
- B12d: ZI [0, 2500]
- B12e: ZI [500, 1000] with $\eta = 0.4$

In each of these profiles, all N traders play the specified strategy. The surplus of each profile is shown in Figure 4.7, and the corresponding spread measures are in Figure 4.8. I summarize quoted spread for a scenario run as the median spread over all time steps, and

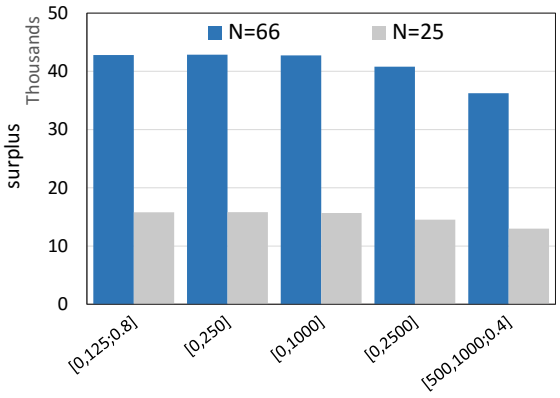
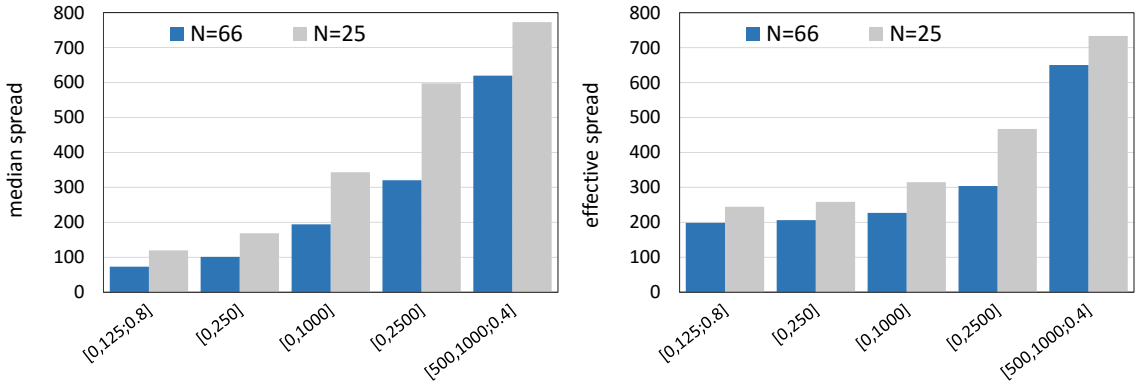


Figure 4.7: Overall surplus in five pure-strategy profiles for $N = 66$ and $N = 25$ in configuration B12. The ZI strategies are written in the form $[R_{\min}, R_{\max}; \eta]$, unless $\eta = 1$, in which case it is omitted from the label.



(a) Quoted spread

(b) Effective spread

Figure 4.8: Quoted spread and effective spread in five pure-strategy profiles for $N = 66$ and $N = 25$ in configuration B12. The ZI strategies are written in the form $[R_{\min}, R_{\max}; \eta]$, unless $\eta = 1$, in which case it is omitted from the label.

I take the mean BID-ASK difference over all times when a trade occurs as my aggregate measure of effective spread. I find that for both $N = 66$ and $N = 25$, the lowest surplus is observed in B12e, whereas surplus is relatively constant for B12a–c. Both spread measures, however, widen significantly going from B12a to B12e, which accurately reflects the welfare improvement in B12c–e, but fails to capture the flat welfare across profiles B12a–c. These results demonstrate that the accuracy of quoted and effective spread measures as predictors of welfare can be limited.

4.8 Conclusions

Market makers are generally considered to serve a valuable function in continuous market mechanisms by providing liquidity to bridge ebbs and flows of trader orders. The precise impact of this behavior, however, depends on specific features of market environments and trading strategies. I conducted a systematic agent-based simulation study to compare several parameterized environments with and without market-maker agents. I modeled a single security traded in a CDA populated by multiple background traders, and I characterized the strategic play in the induced empirical game model. This enabled the comparison of outcomes in equilibrium, that is, allowing the background traders and market makers to strategically react to each others' presence.

My analysis demonstrates the generally beneficial effects of market making on efficiency, and shows that whether these benefits accrue to background investors depends on market characteristics. Specifically, I find a tendency of a market maker to improve the welfare of impatient investors (those in thin markets or relatively few opportunities to trade with each other), but not in general. My results also show that liquidity proxy measures such as spread are inadequate substitutes for direct evaluation of investor welfare.

My study has several limitations, which must be taken into account in assessing my conclusions. First, my methods involve sampling, approximation, and limited search, all of which bear on the accuracy of equilibrium determinations. Sampling error is mitigated through the large number of simulation runs I gather over a breadth of environments and profiles, so not a fundamental concern for my conclusions here. The player reduction method I employ (DPR) has been shown to produce good approximate equilibrium estimates on other problems (Wiedenbeck and Wellman, 2012), and for my purposes approximate equilibria provide a sufficient basis for outcome comparison. However, DPR estimates are not guaranteed approximations. Even within the DPR game, I am unable to evaluate all profiles and cannot be sure that I have found all equilibria.

A second area of limitation is the relatively narrow exploration of strategies. The equili-

bration process selects a combination of ZI parameters (from those included in the strategy set) that is best suited for the given environment. Nevertheless, further investigation may yield improved versions of ZI or other strategies (for example adaptive variants (Cliff, 2009; Vytelingum et al., 2008)) that could alter equilibrium findings. Similar improvements may be found on the MM side, for instance with strategies incorporating learning (Abernethy and Kale, 2013). I include only one MM in the analysis reported here, but Mason Wright has led an extension of this study to evaluate market maker competition, finding that this leads to further background-trader gains (Wah et al., 2016).

Finally, my exploration of environments is also far from exhaustive. Whereas covering all plausible environments is infeasible, a broader range of variation on number of players, valuation distributions, and fundamental dynamics could go a long way in illuminating and validating robust conditions for qualitative welfare effects of market making in continuous double auctions.

CHAPTER V

Latency Arbitrage, Market Fragmentation, and Efficiency: A Two-Market Model

In this chapter, I present a study on the impact of *latency arbitrage*, a type of high-frequency trading strategy in which traders exercise superior speed in order to exploit price disparities between markets. I examine the effect of latency arbitrage on allocative efficiency and liquidity in fragmented financial markets. I propose a simple model of latency arbitrage in which a single security is traded on two exchanges, with aggregate information available to regular traders only after some delay. An infinitely fast arbitrageur profits from market fragmentation by reaping the surplus when the two markets diverge due to this latency in cross-market communication. I employ the discrete-event simulation system presented in Section 3.1 to capture this processing and information transfer delay, and I simulate the interactions between high-frequency and Zero Intelligence trading agents within three different market environments. I then evaluate allocative efficiency and market liquidity arising from the simulated order streams, and I find that market fragmentation and the presence of a latency arbitrageur reduces total surplus and negatively impacts liquidity. Replacing continuous-time markets with frequent call markets eliminates latency arbitrage opportunities and achieves further efficiency gains through the aggregation of orders over short time periods.

This chapter extends and supersedes a paper by Wah and Wellman (2013) presented

at the *14th ACM Conference on Electronic Commerce*. In that study, traders employed a fixed strategy for all market configurations and latency settings. The analysis presented in this chapter employs empirical game-theoretic methods (Section 3.2) to perform strategy selection for traders. The qualitative conclusions presented by Wah and Wellman (2013) still hold; the results I report here serve to confirm those main points in a more strategically valid evaluation.

5.1 Introduction

Although algorithmic trading has been a reality for many years now, the pervasiveness, speed, and autonomy of trading algorithms are reaching new heights. *High-frequency trading* (HFT)—characterized by large numbers of small orders in compressed periods, with positions held for extremely short durations—is estimated to have accounted for as much as 78% of total trading volume in 2009, up from nearly zero in 1995 (Schneider, 2012).[1] The practice of HFT has generated several public controversies regarding its ramifications for the transparency and fairness of market operations as well as its effects on market volatility and stability.

Many HFT strategies exploit advantages in *latency*—the time it takes to access and respond to market information. Trading on these advantages has been estimated to account for $21 billion in profit per year (Schneider, 2012).[2] HF traders achieve such advantages by investing in specialized computer hardware and software, co-locating servers in exchanges' data centers, and constructing dedicated communication lines (Goldstein et al., 2014).

The HFT strategy I examine here is *latency arbitrage*, where an advantage in access and response time enables the trader to book a certain profit. Arbitrage is the practice of

[1] Definitive figures are elusive, but proportions exceeding two-thirds are widely reported, for instance 73% in "SEC runs eye over high-speed trading," *Financial Times*, 29 July 2009. This no doubt includes straightforward monitoring for arbitrage opportunities—for example between index securities and their defining constituents, which itself has long represented a large fraction of exchange trading volume.

[2] Profit figures are considerably more uncertain than volume estimates. Kearns et al. (2010) present an interesting approach to derive an upper bound on HFT profits. Presumably the billions HFT firms invest annually in technology and infrastructure (Adler, 2012) represent a lower bound on gross trading profit.

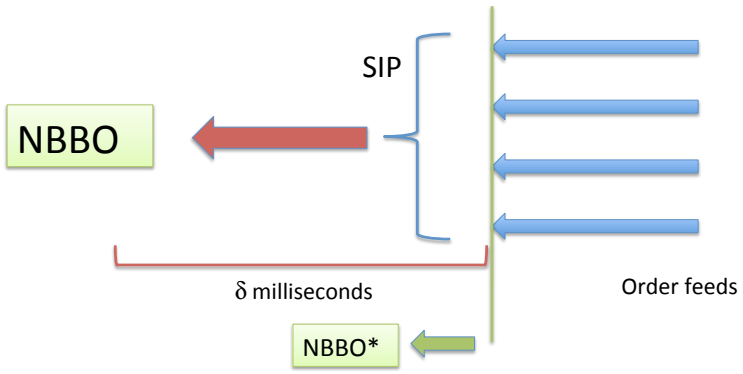


Figure 5.1: Exploitation of latency differential. Rapid processing of the order stream enables private computation of the NBBO before it is reflected in the public quote from the SIP.

exploiting disparities in the price at which equivalent goods can be traded in different markets. Such disparities can arise in financial markets in several ways, and the term "latency arbitrage" has been applied to a variety of practices that exploit speed advantages. Cross-market latency arbitrage opportunities are quite prevalent across U.S. stock exchanges, with total potential yearly profit in 2014 exceeding $3 billion (Wah, 2016). In this chapter, I model a specific type of latency arbitrage in which disparities arise from the fragmentation of securities markets across multiple exchanges. This fragmentation has been a major trend, particularly in the United States over the last decade (Arnuk and Saluzzi, 2012). U.S. securities regulations have attempted to mitigate the effect of fragmentation through the formulation of Regulation NMS, which mandates cross-market communication and the routing of orders for best execution (Blume, 2007; Securities and Exchange Commission, 2005). Orders stream into exchanges, which are required to feed summary information about their best buy and sell orders to an entity called the Security Information Processor (SIP). The SIP continually updates public price quotes called the "National Best Bid and Offer" (NBBO).

I illustrate this process and the potential for latency arbitrage in Figure 5.1. Given order information from exchanges, the SIP takes some finite time, say δ milliseconds, to compute and disseminate the NBBO. A computationally advantaged trader who can process the or-

der stream in less than δ milliseconds can simply out-compute the SIP to derive NBBO*, a projection of the future NBBO that will be seen by the public. By anticipating future NBBO, an HFT algorithm can capitalize on cross-market disparities before they are reflected in the public price quote, in effect jumping ahead of incoming orders to pocket a small but sure profit. Naturally this precipitates an arms race, as an even faster trader can calculate an NBBO** to see the future of NBBO*, and so on.

The latency arms race as sketched above is fundamentally an outgrowth of *continuous trading*: a property of mechanisms that distinguish precedence according to arbitrarily small time differences. By moving to a discrete-time model—which introduces short but finite clearing intervals (as in a *frequent call market*, or frequent batch auction)—I can neutralize small disparities in information access and response time. A driving question of this work is how such a mechanism-design intervention would affect market performance.

More broadly, I seek to understand not only the effects of latency arbitrage on market efficiency and liquidity, but also the interplay between fragmentation, clearing mechanisms, and latency arbitrage strategies in producing this performance. Such questions about HFT implications are inherently computational, as the very speed of operation renders details of internal market operations—especially the structure of communication channels—systematically relevant to market performance. In particular, the latencies between market events (transactions, price updates, order submissions) and when market participants observe these activities become pivotal, as even the smallest latency differential can significantly affect trading outcomes. Lacking suitable data to study these questions empirically,[3] I pursue a simulation approach.

I present a simple model that captures the effect of latency across two markets with a single security. My model captures the interplay of latency and fragmentation as well as

[3]Order activity at the temporal granularity of interest here is generally unavailable for public research, and it is unclear whether data on communication latencies and the end-to-end routing of orders among brokers and exchanges is available from any source. What high-frequency trading data does exist commercially is prohibitively expensive. Moreover, even full details on conceivably observable trading activity could not directly resolve counterfactual questions, such as the response of financial markets to possible shocks or the effects of alternative market rules and regulations.

the regulatory environment responsible for current equity market structure, and I have the first results quantifying the effect of latency arbitrage on surplus allocation as a function of latency and market rules. Using an agent-based approach, I simulate the interactions between high-frequency and background traders, and I employ empirical game-theoretic analysis to identify equilibria under different market conditions. I evaluate efficiency (as measured by total surplus) arising from the simulated orders, under a range of latency settings. My main finding is that latency arbitrage not only reduces profits of the background traders, but also diminishes surplus overall. Perhaps surprisingly, market fragmentation per se does not harm efficiency; in fact some degree of fragmentation mitigates inefficient trades that are often executed by a continuous mechanism. The discrete-time frequent call market eliminates latency arbitrage by construction and, by virtue of temporal aggregation, yet more effectively matches orders, producing significantly greater surplus.

This chapter is structured as follows. In Section 5.2, I discuss related work on agent-based financial markets and models of HFT and market structure. I describe my two-market model in Section 5.3. In Section 5.4, I discuss my experiments. I present my results in Section 5.5 and conclude in Section 5.6.

5.2 Related Work

5.2.1 Agent-Based Financial Markets

There is a substantial literature on agent-based modeling (ABM) of financial markets (Buchanan, 2009; Farmer and Foley, 2009; LeBaron, 2006), much of it geared to reproduce and thereby explain stylized facts from empirical studies of market behavior. For example, simulated markets have been constructed to reproduce phenomena observed in real stock markets, such as bubbles and crashes (LeBaron et al., 1999; Lee et al., 2011). Because agent behavior is shaped by the market environment, which includes interactions with other agents over time, such models can support causal reasoning (as in the study by Thurner et al.

(2012) establishing the effect of leverage on price volatility). One prominent example of an agent-based financial market is the Santa Fe artificial stock market (Palmer et al., 1994; LeBaron, 2004). ABM has also been used to model financial markets for applications such as portfolio selection (Jacobs et al., 2004) and determining the distributions of order and trading waiting times in a limit order book (Raberto and Cincotti, 2005).

5.2.2 High-Frequency Trading Models

Much of the current literature on the effects of HFT relies on the evaluation of historical order data. Hasbrouck and Saar (2013) use NASDAQ order data to construct sequences of linked messages describing trading strategies. They find that this low-latency activity improves short-term volatility, spreads, and market depth. Brogaard (2010) analyzes a 120-stock NASDAQ dataset that distinguishes HFT from non-HFT activity in order to assess the impact of high-frequency trading on liquidity, price discovery, and volatility. Prior work suggests that algorithmic trading improves liquidity (Hendershott et al., 2011); Angel et al. (2011) reach similar conclusions, finding that the emergence of automated trading and HFT has improved various market measures such as execution speed and spreads. Additional work suggests a link between HFT and increased volatility (Arnuk and Saluzzi, 2012). Foucault et al. (2015) examine latency arbitrage opportunities in currency markets, and provide evidence of a tradeoff between pricing efficiency and liquidity. In another study, Baron et al. (2012) find that some kinds of HFT activities directly harm ordinary investors.

Others rely on theoretical analysis to determine the optimal behavior of high-frequency traders. Avellaneda and Stoikov (2008) derive an optimal limit order submission strategy for a single high-frequency trader acting as a liquidity provider, running numerical simulations to assess the agent's performance under varying strategies. Cohen and Szpruch (2012) propose a single-market model of latency arbitrage with one limit order book and two investors operating at different speeds. The fast trader employs a strategy that deter-

mines in advance the quantity the slow investor intends to trade, using this information to generate a risk-free profit. Jarrow and Protter (2012) develop a model of traders with differentials in speed and access to information, showing that HFT transactions can degrade price discovery, exacerbate volatility and increase mispricings—which HF arbitrageurs can then exploit.

In a rare application of ABM to HFT, Hanson (2012) finds that market liquidity and total surplus vary directly with the number of HF traders.

5.2.3 Modeling Market Structure and Clearing Rules

Several prior works seek to identify the effects of market fragmentation and clearing rules, mainly via anecdotal evidence elicited from historical data. On the theoretical side, Mendelson (1987) investigates the effect of consolidation versus fragmentation of periodic call markets, without consideration of arbitrage between the submarkets. O'Hara and Ye (2011) use historical quote data and execution metrics to demonstrate that market fragmentation does not appear to harm measures such as spreads, execution speed, and efficiency. Bennett and Wei (2006) compare the execution costs of stocks that have switched from the NASDAQ to the more consolidated NYSE, finding evidence that execution costs decline with order flow consolidation. Amihud et al. (2003) examine the response of equities on the Tel Aviv Stock Exchange to the exercise of corporate warrants, concluding that consolidation improves liquidity.

However, few prior studies attempt to directly model the communication latencies arising from market fragmentation and the resultant arbitrage opportunities, with the exception of Ding et al. (2014), who analyze NBBO latencies and the ability of HFTs to generate a synthetic NBBO. They conclude that price dislocations between the official and synthetic NBBOs can be exploited by HFTs for profit.

Switching to a discrete-time clearing mechanism, as in a frequent call market, has already been proposed as a means to eliminate the exploitation of latency differentials across

multiple exchanges (Wellman, 2009; Schwartz and Peng, 2013; Sparrow, 2012). Budish et al. (2015) analyze a theoretical model of a continuous limit order book, showing that HFT profits in equilibrium come from investors via wider spreads and that frequent batch auctions reduce the value of very small speed advantages. Others have proposed variants on the frequent call market with randomized clearing intervals (Sellberg, 2010; Industry Super Network, 2013), or randomized batching in conjunction with pro rata trade allocation rules, which may promote more equitable allocation of trades among investors (Farmer and Skouras, 2012; McPartland, 2013).

A number of other studies have focused not on the role of call markets in mitigating the harmful effects of HFT, but on the differences in market quality offered in a discrete-time versus a continuous market (Pancs, 2013; Pellizzari and Dal Forno, 2007) or an alternative market rule such as selective delay, in which cancellation orders are processed immediately but all other order types have a small delay (Baldauf and Mollner, 2014).

Empirical work on the effects of switching to periodic clearing is limited and again relies largely on the analysis of historical events (Webb et al., 2007; Kalay et al., 2002). For example, Amihud et al. (1997) find that switching from a daily call auction to a combination of discrete and continuous trading in the Tel Aviv Stock Exchange is associated with improvements in liquidity.

5.2.4 Two-Market Model in Relation to Prior Work

To study latency arbitrage as made possible by market fragmentation, I construct an agent-based model populated by representative trading strategies interacting within carefully specified market mechanisms. My model comprises a latency arbitrageur and multiple non-HF traders, with a single security whose trading is fragmented across two markets. This two-market model captures the connections between market fragmentation, communication latencies, regulations, and latency arbitrage. As discussed above, previous analytical or agent-based HFT models employ a single market or order book—rendering them

incapable of capturing the effect of fragmentation—and they fail to incorporate the communication delays enabling cross-exchange arbitrage.

I implement my model in the discrete-event simulation system described in Section 3.1, which explicitly models the communication patterns between background investors, exchanges, and the SIP operating in current U.S. equity markets. I then compare allocative efficiency in equilibrium in the two-market model with the welfare in other models of market structure, including a centralized continuous double auction market and a frequent call market.

5.3 Two-Market Model

I present a simple model for latency arbitrage across two markets populated by a single high-frequency trader and multiple background traders. I describe the specifics of this model in Section 5.3.1. The valuation model and class of strategies employed by the background investors are described in Sections 2.2 and 2.3, respectively. In Section 5.3.2, I discuss the behavior of the latency arbitrageur. I present an example of how a latency arbitrage opportunity may arise in this two-market model in Section 5.3.3.

5.3.1 Model Description

My model of latency arbitrage consists of one security traded on two markets, each employing a *continuous double auction* mechanism (Section 2.1). The two markets are linked by a public NBBO signal (see Figure 5.2). Limit orders lodged in either market are forwarded to the SIP, which calculates and reports an NBBO—based on the quotes from the two markets—with some finite delay δ. This latency reflects the time required to receive information about activities in the two markets and compute an updated public price signal.

Retail and institutional investors generate limit orders according to an evolving fundamental (driven by news) and other private factors. Each non-HF investor is primarily associated with one of the markets. An order is sent to the trader's primary market unless

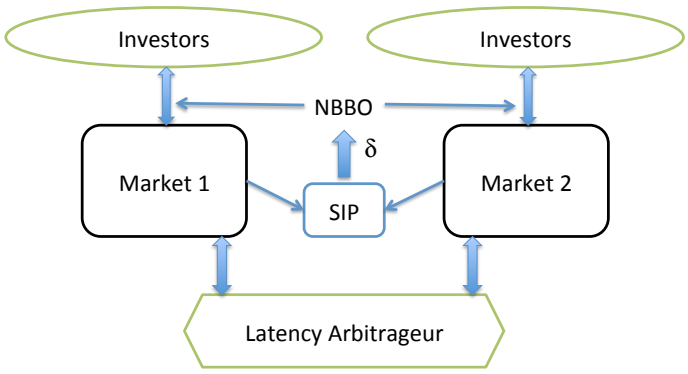


Figure 5.2: Two-market model with one infinitely fast latency arbitrageur and multiple background investors. A single security is traded on the two markets. Each background investor is associated primarily with one of the two markets, and its order is routed to its alternate market if and only if the NBBO quote indicates an immediate execution. The latency arbitrageur has undelayed access to both markets, so it can immediately detect arbitrage opportunities arising from the delay in NBBO calculation.

the NBBO indicates that it could be executed in the alternate market at a price better than that available on the primary market.

More precisely, let BID^j and ASK^j, where $j \in \{1, 2\}$, denote the current BID and ASK quotes, respectively, in market j. Similarly, let BID^N and ASK^N represent the NBBO quote. Background traders have direct access to the quotes on their primary market and the NBBO, but not to those on the alternate market. Suppose a trader associated with market 1 generates a limit order to buy a unit at price p. This order is routed to market 2 if and only if $p \geq ASK^N$ and $ASK^N < ASK^1$. Otherwise, the order goes to market 1, the trader's primary market. Note that the conditions for submitting to the alternate market entail that the trader's order would execute there immediately, if in fact the NBBO reflects the current global state. If the order is routed to the primary market, it may execute right away (if $p \geq ASK^1$); otherwise, it is added to market 1's order book. The rule for routing sell orders is analogous.

The latency arbitrageur in this model can determine the best prices in each market before the NBBO updates, due to its ability to receive and process order streams faster than background investors. It can thus immediately detect an arbitrage situation, which occurs

whenever $BID^1 > ASK^2$ or $BID^2 > ASK^1$. I assume the arbitrageur can respond infinitely fast, so it quickly takes the profit from such arbitrage situations by submitting executable orders to the two markets. Note that the arbitrage opportunity can arise only to the extent that the NBBO information is out of date. If the SIP were able to compute and publish the NBBO with zero latency, then a new order would always be routed correctly and would thereby execute immediately if there were a matching order in either market. Any finite delay, however, opens the possibility that an order is routed to the investor's primary market, despite there being a matching order in the alternate market that had arrived too recently to be admitted in the available NBBO. An out-of-date NBBO can also cause an order to be improperly routed to the alternate market despite it no longer matching there, even if there is a matching order in the primary market.

5.3.2 Latency Arbitrageur

The latency arbitrageur (LA) in the two-market model operates as follows. LA first obtains current price quotes in both markets, then checks whether an arbitrage situation exists. I denote the best price available to sell at by

$$BID^* \equiv \max\{BID^1, BID^2\},$$

and I denote the best price available to buy by

$$ASK^* \equiv \min\{ASK^1, ASK^2\}.$$

Given a threshold $\alpha \geq 0$, LA deems the current state a worthwhile arbitrage opportunity if and only if $BID^* > (1+\alpha)\, ASK^*$. To execute the arbitrage, LA submits orders exploiting the price differential to the two markets simultaneously. Under my assumption that LA is infinitely fast, bidding any price at or better than the current quote would lead to successful execution at the quoted prices. In my implementation, LA calculates the midpoint m be-

tween BID^* and ASK^*, then submits an order to buy at $\lfloor m \rfloor$ to the market with the better ASK price and an order to sell at price $\lceil m \rceil$ to the market with the better BID price. LA surplus (i.e., profit) for these trades is $BID^* - ASK^*$.

5.3.3 Example

Figure 5.3 illustrates how a latency arbitrage opportunity may arise in my two-market model. At time t, the NBBO quote is $BID^N = 104$ and $ASK^N = 110$. Consider background trader i, who wishes to submit a sell order at 105 to market 1, its primary market. To determine the order routing, BID^1 is compared with the NBBO. As $BID^N > BID^1$, the alternate market appears to be superior. However, a sell offer at 105 would not transact immediately (since $BID^N = 104$), so agent i's order is routed to market 1. At the beginning of time $t + 1$, for latency $\delta > 1$, the SIP has not yet updated the NBBO to include the order submitted at time t. Thus, the NBBO available to background investors is out of date: the correct quote would be $(104, 105)$, but the NBBO at time $t + 1$ is still $(104, 110)$ and matches ASK^2 in market 2, incoming agent $i + 1$'s primary market. Consequently, agent $i + 1$'s buy order at price 109 is routed to its primary market. At this point, BID^2 (at price 109, submitted by agent $i + 1$) exceeds ASK^1 (at price 105, submitted by agent i), which defines an arbitrage opportunity. Since LA is infinitely fast, it capitalizes on this disparity by submitting bids to buy at 107 in market 1 and sell at 107 in market 2, realizing a profit of 4.

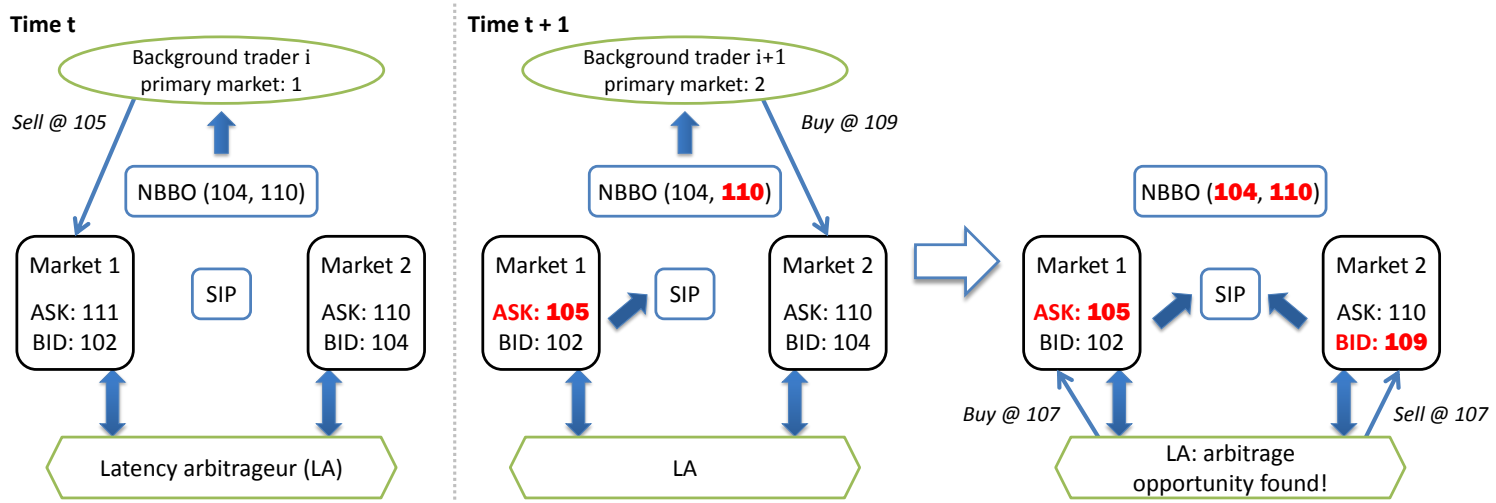


Figure 5.3: Emergence of a latency arbitrage opportunity over two time steps in the two-market model. All orders are for single-unit quantities. A red, bolded price highlights a discrepancy between the actual market state and the NBBO, represented in the diagram as $\left(BID^N, ASK^N\right)$. At time t, the NBBO is up to date. Background trader i wishes to sell at price 105. Since $BID^N < 105$ (which indicates non-immediate execution), the investor's order is routed to market 1. At time $t+1$, the NBBO is out of date, as the SIP updates the public quote with some delay δ. Background trader $i+1$ wishes to buy at 109; based on the NBBO, its order is routed to market 2, its primary market. (Had its order been routed to market 1, its bid would have transacted immediately.) The submission of its order to the inferior market opens up an arbitrage opportunity between the two markets ($BID^2 > ASK^1$), which LA immediately exploits for a guaranteed profit.

5.4 Experiments

To isolate the ramifications of market fragmentation, I consider two forms of centralized market configurations in my simulations: a CDA and a frequent call market. Recall that in contrast to a continuous-time market, clearing in a frequent call market takes place at designated intervals (Section 2.1). A frequent call market eliminates latency arbitrage opportunities, as the periodic clearing mechanism makes it impossible to gain or exploit informational advantages over other market participants within the clearing interval.

My experiments evaluate a variety of market configurations with respect to several performance measures described in Section 2.4. The configurations address the following central issues:

- **Presence of latency arbitrage**: I include configurations of the two-market model with and without LA.
- **Market fragmentation**: Along with the two-market model, I evaluate a centralized configuration where the two markets are consolidated as one.
- **Market clearing rules**: Along with continuous markets, I include a discrete-time call market setting. To facilitate direct comparison, in each run I set the clearing interval of the call market to equal the NBBO update latency.

5.4.1 Environment Settings

I evaluate and compare the performance of the four market structure configurations (two-market model with and without LA, CDA, and frequent call market) within three distinct environments. For the fragmented cases, an equal proportion of background traders is assigned primary affiliation with each market in a model. In the centralized call market, orders transact at a uniform price each time the market clears; this price is computed to best match supply and demand (Section 2.1).

In defining my environments, I selected environment parameters that generate sufficient arbitrage opportunities and also replicate the original findings for fixed-strategy, non-equilibrium comparisons from my previous study (Wah and Wellman, 2013). To do so, I explored a number of environments, varying the number of traders, trading horizon length, degree of mean reversion, and variance in both the fundamental and private values. In these runs, all traders employed a fixed strategy with $\eta = 1$, similar to the agents in the EC 2013 paper. I selected the environments reproducing the qualitative effects previously observed, and analyzed the impact of latency arbitrage, fragmentation, and discrete-time clearing in *equilibrium* via EGTA.

The threshold α for LA is fixed at 0.001. I set the mean fundamental value $\bar{r} = 10^5$, and the variance parameters $\sigma^2_{PV} = 5 \times 10^6$ and $\sigma^2_s = 5 \times 10^6$. All bids have single-unit quantities, and I assume zero transaction costs. Background traders play strategies from the set listed in Table 5.1.

The environments differ in number of background traders (N), background-trader reentry rate (λ_{BG}), value of the mean-reversion parameter (κ), and time horizon (T). For each market configuration in an environment, I explore a range of latency settings, with a minimum difference (or order of magnitude) of $\Delta_\delta \in \{10, 100\}$. The configurations of parameter settings are as follows.

Environment 1 $N = 24$, $\lambda_{BG} = 0.05$, $\kappa = 0.05$, $T = 15000$, $\Delta_\delta = 100$

Environment 2 $N = 238$, $\lambda_{BG} = 0.005$, $\kappa = 0.02$, $T = 10000$, $\Delta_\delta = 10$

Environment 3 $N = 58$, $\lambda_{BG} = 0.005$, $\kappa = 0.02$, $T = 5000$, $\Delta_\delta = 10$

The arrival rate parameter is either $\lambda_{BG} = 0.05$ or $\lambda_{BG} = 0.005$; each ZI agent arrives, on average, every 20 or 200 time steps.

Table 5.1: ZI strategy combinations included in empirical game-theoretic analysis of market structure games with varying latencies.

$R_{\min}$	$R_{\max}$	η
0	125	1
0	250	1
0	500	1
250	500	1
0	1000	1
500	1000	0.4
500	1000	1
0	1500	0.6
1000	2000	0.4
0	2500	0.4
0	2500	1

5.4.2 EGTA Process

I examine 23 empirical games within environment 1, which cover the four market configurations across 8 latency settings, with latency $\delta \in \{0, 100, 200, 300, 400, 600, 700, 900\}$. For environment 2, I include 8 empirical games (with latency $\delta \in \{0, 50, 100\}$), and I examine 14 games within environment 3 (with latency $\delta \in \{0, 25, 50, 75, 100\}$). The games in a given environment include one centralized CDA game (which is independent of latency), and one game for each of the other three market configurations (two fragmented cases and one with periodic clears) per latency setting simulated. At latency 0, the centralized frequent call market is equivalent to the CDA, and the two models with fragmentation are equivalent as there are no arbitrage opportunities at zero latency.

The market structure games are modeled as role-symmetric games with a single role, or equivalently as symmetric games (Section 3.2). I apply deviation-preserving reduction (Section 3.2.2) to generate an approximation of the full game with fewer players. In general, I choose the number of traders N to facilitate reduction to a k-player game. I estimate 4-player reduced games from full games with $N \in \{24, 238, 58\}$ players.

5.5 Results

I find that the presence of a latency arbitrageur reduces total surplus (Section 5.5.1) and has a mixed effect on market liquidity (Section 5.5.2). Eliminating fragmentation can improve surplus and execution metrics. Replacing continuous markets with frequent call markets eliminates latency arbitrage opportunities and achieves substantial efficiency gains in all three environments (Section 5.5.3).

I identified 1–3 equilibria for each of the 23 games in environment 1 (Tables 5.2 and B.1), the 8 games in environment 2 (Tables 5.3 and B.2), and the 14 games in environment 3 (Tables 5.4 and B.3). For each equilibrium, I estimated background-trader surplus, as well as LA profit if applicable, by sampling 500 profiles according to the equilibrium mixture, and running 100 simulations per sampled profile (50,000 full-game simulations in total).

Table 5.2: Symmetric equilibria for market structure games for environment 1, one per latency (or clearing interval) setting per market configuration, $N = 24$, calculated from the 4-player DPR approximation. Each row of the table describes one equilibrium found and its average values for background-trader surplus, LA profit, and two strategy parameters: R_{mid} (the midpoint of ZI range $[R_{\min}, R_{\max}]$) and threshold η. Values presented are the average over strategies in the profile, weighted by mixture probabilities. Surplus values are means from thousands of simulations of the full game, where strategies are randomly sampled from the equilibrium mixed strategy profile.

Model	Latency	Surplus	Profit	R_{mid}	η
CDA	–	10114	–	1298	0.458
CDA	–	10383	–	1377	0.4
2M	0	11807	–	1250	0.4
2M	0	11393	–	1034	0.506
Call	100	13471	–	682	0.695
2M (no LA)	100	9400	–	1439	0.4
2M (no LA)	100	10373	–	1008	0.4
2M (LA)	100	5919	3487	1266	0.4
Call	200	13308	–	687	0.703
2M (no LA)	200	10621	–	1144	0.4
2M (LA)	200	6358	3164	1420	0.4
Call	300	13107	–	721	0.679
2M (no LA)	300	10386	–	1402	0.4

Continued on next page

Table 5.2 – *Continued from previous page*

Model	Latency	Surplus	Profit	R_{mid}	η
2M (no LA)	300	11244	–	913	0.4
2M (LA)	300	6398	3224	1414	0.4
Call	400	13004	–	383	1
Call	400	12771	–	640	0.747
Call	400	12686	–	460	0.961
2M (no LA)	400	10438	–	1399	0.4
2M (LA)	400	6130	4018	1080	0.4
Call	600	12932	–	321	1
Call	600	12403	–	704	0.76
Call	600	12526	–	675	0.701
2M (no LA)	600	10182	–	750	0.4
2M (no LA)	600	11128	–	845	0.4
2M (LA)	600	7459	4349	1257	0.429
2M (LA)	600	6457	4460	932	0.4
2M (LA)	600	6509	3276	1411	0.429
Call	700	12910	–	294	0.957
Call	700	12868	–	287	0.958
2M (no LA)	700	9138	–	1343	0.442
2M (no LA)	700	11302	–	881	0.4
2M (LA)	700	5256	2958	1453	0.4
Call	900	12613	–	251	1
2M (no LA)	900	8641	–	1459	0.498
2M (no LA)	900	12358	–	1250	0.4
2M (no LA)	900	10710	–	1384	0.4
2M (LA)	900	4807	3121	1403	0.426
2M (LA)	900	6819	4825	1184	0.479

Table 5.3: Symmetric equilibria for market structure games for environment 2, one per latency (or clearing interval) setting per market configuration, $N = 238$, calculated from the 4-player DPR approximation. Data presented is as for Table 5.2.

Model	Latency	Surplus	Profit	R_{mid}	η
CDA	–	136079	–	1250	0.565
CDA	–	136140	–	1250	0.605
2M	0	134339	–	1077	0.488
Call	50	141816	–	1250	0.4

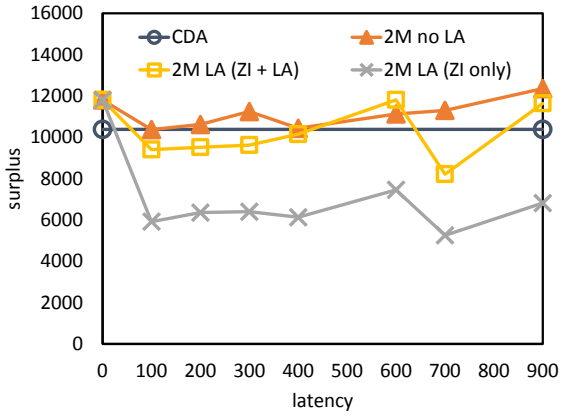
Continued on next page

Table 5.3 – *Continued from previous page*

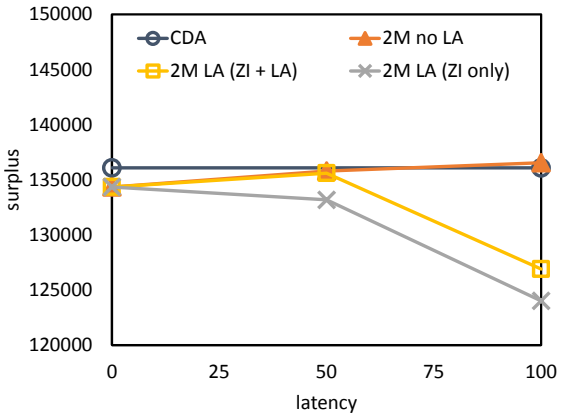
Model	Latency	Surplus	Profit	R_{mid}	η
2M (no LA)	50	135789	–	1068	0.497
2M (LA)	50	133177	2417	1062	0.513
Call	100	136961	–	1275	0.496
2M (no LA)	100	136542	–	1189	0.544
2M (LA)	100	124012	2888	1308	0.4

Table 5.4: Symmetric equilibria for market structure games for environment 3, one per latency (or clearing interval) setting per market configuration, $N = 58$, calculated from the 4-player DPR approximation. Data presented is as for Table 5.2.

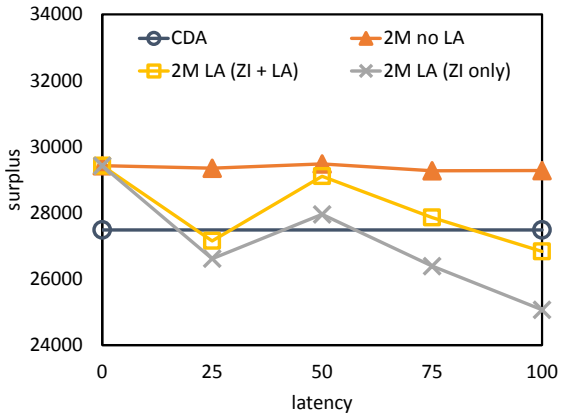
Model	Latency	Surplus	Profit	R_{mid}	η
CDA	–	27482	–	1312	0.4
2M	0	29424	–	1234	0.41
Call	25	30136	–	1191	0.559
2M (no LA)	25	29347	–	1250	0.487
2M (LA)	25	12300	161	1412	0.4
2M (LA)	25	26612	538	1303	0.4
Call	50	30310	–	1250	0.4
2M (no LA)	50	18704	–	1445	0.531
2M (no LA)	50	29479	–	1250	0.431
2M (LA)	50	16720	523	1377	0.524
2M (LA)	50	27953	1154	1228	0.413
Call	75	30587	–	1115	0.472
2M (no LA)	75	29271	–	1250	0.506
2M (LA)	75	26388	1470	1285	0.4
Call	100	27665	–	1295	0.4
2M (no LA)	100	19833	–	1430	0.565
2M (no LA)	100	29277	–	1250	0.497
2M (LA)	100	15965	1142	1398	0.449
2M (LA)	100	25070	1763	1292	0.409



(a) Environment 1



(b) Environment 2



(c) Environment 3

Figure 5.4: Total surplus in the two-market (2M) model, both with and without a latency arbitrageur, and in the centralized CDA market, for the three environments. In the two-market model with LA, both the total surplus (ZI + LA) and background-trader surplus (ZI only) are plotted. Each point reflects the average over 50,000 simulation runs of the maximum-welfare equilibrium for each market configuration and latency setting.

5.5.1 Effect of LA on Market Efficiency

Figure 5.4 shows the total surplus, for the centralized CDA and the two-market model with and without a latency arbitrageur, over multiple latency settings in the three environments. The total surplus of the two-market model without LA, as well as that of the centralized CDA market (an unfragmented continuous-time market), generally exceeds that of the two-market model with LA, whether or not the profits of LA are counted. This holds across the three environments. In other words, the latency arbitrageur takes surplus away from the background investors, and the amount it deducts exceeds the gross trading profit

it accrues.

Note that when latency is zero, the two fragmented models and the CDA market are effectively identical. The NBBO is always correct if there is no delay, so it is not possible for any latency arbitrage opportunities to emerge. It follows that the various market configurations at zero latency produce similar total surplus in equilibrium. Differences in equilibrium between the centralized and fragmented models may arise at zero latency when traders employ strategies with $\eta < 1$. In fragmented markets, traders decide whether to submit executable orders based on the current best quote in their assigned market, not the NBBO, so trader performance in the two-market model may differ from the centralized CDA.

In environment 1, LA significantly degrades efficiency in the two-market model, and total LA profit accounts for half of aggregate surplus once nonzero latency is introduced. Environments 2 and 3, however, have reduced mean reversion, which increases background traders' risk of adverse selection and having the LA pick off their standing orders. As a result, background traders in these two environments shade more in response to the LA. This can be seen by higher R_{mid} values in the RSNE found. Prior work by Zhan and Friedman (2007) has shown that some degree of bid shading can mitigate inefficient trades in CDAs. The infinitely fast arbitrageur immediately exploits arbitrage opportunities due to orders that are routed incorrectly; the LA's trades tend to be inefficient, contributing to the lower overall welfare observed in the two-market model with LA. Therefore, increased bid shading in the low mean reversion environments can alleviate some of these inefficiencies, which improves background-trader surplus and reduces LA profits.

Centralizing the markets in a consolidated CDA generally outperforms the fragmented market with LA in environments 1 and 2. This effect is muted in thinner markets when there are fewer trading opportunities, such as environment 3. As for the case without latency arbitrage, it may seem counterintuitive that welfare in the two-market model without LA is higher than in the centralized CDA in some environments. As discussed by Wah and

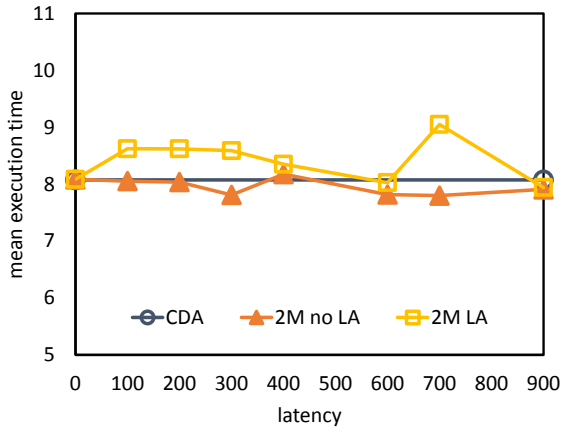
Wellman (2013), it turns out that fragmentation can actually provide a benefit for continuous markets, as the separated markets are less likely to admit inefficient trades (i.e., where both traders' values fall on the same side of the longer-term equilibrium price) that arise due to the vagaries of arrival sequences. LA can defeat this benefit by ensuring that any orders that would match in the central CDA also trade in the fragmented case, albeit with LA rather than with a counterpart investor (see Section 4.2 for an example illustrating the problem of allocative inefficiency in CDAs). This primarily applies in environment with sufficient trading opportunities, such as environments 1 and 2. In a thicker market as in environment 2, fragmentation does not always boost surplus in the two-market model without LA, as there are many traders in each market who can act as counterparties for trade.

5.5.2 Effect of LA on Liquidity

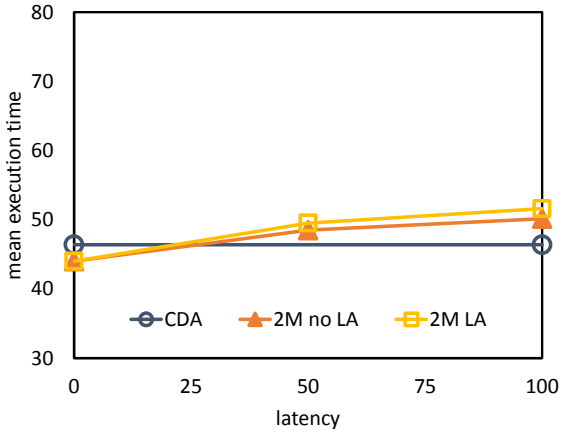
I also evaluate the effect of latency arbitrage on market liquidity, as measured via execution times and BID-ASK spreads. Figure 5.5 shows that execution time tends to be highest in the two-market model with LA. The fastest trade execution in environment 1 is achieved in the two-market model without LA, which differs from findings in the literature that trading at lower latencies improves overall execution time (Angel et al., 2011; Garvey and Wu, 2010; Riordan and Storkenmaier, 2012). This is largely due to the different strategies selected in equilibrium in this environment; traders tend to shade their bids less (i.e., R_{mid} is lower) in the fragmented model without LA, hence orders are more likely to execute sooner rather than later. The improvement in execution time is at best approximately 1–2 time steps, however, which is generally unobservable by non-HF traders.

Traders in the other two environments, however, do not shade more in equilibrium in the two-market model without LA. In these cases, the fastest execution is achieved in the centralized CDA, which makes sense, given the absence of communication latencies and thinness induced by fragmentation.

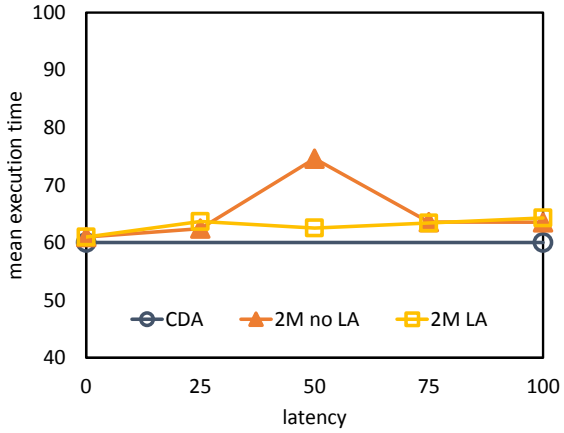
Spreads are a measure of liquidity costs in the market, as the BID-ASK differential



(a) Environment 1



(b) Environment 2



(c) Environment 3

Figure 5.5: Mean execution time in the two-market (2M) model, both with and without a latency arbitrageur, and in the centralized CDA market, for the three environments. Execution time is the difference between bid submission and transaction times.

will be smaller in a more liquid market (see Section 4.7.4 for a discussion of their effectiveness as a proxy measure of welfare). A tighter spread indicates lower liquidity cost, which directly translates to greater market liquidity. The widest spreads are generally in the two-market model with LA (Figure 5.6). LA also slightly exacerbates NBBO spreads, which are as a whole narrower than spreads of individual markets. The increase in spread could reflect an implicit transaction cost responsible for part of the surplus reduction observed above.

5.5.3 Frequent Call Market

Lastly, I evaluate the effect of switching to a discrete-time frequent call market. In my frequent call market configuration, the latency setting dictates the clearing period. Figure 5.7 shows that the total surplus in the centralized call market far exceeds that of the two-market model with LA, and the call market surplus is higher for all latency settings greater than 0 (there are only two market configurations at zero latency, the fragmented model without LA and the centralized CDA). By aggregating orders over time, call markets perform a more informed clear. They increase the probability that trades occur between intra-marginal traders—those with private valuations inside the equilibrium price range—and thus are less prone to executing inefficient trades than CDAs (Gode and Sunder, 1997).

As shown in Figure 5.8, the mean execution time in the centralized call market is much higher than that of the two-market model with LA. Unsurprisingly, I find that execution time in the centralized call market is higher than that observed in the other market configurations. As market clears occur less frequently in this market configuration, it takes longer for a bid to match and be removed from the order book. In environment 1, execution time in the frequent call market plateaus at approximately 20 time steps, which is equivalent to the average time between trader reentries. In the other two environments, the execution time in the call market increases monotonically with the length of the clearing interval, since trader reentries occur less frequently than market clears.

In Figure 5.9, I observe that the tightest spread is realized in the centralized call market, for all three environments. Spreads in the frequent call market are measured at the end of each market clear. They represent the market liquidity after orders have traded in each interval. Since the call market generally matches orders to trade more efficiently than the CDA, the spreads in the centralized call market tend to be tighter. The median spread decreases to some degree with latency due to the accumulation of bids in the order book, which is indicative of greater liquidity in the market. The temporal aggregation in the centralized call market is also responsible for similarly tight NBBO spreads (Figure 5.9(b)).

5.5.4 Relationship between Transactions and Surplus

Figure 5.10 shows the total number of transactions in each market configuration, for the three environments, averaged over all observations at a given latency. In all three environments, the total number of transactions in the centralized CDA and the two-market model without LA are generally comparable, though slightly lower in the latter. This is consistent with my observations of surplus patterns in Figure 5.4. The two-market model without LA results in higher surplus despite a reduction in number of transactions, indicating that each transaction in the fragmented model is associated with more surplus on average than in the centralized CDA.

The number of LA transactions does not increase with latency, although the number of arbitrage opportunities grows as the NBBO update delay increases. Since the background traders strategically respond to the presence of the LA by submitting executable orders over limit orders, they are less likely to be picked off by the LA.

In addition, the highest number of trades for a market configuration at a given latency setting in environment 1 is generally (although not always) observed in the call market. This is a result of the reduced R_{mid} values observed in the call market equilibria; traders in the frequent call market tend to shade their bids less in equilibrium, and consequently are more likely to trade. In contrast, transaction volume is generally lower in the frequent call market in the low mean reversion environments. Given the corresponding surplus improvement (Figure 5.4), this indicates that discrete-time clearing leads to higher surplus per trade.

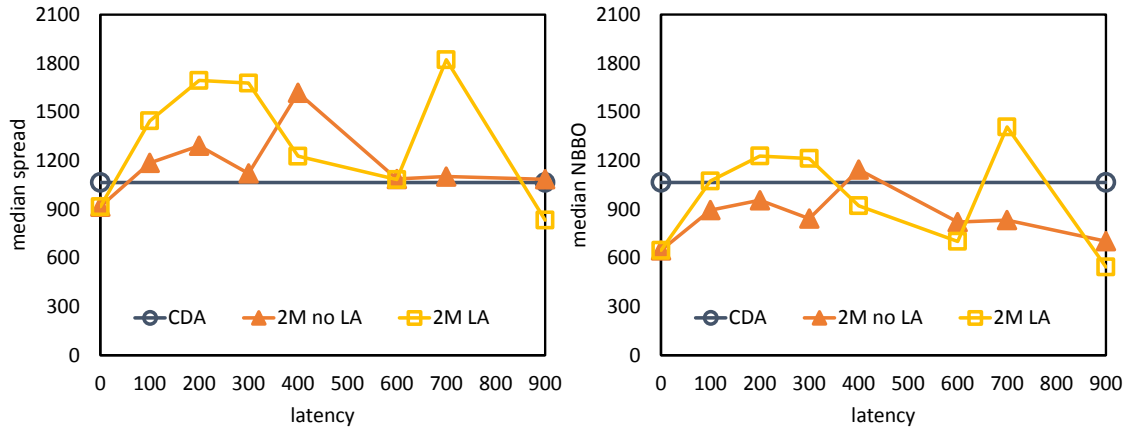
5.6 Conclusions

To understand an important phenomenon in high-frequency trading, I presented a two-market model of latency arbitrage. I implemented this model in a system combining agent-based modeling and discrete-event simulation. I employed empirical game-theoretic anal-

ysis to compute equilibria in games with variations in market structure and within three parametrically distinct environments, and I compared equilibrium outcomes in order to evaluate the interplay of latency arbitrage, market fragmentation, and market design, as well as their consequences for market performance. My results demonstrate that market efficiency in equilibrium is negatively affected by the actions of a latency arbitrageur, with no countervailing benefit in liquidity or any other measured market performance characteristic. Taking into consideration the substantial operational costs of the latency arms race would only amplify my conclusions about the harmful implications of this practice.

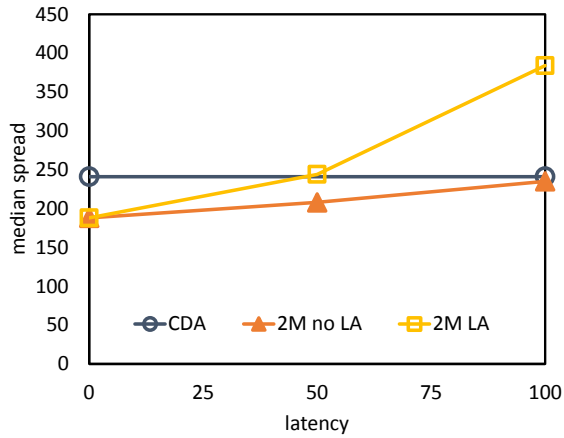
Virtually all modern financial markets employ continuous trading, which enables speed-advantaged traders to make risk-free profits over fragmented markets and which degrades overall efficiency. A frequent call market prevents high-frequency traders from gaining a latency advantage, thereby eliminating latency arbitrage opportunities and increasing surplus for background traders. Aggregating orders over small, regular time intervals provides efficiency gains over fragmented and continuous markets, and in fact these benefits appear to overshadow the gains attributable specifically to neutralizing latency arbitrage.

As with any simulation model, my results are valid only to the extent my assumptions capture the essence of real-world markets. Additional avenues for further study include examining the effect of more sophisticated HFT and background-trader strategies (such as those using historical information or responding to LA price signals), introducing other types of traders such as market makers, and further quantifying the impact of price discovery on efficiency.

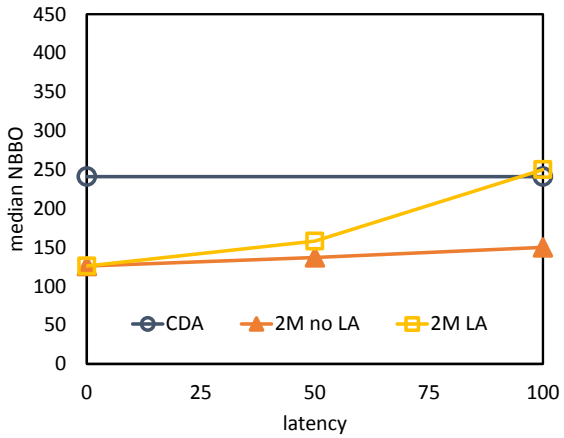


(a) Environment 1 median spread

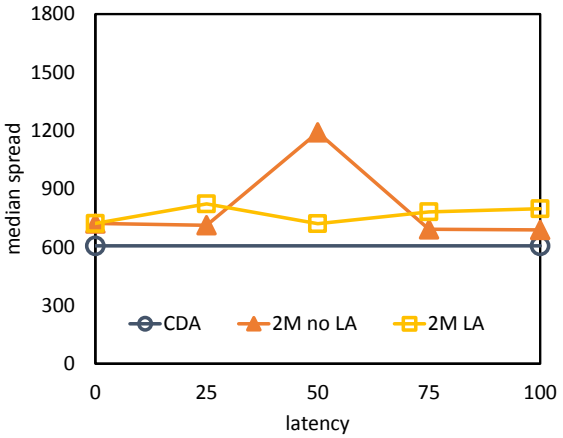
(b) Environment 1 median NBBO



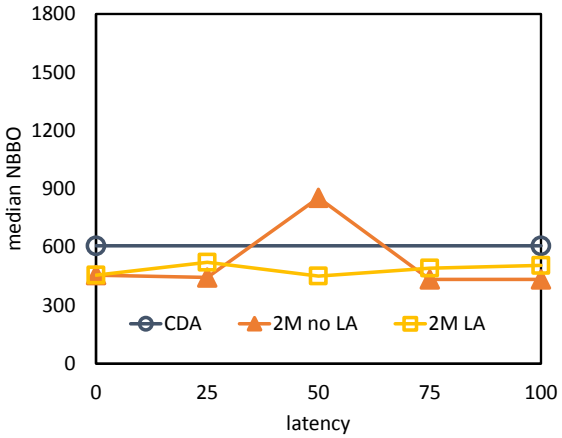
(c) Environment 2 median spread



(d) Environment 2 median NBBO

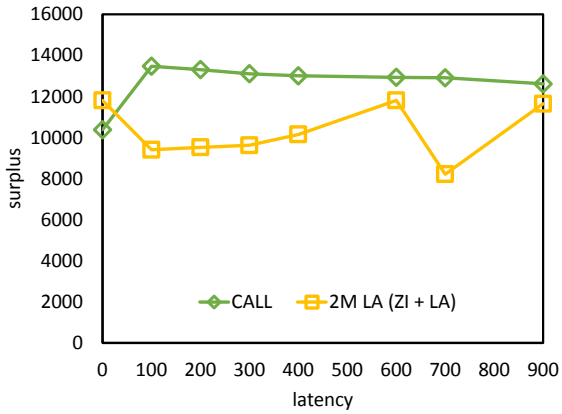


(e) Environment 3 median spread

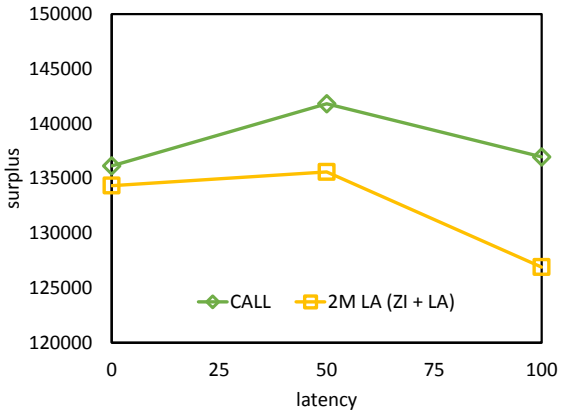


(f) Environment 3 median NBBO

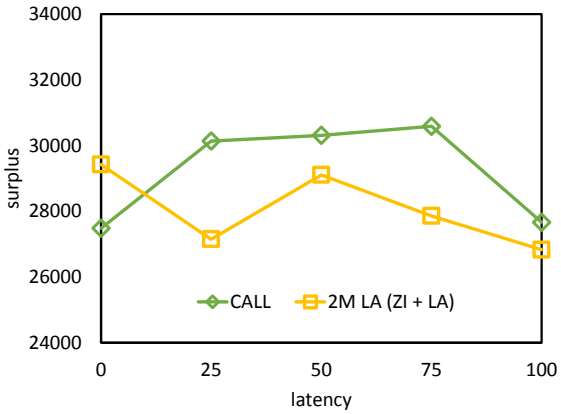
Figure 5.6: Median spread and NBBO spread in the two-market (2M) model, both with and without a latency arbitrageur, and in the centralized CDA market. Spread is the amount by which ASK exceeds BID. NBBO spread is the difference between BID and ASK of the NBBO quote. The spreads in the two-market models (2M) are the average of the median spread in the individual markets. Each point reflects the average over 50,000 simulation runs of the maximum-welfare equilibrium for the market configuration and latency setting.



(a) Environment 1

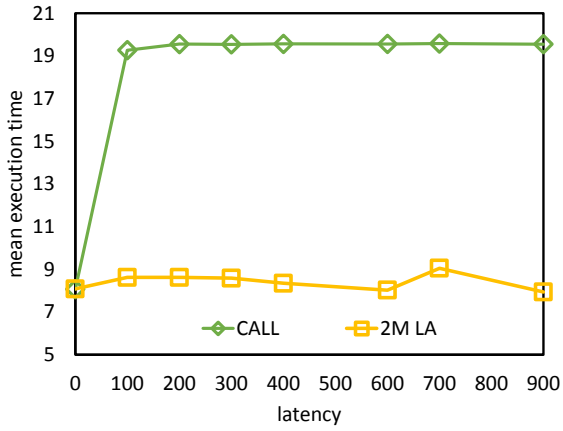


(b) Environment 2

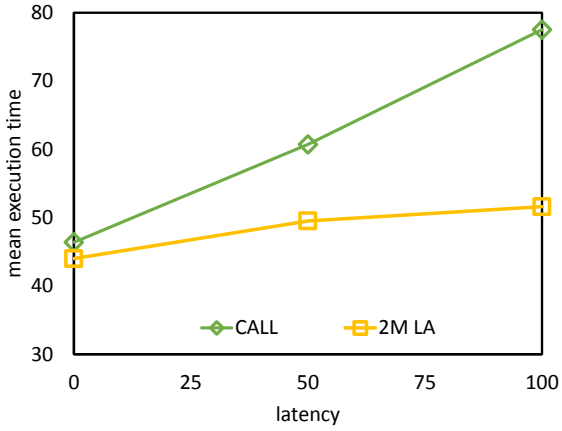


(c) Environment 3

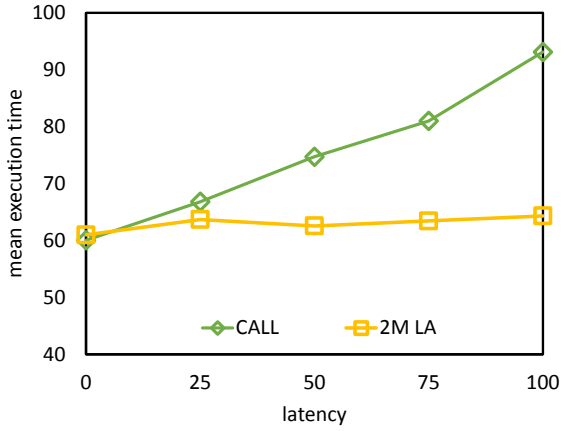
Figure 5.7: Total surplus for the centralized frequent call market and the two-market (2M) model with LA, for the three environments. Each point reflects the average over 50,000 simulation runs of the maximum-welfare equilibrium for each market configuration and latency setting.



(a) Environment 1

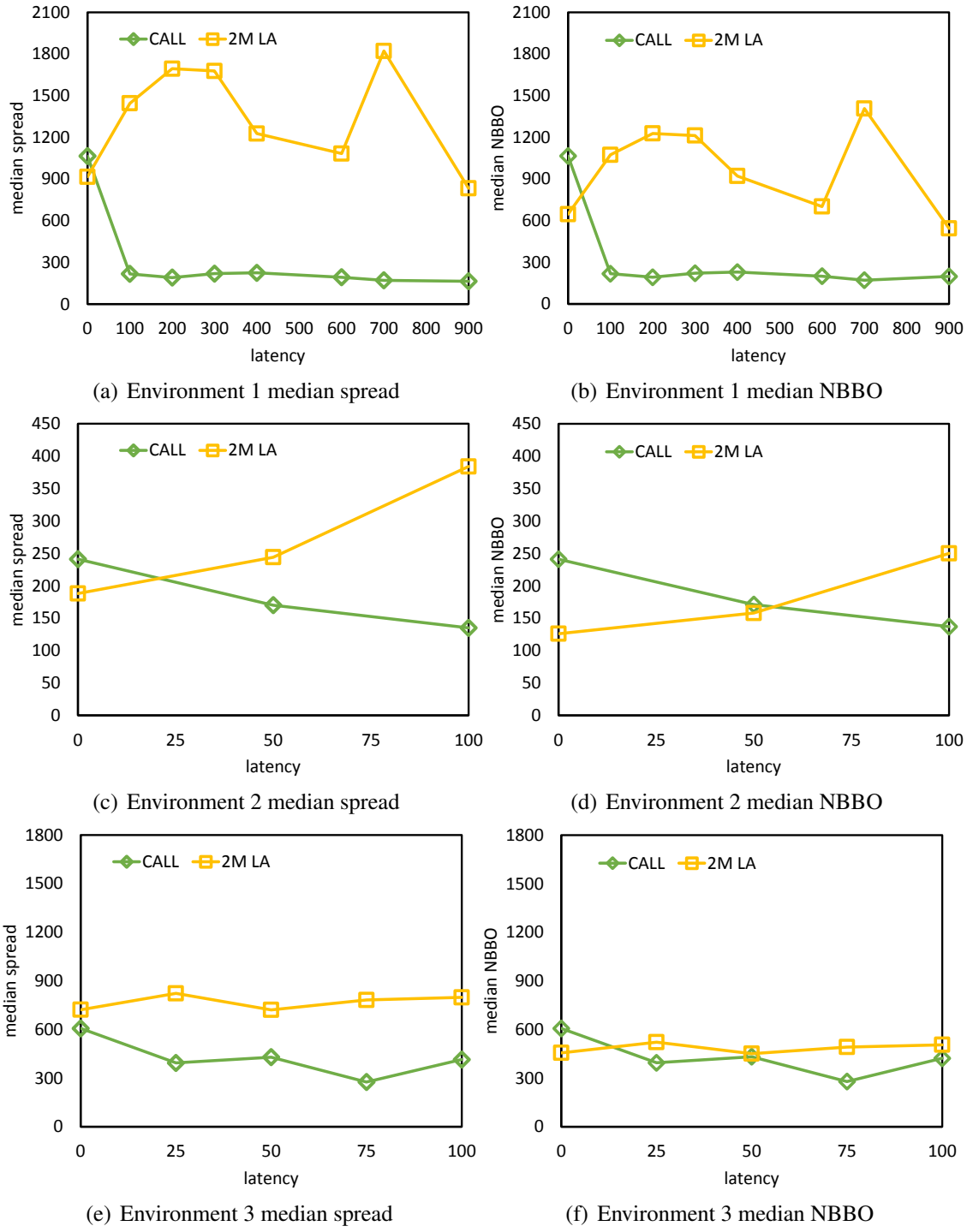


(b) Environment 2



(c) Environment 3

Figure 5.8: Execution time for the centralized frequent call market and the two-market (2M) model with LA, for the three environments. Each point reflects the average over 50,000 simulation runs of the maximum-welfare equilibrium for each market configuration and latency setting.



(a) Environment 1 median spread

(b) Environment 1 median NBBO

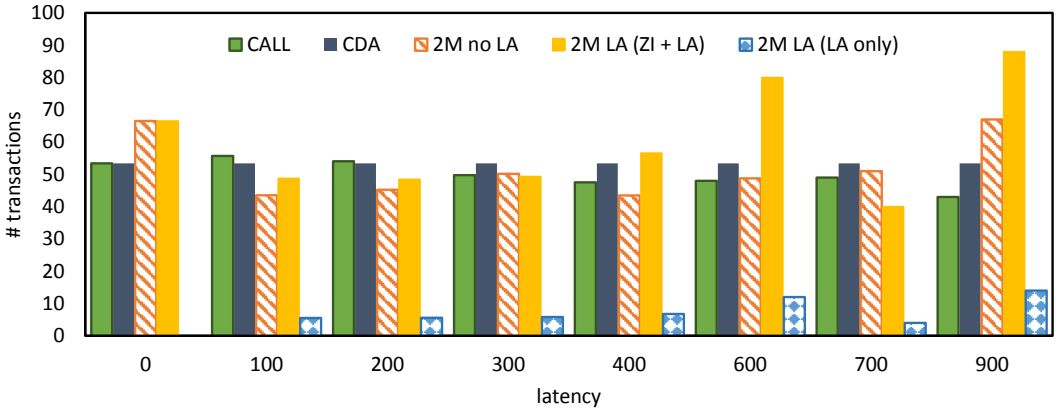
(c) Environment 2 median spread

(d) Environment 2 median NBBO

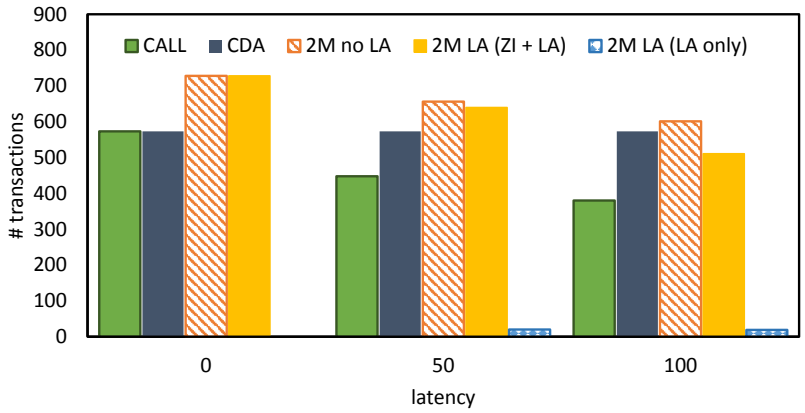
(e) Environment 3 median spread

(f) Environment 3 median NBBO

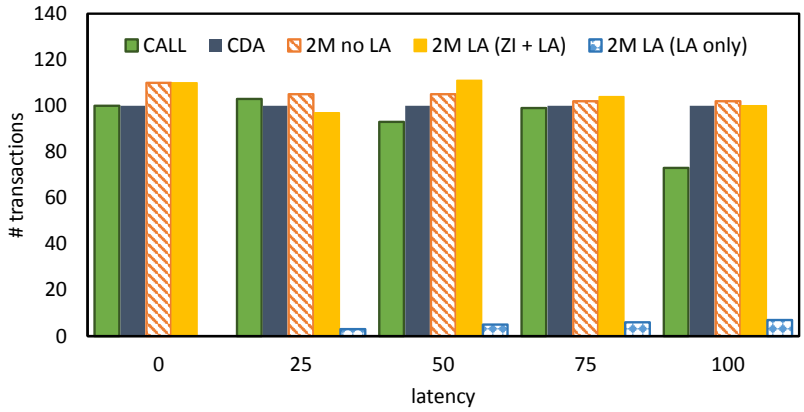
Figure 5.9: Median spread and NBBO spread for the centralized call market and the two-market (2M) model with LA, for the three environments. Each point reflects the average over 50,000 simulation runs of the maximum-welfare equilibrium for each market configuration and latency setting.



(a) Environment 1



(b) Environment 2



(c) Environment 3

Figure 5.10: Total number of transactions in each of the four market configurations, as well as the number of LA transactions in the two-market model with LA, for the three environments. Each bar reflects the average over 50,000 simulation runs of the maximum-welfare equilibrium for each market configuration and latency setting.

CHAPTER VI

Strategic Market Choice: Frequent Call Markets versus Continuous Double Auctions for Fast and Slow Traders

Frequent call markets have been proposed as a market design solution to the latency arms race perpetuated by high-frequency traders in continuous markets, but the path to widespread adoption of such markets is unclear. If such trading mechanisms were available, would anyone want to use them? This is ultimately a question of market choice, thus I model it as a game of strategic market selection. My market environment is populated by fast and slow traders who choose to trade in either a frequent call market or a continuous double auction. I employ the empirical game-theoretic methods presented in Section 3.2 to determine the market type selected in equilibrium. I also analyze best-response patterns to characterize the frequent call market's basin of attraction. My findings show that in equilibrium, welfare of slow traders is generally higher in the call market. I also find strong evidence of a predator-prey relation between fast and slow traders: the fast traders prefer to be with slower agents, and slow traders seek the protection of the frequent call market. The results in this chapter were presented at the *3rd EAI Conference on Auctions, Market Mechanisms, and their Applications* (Wah et al., 2015).

6.1 Introduction

Incremental speed yields advantage in trading due to the continuous nature of market mechanisms. Currently, most stock markets operate as continuous double auctions or CDAs (Friedman, 1993). Recall that in a CDA, orders are matched strictly on a first-come basis. This time priority rule induces a winner-take-all scenario, where the fastest trader can readily expropriate all gains from new information. The speed differential between high-frequency traders and slower, non-HF investors subjects the latter to *adverse selection*, in which the slower traders' resting orders are more likely to trade when information moves against them.

An alternative to the continuous double auction is a frequent call market (or frequent batch auction) described in Section 2.1. Recall that in a frequent call market, orders are matched at discrete, regular intervals, and there is no time-priority within each clearing interval. Each interval is a sealed-bid auction: participants do not know what orders other traders have submitted, ergo orders in the frequent call market cannot be targeted specifically by incoming informed orders. Even if a fast trader knew somehow about a stale order sitting in the book, it could not exploit that completely because the prices are set via a competition among all traders able to submit orders within the clearing interval.

Allowing orders to accumulate over short time periods in a frequent call market has already been advocated as a means to impede harmful HFT strategies that exploit other traders through speed (Wellman, 2009; Sparrow, 2012; Schwartz and Wu, 2013). Not only do frequent call markets offer significant gains in social welfare over CDA markets by aggregating multiple orders and matching at a uniform price (Wah and Wellman, 2013), they effectively eliminate the advantage of almost imperceptible improvements in latency by shrinking the window of speed advantage to a tiny fraction of each clearing interval (Budish et al., 2015). Since the order book is not visible—each clearing interval is sealed-bid—and the best prices in the frequent call market are only available after orders have been matched and cleared, traders in a frequent call market are incentivized to compete not

on time but rather on price.

In recent years, frequent call markets have steadily gained traction as a potential mechanism design solution to the exigencies of today's financial markets. Budish et al. (2015) show how correlations of related securities break down at small time scales, opening up opportunities for arbitrage based on tiny speed advantages. They present a model with competitive HFTs in a call auction, and show how the frequent batch design neutralizes such speed advantages. Farmer and Skouras (2012) likewise advocate frequent sealed-bid auctions as a means to end the technological arms race, suggesting that clear times be randomized. McPartland (2013) proposes matching orders every half-second and switching to a cardinal time-weighted pro rata trade allocation formula to eliminate the advantage of speed in tie-breaking. This author also recommends randomization of the trade match algorithm, that is, matching orders to trade at a random time within each fixed-length clearing interval. Other variants of randomized frequent call markets to deter HFT sniping have been suggested by Sellberg (2010) and Industry Super Network (2013).

Regulators are starting to take notice. For instance, U.S. Securities and Exchange Commission Chair Mary Jo White has indicated receptiveness towards "flexible competitive solutions…[which] could include frequent batch auctions or other mechanisms designed to minimize speed advantages" (White, 2014). New York Attorney General Eric Schneiderman endorsed frequent batch auctions in remarks during a March 2014 New York Law School panel on Insider Trading 2.0 (Schneiderman, 2014):

> *Currently, on our exchanges, securities are traded continuously, which means that orders are constantly accepted and matched with ties broken based on which orders arrived first. This system rewards high-frequency traders who continuously flood the market with orders, emphasizing speed over price.… If you had frequent batch auctions, there's no point in trying to get faster than whatever the interval is. It would discourage the risk taking that can cause flash crashes because, in the quest for greater and greater speed, there is, in*

and of itself, a threat to market stability.

Skeptics of frequent call markets raise various objections to their feasibility. Some doubt whether continuous- and discrete-time markets can coexist, and posit that it will be necessary to ensure that fragmented call markets clear in a synchronized manner (Rosov, 2014). Others question the ease of implementing these call markets (Baldauf and Mollner, 2015). Featherstone (2014) argues that frequent batch auctions are an unattractive alternative to current continuous markets, contending that discrete-time markets will diminish trading and adversely affect price stability, while simultaneously creating the incentive to snipe within the clearing interval in the event of new information arriving before the clear. Similarly, Ross (2014) surmises that the introduction of frequent batch auctions would engender a race to place the first order in the book for each call auction.

These and most other arguments I have encountered appear to be based on misconceptions or unfounded speculation. The call market does not need to give time priority for orders within the clearing interval, and so it is easy to avoid races to submit orders and to instead channel competition to the price dimension. Traders submitting the best price, whether fast or slow, will execute, and orders clear at a uniform price that no market participant knows in advance, making ties in price unlikely anyway (especially if prices are fine-grained). For the same reason, synchronization of multiple frequent call markets is unnecessary, given that the gain from a speed advantage is already reduced by the lack of visibility into the order book. In addition, implementation of frequent call markets is clearly feasible; many modern stock markets open and close trading each day with a call auction (Madhavan, 1992; Vives, 2010).

Yet frequent call markets have hitherto not been widely adopted. This may be simply a matter of inertia; as markets have evolved from in-person to electronic, imposing an explicit time delay would take a deliberate intervention. Such time delays are intuitively retrograde to many, as they seem to compromise the general investor demand for trading immediacy (Economides and Schwartz, 1995).

This explains why existing continuous markets might not change their policies, but what about introducing new markets with the frequent call mechanism? I see no economic reason why a discrete-time market could not coexist alongside continuous market mechanisms (Wah et al., 2013), but admittedly the burden of demonstration may rest on those of us arguing for feasibility. To provide such a demonstration, I consider the question of market choice: given availability of both mechanisms, will traders elect to submit orders to a frequent call market over a continuous market, and if so, under what conditions?

These are the questions addressed by my study. I formulate the frequent call market vs. CDA scenario as a game of market choice in which fast and slow traders—who differ on the frequency with which they arrive to trade—specify, as part of their strategy, a selected market mechanism. This strategic market choice game is described in the following section. I discuss my experiments in Section 6.3 and my results in Section 6.4. I survey additional related work in Section 6.5. Section 6.6 offers my conclusions.

6.2 Strategic Market Choice

To determine whether a frequent call market operating alongside a continuous market can successfully attract investors, I present a market choice game in which traders specify the preferred trading mechanism as part of their strategy. The players in my game are traders, grouped in two roles: FAST and SLOW. These roles differ only in the frequency with which traders enter to submit an order.

In my model, there is a single security traded simultaneously in a continuous double auction market (CDA) and a frequent call market (CALL). The environment is populated by multiple trading agents, representing investors. The trader valuation model and the class of strategies employed are described in detail in Section 2.2. Traders can elect to submit to either the frequent call market or the continuous market. Resting orders in the book are subject to adverse selection, since newly arriving traders have more current information about the fundamental, which they can exploit to pick off stale orders. Since SLOW traders

arrive less frequently into the market than their FAST counterparts, SLOW-agent orders are on average based on older information, and thus are exposed to a greater degree of adverse selection.

I select a fixed, deterministic rate of clearing for the frequent call market in my market choice game. Though several have proposed randomizing the clearing interval to deter sniping (Farmer and Skouras, 2012; Industry Super Network, 2013; Sellberg, 2010), I have argued (Wah et al., 2013) that such randomization accomplishes no reduction in incentive for HFT speed advantages. A deterministic clear time offers the prospect of sniping within a small time window at the end of the clear interval, whereas a random clear time offers a small probabilistic prospect for advantage over the entire interval. In expectation, the value of this advantage is the same. Moreover, as the model in this paper does not include strategic timing, sniping is effectively ruled out by assumption.

In a market choice game with players who strategically decide among market mechanisms, there trivially exist equilibria in which all traders select any one given market, regardless of its merits. These equilibria arise because when all *other* agents are in that market, the remaining trader has no possibility to trade anywhere else. Thus the trader's only option for positive payoff is to join the focal market. To render these equilibria non-inevitable, I introduce to each market a set of *environment agents*, providing a base set of available trading partners. The environment traders follow designated strategies for their assigned market and are not considered players in my game model. As such, their behavior plays no part in game-theoretic analysis and their trading gains are ignored in surplus calculations. I denote the number of environment agents in each market by E.

I employ an empirical simulation-based approach (presented in Chapter III) to explore the strategy space. This facilitates identification of the market conditions under which traders may prefer one market mechanism over the other. From the empirical game induced over thousands of simulations of selected strategy profiles, I determine the market chosen in equilibrium, and I analyze the corresponding gains from trade. I characterize the frequent

call market's basin of attraction through analysis of trader best responses that specify the frequent call market over the CDA.

My findings show that in equilibrium, welfare of SLOW traders is generally higher in the frequent call market than in the continuous double auction. I also find strong evidence of a predator-prey interaction between FAST and SLOW traders. The FAST traders prefer to be in the same market as their prey, whereas the SLOW traders congregate in the frequent call market as long as it is sufficiently thick.

6.3 Experiments

6.3.1 Environment Settings

I evaluate the performance of traders in four environments. Recall that agents arrive according to a Poisson process, and on each arrival they submit a single-unit limit order to their associated market—replacing any prior outstanding order. Reentry rates are fixed across the environments, with FAST traders arriving in the market at rate $\lambda_F = 0.004$, and SLOW traders entering at rate $\lambda_S = 0.002$. In all settings, there is one CDA and one frequent call market, which clears every 100 time steps. Each simulation run lasts $T = 12000$ time steps. The mean-reverting global fundamental has a mean value $\bar{r} = 10^5$. The variance for the private value vector is $\sigma^2_{PV} = 5 \times 10^6$. The fundamental shock variance is $\sigma^2_s = 1 \times 10^6$.

The strategy of environment agents is fixed; they play a ZI strategy with range $[0, 1000]$ and $\eta = 1$. The environment agents enter their respective markets with rate $\lambda_E = 0.005$. The environments differ in the value of the mean-reversion parameter (κ) and the number of environment agents $E \in \{8, 14, 42\}$. The configurations are as follows:

Environment I $E = 8$, $\kappa = 0.05$

Environment II $E = 8$, $\kappa = 0.01$

Environment III $E = 14$, $\kappa = 0.01$

Environment IV $E = 42$, $\kappa = 0.01$

The empirical games for these environments include 12 strategies (Table 6.1) for traders, 6 in each market. The market choice decision is made before trading commences at time 0, and once selected, the market for a given agent is fixed for the duration of the trading horizon T. Player agents choose between CDA and CALL, and environment agents are each assigned to one of these.

6.3.2 EGTA Process

In the market choice game, players are partitioned into roles $R = \{\text{FAST}, \text{SLOW}\}$, and players in either role can select among a set of strategies S. I determine equilibria for my game of strategic market choice through empirical game-theoretic analysis, described in Section 3.2. I collect data for multiple combinations of the trader strategies: a minimum of 5,000 samples per profile evaluated, with 20,000 samples for most profiles and averaging at least 10,739 samples per profile in each environment. From these payoff estimates, I compute RSNE for each environment, and use these as a foundation for my analysis of the welfare effects in equilibrium for FAST versus SLOW traders, the attractiveness of the CALL over the CDA, and the loss in deviating from the equilibrium market type.

I apply deviation-preserving reduction (Section 3.2.2) to reduce an $(N_{\text{FAST}}, N_{\text{SLOW}})$-player game to a $(k_{\text{FAST}}, k_{\text{SLOW}})$-player reduced game. I deliberately select values for N_r and k_r, $r \in \{\text{FAST}, \text{SLOW}\}$, to ensure that the fractions defining the game reduction come out as integers. Specifically, my market choice game is comprised of 42 players, with $N_{\text{FAST}} = N_{\text{SLOW}} = 21$, which I approximate by a DPR game with $k_{\text{FAST}} = k_{\text{SLOW}} = 3$. I use simulation data from the full $(21, 21)$-player game to estimate the payoffs of the $(3, 3)$-player reduced game.

Table 6.1: ZI strategy combinations included in empirical game-theoretic analysis of market choice games.

$R_{\min}$	$R_{\max}$	η	Market type
0	125	1	Both
0	250	1	Both
0	500	1	Both
0	1000	1	Both
500	1000	0.4	CDA
500	1000	1	CALL
0	2500	1	Both

6.3.3 Social Optimum

I assess efficiency by comparison of market outcomes with the social optimum. I define this optimum for a population of 42 traders, based on the distribution of the private component of agents' valuations, with parameters $q_{\max} = 10$ and $\sigma^2_{PV} = 5 \times 10^6$. To calculate an optimal allocation for a particular array of draws from this distribution, I simply find the competitive equilibrium using the call market clearing function. Each trader submits its valuation vector as a demand curve, with $q_{\max}$ sell orders at prices $\bar{r} + \theta^s_i$, $s \in \{-q_{\max} + 1, \ldots, 0\}$ and $q_{\max}$ buy orders at prices $\bar{r} + \theta^b_i$, $b \in \{+1, \ldots, q_{\max}\}$, with each order for a single unit of the security. Over 20,000 samples, I find a mean social welfare of 27887. As they are not considered players in the market choice game, I do not include environment agents in the determination of the socially optimal allocation. Figure 6.1 shows the histogram of trades per player in the social optimum.

6.4 Results

6.4.1 Basin of Attraction

The main results of this study are shown in the heat maps of Figure 6.2, which illustrate the trader population conditions under which the CALL serves as an attractor. I characterize the frequent call market's basin of attraction by categorizing the market type a trader selects

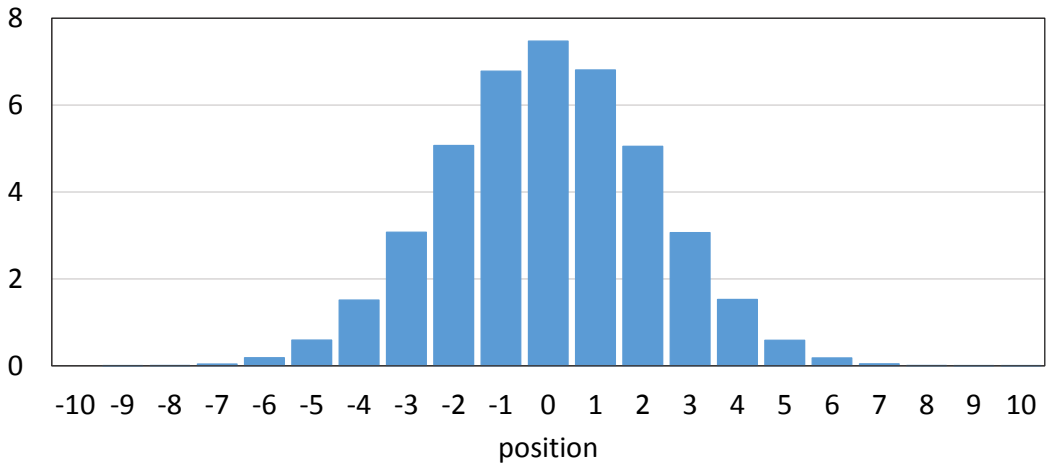


Figure 6.1: Histogram of the net position (i.e., number of units traded) of traders in the socially optimal allocation. The distribution is compiled from 20,000 samples.

when the other traders' strategies are fixed; in other words, I identify and classify the trading mechanism selected in the trader's best response.

Given a trader, I fix the other-agent profile (i.e., the set of strategy counts for the 20 players in the same role and for the 21 players in the other role), and I identify the market type selected in the trader's best response. Since I selectively sample full-game profiles for the $(3, 3)$-player DPR approximation, I can bucket all other-agent profiles into 12 unique categories by role, based on the population of traders in each market type.

For example, if I examine a SLOW trader in the CALL, the 20 other SLOW traders may all be in the same market (CDA or CALL) or they may be equally split between the two markets (10 in the CALL and 10 in the CDA). No other cross-market divisions of same-role players are possible because I selectively collect profiles to reduce via DPR to a $(3, 3)$-player game. The 21 traders in the other role (FAST) may all be in the same market (CDA or CALL), or they may be split between the two markets, with 7 agents in one market and 14 in the other.

For each of the 12 categories of trader population distributions across markets, I count the number of other-agent profiles for which the given player's best response specifies the CALL over the CDA. I report the corresponding percentages in two best-response heat

maps, one per role, for each market choice game.

To ensure full coverage of all population categories, I construct a complete subgame. In each subgame I include the strategies played with the highest probabilities across all RSNE found in that environment. The strategy sets of the complete subgames used to characterize the frequent call market's basin of attraction for each environment are given in Table 6.2.

My results for these subgames are illustrated in the heat maps of Figure 6.2; these characterize the frequent call market's basin of attraction from the perspective of a single trader in each role (SLOW on the left, FAST on the right). For example, the top left entry in a SLOW trader heat map reports the percentage of all other-agent profiles comprised of 20 SLOW traders in the CDA and 21 FAST traders also in the CDA in which a SLOW trader's best response specifies the CALL.

The higher mean reversion in environment I implies that slower traders are less likely to be picked off by speed-advantaged traders, and therefore I find that the SLOW traders display no strong preference to switch to the CALL unless the majority of other traders (regardless of speed) are in the frequent call market as well. When the degree of mean reversion is reduced, the SLOW agents face greater risk of being picked off by FAST agents with newer and better information. Therefore, environments II through IV are much more salient in answering questions about strategic market choice under adverse selection, and I focus the rest of this section on those corresponding subgames.

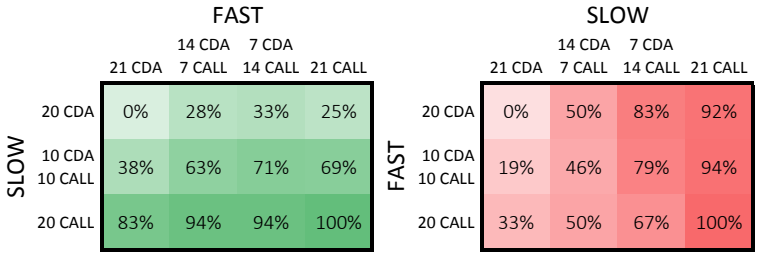
I see from the environment II–IV heat maps that there is safety in numbers for a single SLOW trader deciding between the CALL and CDA: if 20 of the SLOW traders are in a given market, the best response is more often than not to pick the same market as everyone else, whether that is the CDA or CALL. When the SLOW agent population is equally divided between the two markets, however, I observe a gradual mass exodus of SLOW traders from the frequent call market as more FAST traders enter the CALL. The percentage of SLOW-trader best responses selecting the CALL decreases monotonically from around 90% to below 40% as FAST traders leave the CDA for the frequent call market. Despite the

protection afforded to them in the CALL, the SLOW traders would rather take their chances in the CDA than remain in the same market as the FAST traders. However, if the CALL is sufficiently thick (as in environment IV), the SLOW traders prefer the sanctuary of the frequent call market, regardless of where the FAST traders are.

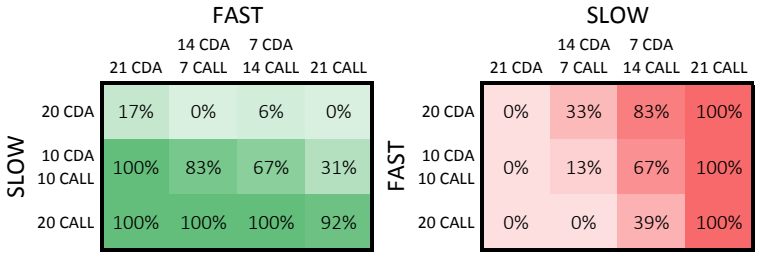
On the other hand, FAST traders clearly stand to gain from the informationally disadvantaged orders submitted by their slower counterparts. Therefore, they exhibit a strong preference for the market selected by the majority of SLOW traders, and they readily follow the SLOW traders to either market. I observe that their preference for the CALL increases strictly monotonically, from 0% to nearly 100%, as the number of SLOW traders in the CALL increases.

These results reveal the dynamics of the predator-prey interaction between the FAST and SLOW traders. The SLOW traders face less risk as part of a large group, but once they are split up between the two markets, those in the CDA tend to flee to the CALL to get away from the FAST traders, while the FAST traders relentlessly pursue the SLOW traders, regardless of market.

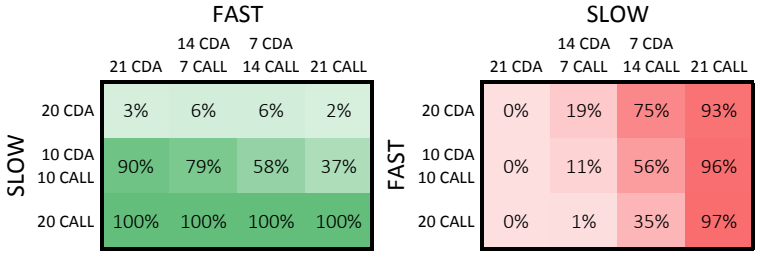
I also analyze the collected profiles in the full games, shown in Figure 6.3. The heat maps for the SLOW traders are similar to those in the complete subgames, but the results for FAST traders are markedly different. This is due to the bias in sampling full-game profiles for my game-theoretic analysis. As sampling all 681,264 profiles in the full game (given two roles, with 12 strategies each) is intractable, my coverage of the profile space is primarily determined by the more promising subgames identified during EGTA.



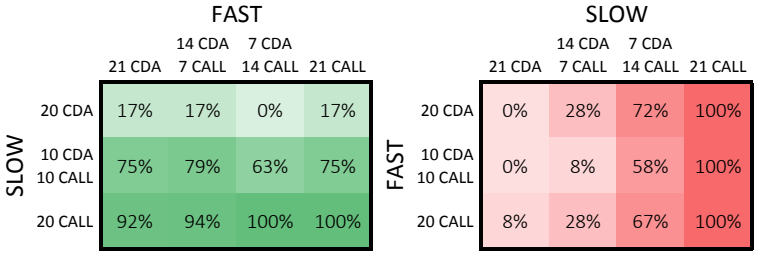
(a) Environment I



(b) Environment II

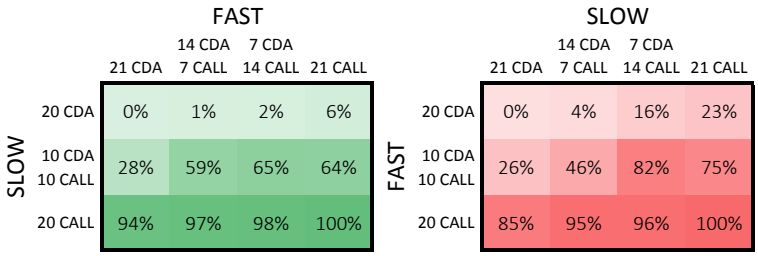


(c) Environment III

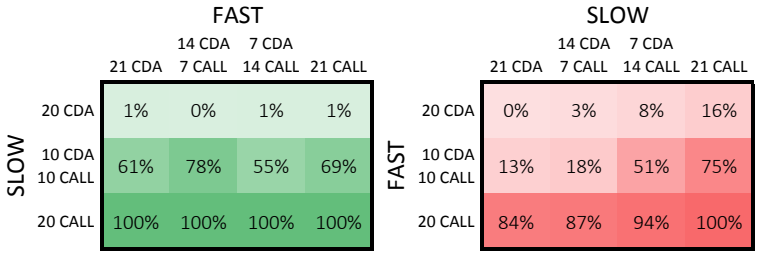


(d) Environment IV

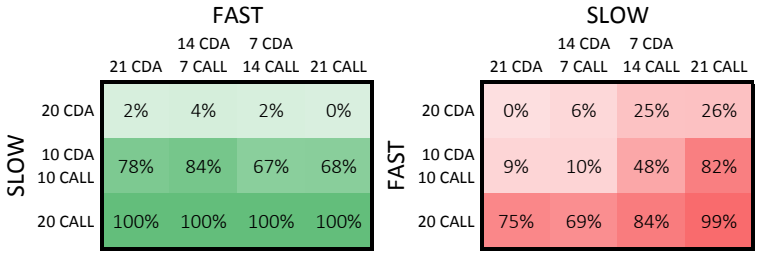
Figure 6.2: Basin of attraction for CALL, as characterized by best-response heat maps of complete subgames for environments I–IV. The subgame for environment III has a 6×6 strategy space, with three strategies in each market, for each role; the subgames for the other environments have strategy spaces of size 4×4. The matrices on the left (in green) are from the perspective of a single SLOW trader; the matrices on the right (in red) are from that of a FAST trader. The rows in each matrix specify the distribution of same-role agents across the two markets, and the columns specify the cross-market distribution of other-role agents. Each entry in the heat map matrix gives the percentage of all other-agent profiles in which a single agent's best response specifies CALL. Heat map colors follow a scale where light corresponds to 0% and dark to 100%.



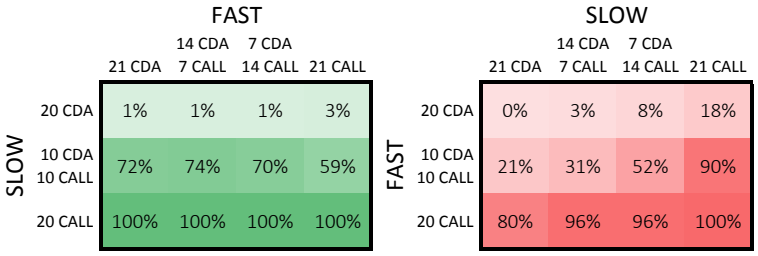
(a) Environment I



(b) Environment II



(c) Environment III



(d) Environment IV

Figure 6.3: Basin of attraction for CALL, as characterized by best-response heat maps of sampled full-game profiles for environments I–IV. Data presented is as for Figure 6.2.

Table 6.2: Complete subgames, one each for environments I–IV, used to analyze the frequent call market's basin of attraction. Each row of the table describes a complete subgame, which includes the strategies played with the highest probabilities across all RSNE found in that environment. The numeric column headings give $R_{\max}$ values for the ZI strategies, for each role. All strategies employ $R_{\min} = 0$, with the exception of the double starred and double dagger (‡) values which use $R_{\min} = \frac{1}{2}R_{\max} = 500$. All strategies employ $\eta = 1$, except for the double dagger (‡) values which use $\eta = 0.4$. Strategies in the table cells are specified according to the market type. The subgame for Environment III has a 6×6 strategy space, with three strategies per market, per role; the subgames for the three other environments each have a 4×4 strategy space.

Env	FAST							SLOW						
	125	250	500	1000	1000‡	1000**	2500	125	250	500	1000	1000‡	1000**	2500
I		CALL	Both		CDA					Both	CALL	CDA		
II	CALL				CDA		Both	CALL				CDA		Both
III	CALL			Both	CDA		Both	CALL			Both	CDA		Both
IV	Both						Both	Both						Both

6.4.2 Equilibrium Analysis

My equilibrium results are shown in Table 6.3 (see Table C.1 for complete specifications of the equilibria found). For each RSNE, I compute surplus for traders in each role by sampling 10,000 full-game profiles based on the equilibrium mixture probabilities, with one simulation run per sampled profile. I successfully find at least one and up to six RSNE in each environment; each equilibrium has one to three strategies played with positive probability for a given role. There is at least one all-CALL RSNE in each environment; all but one environment has at least one all-CDA equilibrium.

I find empirical support for the general welfare benefits of the CALL market, but primarily for SLOW traders: the mean total SLOW-agent surplus accrued over the all-CALL equilibria in a given environment is uniformly higher than that over the all-CDA equilibria in the same environment. Environments I and II have the same environment-agent population, but the lower mean reversion in the latter makes SLOW traders more susceptible to adverse selection. This is reflected in the significant reduction in total SLOW-agent surplus in environment II versus environment I. FAST traders accrue approximately the same level of surplus in both environments. I also observe that although the total welfare in environment I is close to the social optimum described in Section 6.3.3, increased adverse selection reduces overall surplus, and it is in this setting that the CALL provides significant welfare improvement over the CDA.

Within the same environment and with reduced mean reversion, FAST traders also generally shade their bids less in the frequent call market versus the CDA, as can be evidenced by reduced R_{mid} values. This effect does not hold for the SLOW traders, who shade approximately the same regardless of market type. The reduction in FAST-trader bid shading is indicative of the shift from a competition on speed in the CDA to a competition on price in the CALL.

I find at least one all-CDA RSNE in environments I through III. This is due to the low number of environment agents in these games. When there are only 4 environment agents in

each market, as in environments I and II, the CALL market is not thick enough—sufficient volume is required for the call auction to deliver on its promise of welfare improvement. But in environment IV, where $E = 42$, there is ample volume and order activity in the CALL market for traders to strongly prefer it over the CDA, hence I find no all-CDA RSNE in this environment.

Notably, I only find RSNE in which both FAST and SLOW agents select the same market. I can definitively rule out two-market equilibria—in which all agents in one role choose the CALL and all those in the other role choose the CDA—by exploiting an independence property of market choice games. In these games, the payoff for any given strategy depends only on the strategies of traders in the same market. I identify CALL-CDA equilibrium candidates by exploring four subgames for each environment. In each of these subgames, I limit the 21 traders in one role to a single strategy in the first market, while permitting traders in the other role to select any of the six strategies specifying the other market. In essence, I limit these subgames to one market and one role. I compute the equilibria in each of these subgames and form equilibrium candidates of the target form; I can then confirm or refute these candidates within the full strategy space.

For example, I explore a subgame with SLOW traders playing CALL strategies from Table 6.1 and FAST traders playing some strategy s_{CDA} in the CDA. I also explore a subgame with FAST agents playing CDA strategies and SLOW agents playing a fixed strategy s_{CALL} in the CALL market. Analysis of the first (second) subgame gives the equilibria for FAST (SLOW) traders in the CDA assuming no SLOW (FAST) traders are present. I can then form a CALL-CDA equilibrium candidate from any equilibrium in the first subgame (which specifies the strategies for FAST traders in the CDA) and any equilibrium in the second subgame (which specifies the strategies for SLOW agents in the CALL).

I refute all such candidate equilibria in all four environments, hence there are no RSNE in which all FAST traders are in the CDA and all SLOW traders are in the CALL. Such equilibria might be expected given that FAST traders benefit from picking off stale orders

Table 6.3: Role-symmetric equilibria for the four strategic market choice games (one each for environments I–IV), calculated from the $(3, 3)$-player DPR approximation. Each row of the table describes one equilibrium found, including, for each role in the RSNE, the selected market mechanism (CALL or CDA) and the average values for total surplus of players in the role and for two strategy parameters: R_{mid} (the midpoint of ZI range $[R_{\min}, R_{\max}]$) and threshold η. Values presented are averages over strategies in the profile, weighted by mixture probabilities. There is at least one all-CALL RSNE in each environment and one all-CDA RSNE in environments I to III, but I did not find any all-CDA equilibria in environment IV.

Env	Total	FAST				SLOW			
		Market	Surplus	R_{mid}	η	Market	Surplus	R_{mid}	η
I	27288	CALL	14469	129	1	CALL	12819	230	1
I	26697	CALL	14384	210	1	CALL	12314	486	1
I	27261	CDA	14598	250	1	CDA	12662	198	1
I	26785	CDA	14136	435	0.769	CDA	12649	250	1
I	25321	CDA	13502	418	0.943	CDA	11819	750	0.4
I	26133	CDA	13969	559	0.630	CDA	12165	500	1
II	21050	CALL	14697	703	1	CALL	6353	1250	1
II	21242	CDA	15355	710	0.448	CDA	5887	1250	1
III	19992	CALL	13790	644	1	CALL	6202	1250	1
III	20441	CALL	13909	500	1	CALL	6532	1111	1
III	19734	CDA	14483	750	0.4	CDA	5251	1250	1
IV	18067	CALL	12856	970	1	CALL	5211	1250	1

in the CDA. Ultimately, I find no such RSNE because the FAST traders benefit from being in the same market as the SLOW traders. The SLOW traders face greater risk of adverse selection in the CDA, so they select the frequent call market, followed close behind by the FAST traders.

6.4.3 Regret Analysis

I also evaluate the degree to which a trader is attracted to the CALL versus the CDA. To that end, I compute NE regret (Jordan et al., 2010), which captures the loss of utility for a player who deviates from a Nash equilibrium to a specified strategy. The NE regret of a given strategy s is defined as the utility to the player in equilibrium less the payoff it

accrues when it deviates to s. Accordingly, the NE regret of any equilibrium strategy is zero.

To compute the NE regret of deviating to the other market, I use the sampled surplus values in Table 6.3 to determine the per-agent surplus for a trader in a given role, and I subtract from that the payoff of the best-performing strategy in the other market. Again, I can exploit the independence of the two markets in my model, this time to determine the best other-market strategy. For example, for an all-CALL RSNE, I can measure the payoff to an agent that deviates to strategy s_{CDA} in the CDA via the payoff in any profile in which a single trader plays s_{CDA} and the other traders are in the CALL. I average the payoffs accumulated across all such profiles to determine the maximum-payoff other-market strategy for each RSNE, and I use these to compute the minimum NE regret for deviating to the non-RSNE market.

My results are shown in Figure 6.4. SLOW traders generally have lower regret if deviating to the CALL from an all-CDA RSNE than if deviating to a CDA from an all-CALL RSNE. This is indicative of the greater loss they face if they leave the CALL market, as they are at high risk of being picked off by the faster traders in the CDA. The FAST traders, on the other hand, stand to lose more if they deviate from an all-CDA RSNE to the CALL, versus deviating to the CDA from an all-CALL RSNE, because their payoffs are based on exploiting their speed advantage over the SLOW traders. In short, SLOW traders would much rather stay in the CALL market, while FAST traders exhibit a stronger preference for the continuous market. I observe that FAST traders have universally greater regret than the SLOW traders; this is because they already accrue the lion's share of overall welfare, hence they have greater profits to lose. The negative regrets in my results are indicative of the limitations of the DPR approximations I use in deriving equilibria.

Also notable is that the best strategy when deviating to the CDA from an all-CALL RSNE is always the one strategy in which the threshold $\eta < 1$, regardless of environment or trader speed. Because environment agents arrive even more frequently than FAST traders,

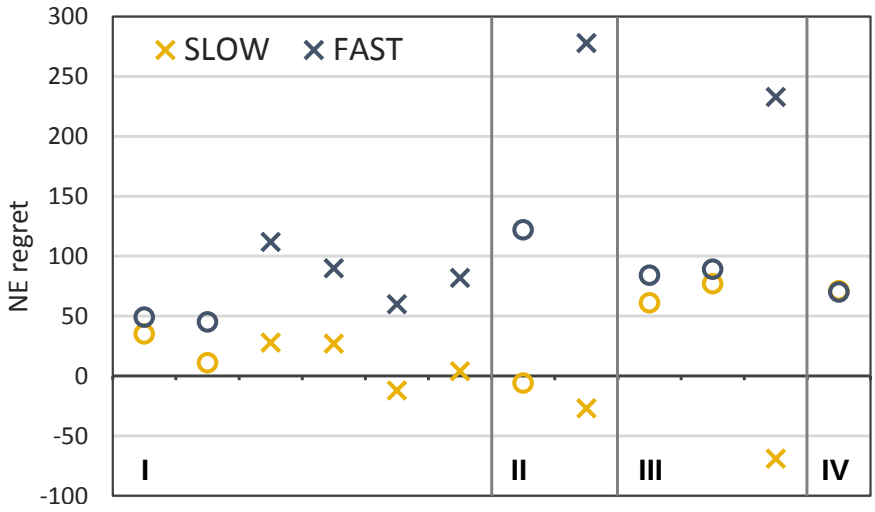


Figure 6.4: NE regret of equilibria in environments I–IV, computed for each RSNE as the per-agent surplus in a role, less the maximum payoff possible if a player in that role deviates to the other market. Os indicate the NE regret in deviating to the CDA from an all-CALL RSNE; Xs indicate the NE regret in deviating to the CALL from an all-CDA RSNE. Note that the FAST and SLOW trader NE regrets in environment IV are overlaid as they are nearly identical.

any player faces significant adverse selection if alone with the environment agents in the CDA market. Adopting a lower η decreases the tendency to leave standing orders, thus avoiding some of the pick-off risk.

6.4.4 Game without Mean Reversion

As discussed in Section 6.4.1, the reduced mean reversion in environments II through IV increases the SLOW traders' risk of adverse selection. I introduce an additional three environments to explore the attractiveness of the CALL market when there is zero mean reversion in the fundamental:

Environment V $E = 8, \kappa = 0$

Environment VI $E = 14, \kappa = 0$

Environment VII $E = 42, \kappa = 0$

In order to for traders to accrue positive gains in my simulations given no mean reversion,

I reduce the variance in the fundamental to $\sigma_s^2 = 1 \times 10^3$. This also means that the specific payoffs of traders in these environments cannot be compared directly to results in the previous sections.

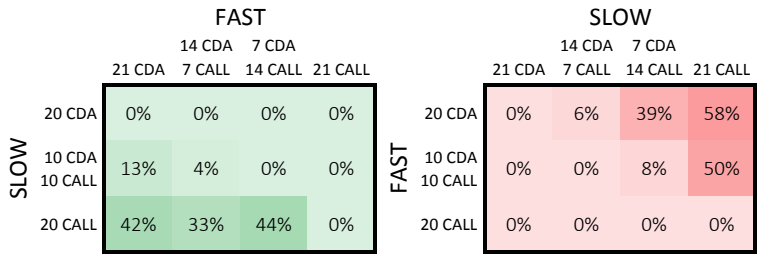
As with the first four environments, I construct complete subgames; the strategy sets of the complete subgames for environments V–VII are given in Table 6.4. Figure 6.6 shows the heat maps for these subgames, and the results for the sampled profiles from the full games are shown in Figure 6.5.

In the settings without mean reversion, I observe qualitatively different results from the previous four environments, particularly for the cases with more environment agents. Both FAST and SLOW traders exhibit a much weaker preference for the CALL market, and this preference only exists under certain market conditions. This is also reflected in the equilibrium results for environments V–VII (Table 6.5): I find one all-CDA RSNE in each environment, in addition to an equilibrium with FAST traders split between the CALL and CDA and all SLOW traders in the CALL (see Table C.2 for complete specifications of the equilibria found).

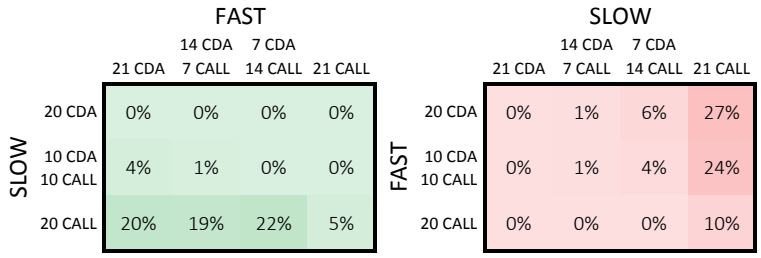
The essence of the predator-prey relationship observed in environments I–IV still remains, most visibly in environment V. A SLOW trader has a greater preference for the CALL when the 20 other SLOW traders are also in the CALL and all 21 FAST traders are in the CDA. Similarly, a FAST trader tends to prefer the CALL when all 21 SLOW traders are also in the CALL. However, this effect diminishes as the number of environment agents E increases from environment V to VII.

Traders in these environments come quite close (or surpass) the social optimum of 27887. As described in Section 6.3.3, this social optimum is based on the 42 traders submitting orders corresponding to their demand curves. Traders can, in aggregate, exceed the social optimum by extracting additional surplus from the environment agents in the CDA. They do so by exploiting the fixed strategy of the environment agents. Despite being faster than the strategic traders, environment agents employ a strategy with threshold parameter

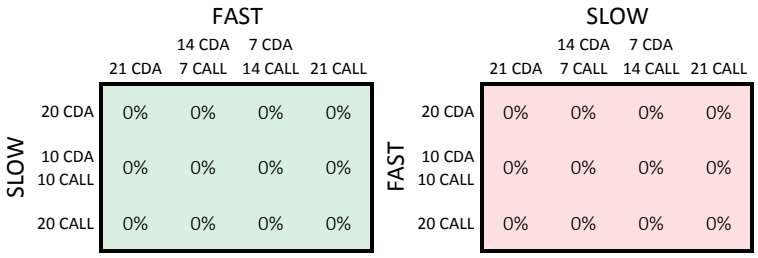
$\eta = 1$, which means these agents generally submit standing orders (i.e., orders that rest in the order book and do not immediately execute) rather than executable orders. There exists a pure-strategy RSNE in each zero mean reversion environment where both FAST and SLOW traders select the only strategy with $\eta < 1$. This reflects an increased number of opportunities—when mean reversion is eliminated—in which strategic traders can take the current price quote over submitting a limit order, so both FAST and SLOW traders can pick off standing orders submitted by the environment agents in these zero mean reversion environments. The number of such opportunities grows with the number of environment agents E, as their fixed strategy ensures they generally submit standing orders. The opportunity to exploit environment-agent orders exists to a much greater degree in the CDA, since any resting order can be readily picked off by submitting an order that will immediately match and trade with the resting order in question. This drastically reduces the incentive of strategic traders to switch to the CALL. In other words, when there is no mean reversion in my games of market choice, the environment agents become the prey in the CDA, and the strategic traders act as predators by picking off standing limit orders.



(a) Environment V

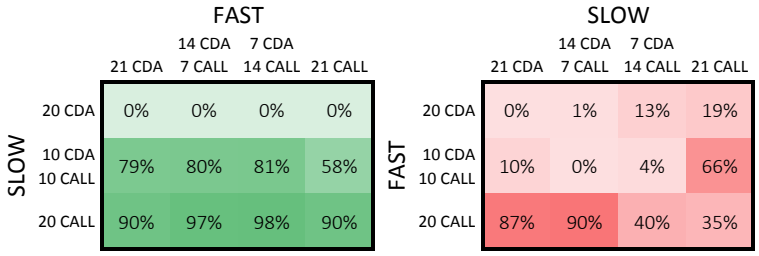


(b) Environment VI

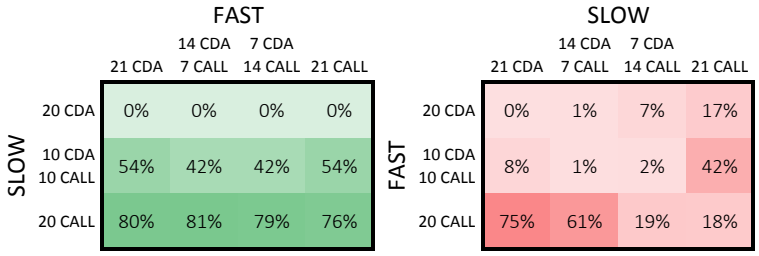


(c) Environment VII

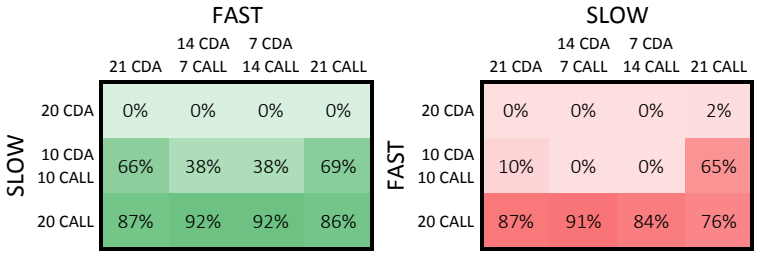
Figure 6.5: Basin of attraction for CALL, as characterized by best-response heat maps of complete subgames for environments V–VII. The subgame for environment VI has a 6×6 strategy space, with three strategies in each market, for each role; the subgames for the other environments have strategy spaces of size 4×4. Data presented is as for Figure 6.3.



(a) Environment V



(b) Environment VI



(c) Environment VII

Figure 6.6: Basin of attraction for CALL, as characterized by best-response heat maps of sampled full-game profiles for environments V–VII. Data presented is as for Figure 6.5.

Table 6.4: Complete subgames, one each for environments V–VII, used to analyze the frequent call market's basin of attraction with zero mean reversion. Data presented is as for Table 6.2. The subgame for Environment VI has a 6×6 strategy space, with three strategies per market, per role; the subgames for the two other environments each have a 4×4 strategy space.

Env	FAST							SLOW						
	125	250	500	1000	1000‡	1000**	2500	125	250	500	1000	1000‡	1000**	2500
V		CALL	Both		CDA					Both		CDA	CALL	
VI		CALL	CDA	CDA	CDA	CALL	CALL	CALL		CDA	CDA	CDA	CALL	CALL
VII		CALL	Both		CDA					CDA		CDA	CALL	CALL

Table 6.5: Role-symmetric equilibria for the three strategic market choice games without mean reversion (one each for environments V–VII), calculated from the $(3, 3)$-player DPR approximation. Data presented is as for Table 6.3. There is at least one all-CDA RSNE in each environment. Environment V has one equilibrium in which the FAST traders are in both the CALL and the CDA, and all SLOW traders are in the CALL.

Env	Total	FAST				SLOW			
		Market	Surplus	R_{mid}	η	Market	Surplus	R_{mid}	η
V	24038	Both	17698	345	0.849	CALL	6430	750	1
V	26457	CDA	18962	750	0.4	CDA	7495	750	0.4
VI	28681	CDA	20012	750	0.4	CDA	8669	750	0.4
VII	29412	CDA	20339	750	0.4	CDA	9073	750	0.4

6.5 Related Work

There are only a few isolated examples of call markets in today's financial markets, most of which clear on a semi-frequent basis. The Taiwan Stock Exchange matches orders by call auction, with clears occurring every 60 to 90 seconds, depending on trading activity (Lee et al., 2004). From mid-1998 to 2002, the Taiwan Futures Exchange employed a periodic call market to match orders; the clearing interval in the auction was incrementally reduced from 30 seconds to 20 and then 10, before finally being eliminated in favor of a predominantly continuous market mechanism (Webb et al., 2007). More recently, both the London Stock Exchange and the NYSE have announced plans to introduce a midday batch auction in hopes of encouraging institutional investors to trade large blocks of shares on their exchanges (Hope, 2015; Stafford, 2014). An intraday call auction has been standard for the past 15 years on Xetra, an electronic trading system for securities operated by Deutsche Börse (Budimir, 2014). Outside the equities space, batching to prevent exploitation by fast traders is currently in place on several foreign-exchange platforms. EBS, one of the largest currency trading platforms, has introduced a so-called latency floor, in which orders are batched in randomized clearing intervals (of lengths ranging from one to three milliseconds) in an effort to curb the advantages of super-fast traders (Clark, 2014). The

competing ParFX platform applies a randomized delay of 20 to 80 milliseconds to all order elements, and Thomson Reuters is currently trialling randomization of order execution on its foreign-exchange platform (Clark, 2014).

Several prior works, already presented in Section 5.2, have argued for frequent call markets as a means to end the latency arms race (Budish et al., 2015; Farmer and Skouras, 2012; McPartland, 2013; Sellberg, 2010; Industry Super Network, 2013). Recall that Budish et al. (2015) show that frequent batch auctions can potentially eliminate the latency arms race by reducing the value of very small speed advantages. Using millisecond-level exchange data, they demonstrate the breakdown of correlation between securities at high frequency, arguing that this phenomenon creates arbitrage opportunities that can be exploited by the fastest traders. In a complementary analysis, Budish et al. (2014) discuss the implementation details of frequent batch auctions in today's regulatory environment. Farmer and Skouras (2012) likewise advocate frequent sealed-bid auctions as a means to end the technological arms race, suggesting that clear times be randomized, whereas McPartland (2013) proposes matching orders every half-second and switching to a cardinal time-weighted pro rata trade allocation formula to eliminate the advantage of speed in tie-breaking.

In a recent study, Li and Das (2016) build an agent-based model with informed high-frequency and slow traders, as well as uninformed liquidity traders, to study the competition between a frequent call market and a CDA market. In their model, traders can pick their preferred market mechanism upon each arrival. By comparing welfare as measured by the price of immediacy, they find that traders are better off in the frequent call market, which also attracts the bulk of order flow.

Others have focused not on the role of call markets in mitigating the harmful effects of HFT, but on the difference in market quality offered in a discrete-time versus a continuous market. Pancs (2013) compares three models—a dark pool, a continuous market, and a periodic call auction—focusing on both allocative efficiency and informational efficiency (which is high when observed transactions reveal traders' private information). This study

finds that the periodic auction is more allocatively efficient than the continuous protocol when the demand for immediacy is low. Pellizzari and Dal Forno (2007) use an agent-based model to compare the efficiency of a call auction (clearing only once), a continuous double auction, and a dealership. They find that the dealer market is the most efficient market structure of the three, offering the lowest volatility and the highest perceived gains by traders.

Baldauf and Mollner (2015) develop a model of order anticipation to examine the impact of exchange competition on the spreads faced by investors. They study selective delay, an alternative trading mechanism in which cancellation orders are processed immediately but all other order types have a small delay, showing that selective delay reduces adverse selection by allowing liquidity providers to cancel stale quotes before being sniped by HFTs. In the specific setting of their work, they demonstrate that selective delay leads to the same outcome as a frequent batch auction. In another study, Baldauf and Mollner (2014) consider a setting in which selective delay and frequent batch auctions result in different equilibrium outcomes. They show that a frequent batch auction in this case results in wider spreads than both selective delay and a continuous market.

Another relevant question is the frequency of clearing in a periodic call market, with some prior work suggesting that more frequent trading leads to increased volatility (Lang and Lee, 1999; Webb et al., 2007). Fricke and Gerig (2015) argue that the optimal speed at which a security clears is related to volatility, trading intensity, and correlation of the security's value with other securities. They estimate that a range of 0.2 to 0.9 seconds is optimal. Du and Zhu (2014) study the effects of trading speed on overall welfare via a series of uniform-price double auctions held at discrete time intervals. They find that the optimal trading frequency varies depending on trader speed: fast traders prefer a higher trading frequency, whereas slow traders prefer a lower frequency (and consequently thicker) market.

Much of the empirical work in the call auction literature examines the effects of discrete-time trading through natural experiments. For example, Kalay et al. (2002) analyze the

move of stocks on the Tel Aviv Stock Exchange from discrete-time trading to continuous trading. They argue that investors prefer stocks that trade continuously, based on observed losses in volume in stocks that trade by call auction. Webb et al. (2007) examine the effect of the decision of TAIFEX, at the time a periodic call auction, to match the trading hours of the Singapore Exchange (SGX), a continuous market, finding that this switch led to a statistically significant reduction in volatility on the SGX. They attribute these results to better price formation in the discrete-time market.

6.6 Conclusions

I examined strategic market choice in four environments with both FAST and SLOW traders who must decide between two market mechanisms: a frequent call market and a continuous double auction. I modeled this interaction as a game of market selection. I employed empirical simulation methods to compare the market type selected in equilibrium, the trading gains accrued, and the regret of deviating from equilibrium. I also analyzed best-response patterns in order to characterize the frequent call market's basin of attraction in multiple environments.

This study offers the first analysis of adoption of frequent call markets, framed as a question of strategic market choice. My findings demonstrate that in equilibrium, SLOW-trader welfare is generally higher in the discrete-time market—further evidence that frequent call markets offer both increased gains from trade as well as protection from speed-advantaged HFTs capable of picking off resting orders. I also find strong evidence of a predator-prey interaction between FAST and SLOW traders. The FAST traders prefer to be in the same market as the SLOW traders, regardless of market, whereas the SLOW traders ultimately seek the protection and efficiency gains of the frequent call market, as long as the CALL is sufficiently thick.

Overall, my results demonstrate that a frequent call market functions as an attractor for SLOW traders, as FAST traders are willing to follow the SLOW traders to either market. The

predators (e.g., the HFT real-world counterparts to FAST traders) will always pursue their prey (e.g., institutional and retail investors), but in a frequent call market, the SLOW traders will be better protected from adverse selection and sniping. This suggests that frequent call markets in the wild could attract sufficient volume for viability, while deterring the wasteful pursuit of tiny latency advantages.

Several limitations should be taken into account in evaluating my results. My trader strategy set is fairly limited, with both FAST and SLOW traders employing the same set of strategies. One particularly unrealistic restriction is that traders cannot alter their market choice once it has been made. In addition, my results in settings without mean reversion demonstrate that fixed strategies are problematic for the environment agents, who can be exploited by strategic traders in the CDA. Interesting extensions might include strategies that permit learning or adaptive selection of the market mechanism, or formulating an iterated form of my market choice game. Similarly, analysis of a broader range of environments could provide further insight on the relative attractiveness of alternative market mechanisms. Additional market conditions of interest include slower environment agents, as well as different clearing frequencies.

CHAPTER VII

Conclusions

Over the past decade, algorithmic trading has become a dominant force in today's fragmented and complex financial marketplace. The highly computerized nature of such trading has made it possible to operate at speeds well beyond human perception: high-frequency traders exploit microsecond-level latency advantages in order to capitalize on opportunities for risk-free profit. The pursuit of higher speed in market access and response has perpetuated a latency arms race in which many market participants have spent (and continue to spend) billions of dollars. Speed offers a competitive edge due to the continuous nature of current markets, and market fragmentation and securities regulations have inadvertently spawned price disparities that can be exploited by the fastest traders. Modifications to current market structure have been proposed in efforts to mitigate the latency arms race, but the full extent of their impact is unclear. As such, characterizing the interrelationship between algorithmic trading and market structure, as well as its economic significance, is of paramount concern.

In this dissertation, I examined the interplay between algorithmic trading and market structure through a computational lens. I designed models to capture core aspects of current financial markets as well as various algorithmic trading strategies. I focused on two trading mechanisms: the continuous double auction, in which orders are matched as they arrive, and the frequent call market, in which orders are matched to trade at regular, pe-

riodic intervals. My market models are populated by background traders, who represent investors—in contrast to market participants whose sole objective is maximizing trading profit. Background traders are permitted to reenter to trade based on an individual valuation (comprised of both private and common components) for the security in question.

I employed a simulation-based approach to model, analyze, and characterize the relationship between trader behavior and market structure. Agent-based modeling facilitated the representation and encoding of markets and traders, and discrete-event simulation allowed me to precisely specify communication latencies and the flow of information in my models. I employed the methodology of empirical game-theoretic analysis to compare strategic interactions and market outcomes in equilibrium.

I explored a variety of market configurations in order to better characterize the impact of algorithmic trading and market structure on allocative efficiency, a measure of how well the market is distributing trades according to underlying private valuations, in addition to other market performance characteristics such as liquidity and price discovery. What follows is a summary of my contributions from the three case studies presented in this thesis.

Welfare Effects of Market Making This study examined the effects of market making on market performance, focusing on allocative efficiency as well as total background-trader surplus. Through liquidity provision, market making is generally considered to perform a valuable function in continuous markets, but the impact of this behavior on welfare depends on the specific market conditions. I modeled a single security traded in a continuous double auction market populated by multiple background traders. I employed empirical simulation-based methods to derive equilibria with and without a single market maker in a number of different market environments. My results show that not only is the market maker profitable in equilibrium, but its presence also significantly improves efficiency. Whether this effect is also reflected in background-trader gains depends on characteristics of the environment, such as market thinness (as captured by the number of traders) and

investor impatience (as captured by the length of the trading horizon and the background-trader reentry rate). I find that market making tends to improve welfare of impatient investors, but not in all cases. In addition, my analysis demonstrates that liquidity proxy measures, such as spreads and execution times, are not adequate substitutes for directly evaluating investor welfare.

A Two-Market Model of Latency Arbitrage and Market Fragmentation In this study, I developed a simple two-market model of latency arbitrage that captures the effects of market fragmentation, current U.S. securities regulations, and market clearing rules. These arbitrage opportunities arise due to the fragmentation of markets across multiple exchanges and delays in updating the public price quote, which can cause orders to be routed to the incorrect market. I show that the presence of a latency arbitrageur significantly degrades overall allocative efficiency, and I present frequent call markets as a means to eliminate the speed advantages of HFTs. My results demonstrate that periodic clearing, as in a frequent call market, not only eliminates latency arbitrage opportunities, but also improves welfare by aggregating orders over each clearing interval.

Frequent Call Markets vs. Continuous Double Auctions for Fast and Slow Traders In my third and final study, I investigated the potential for frequent call markets to coexist with continuous trading—the dominant mechanism in current financial markets—via a game of strategic market choice. I constructed a model of a single security traded simultaneously in a CDA market and a frequent call market. The market environment is populated by multiple investors, grouped in one of two roles (FAST and SLOW) that differ only in the reentry frequency. Traders select a market type (frequent call market or CDA) as part of their strategy. My results provide strong evidence that a frequent call market could coexist and attract sufficient volume alongside continuous markets. In equilibrium, the welfare of slower traders is generally higher in the frequent call market. I also identify a predator-prey relationship between the two types of traders: the faster traders prefer to be where

the slower traders are, whereas the slow traders achieve greater welfare in the discrete-time market.

Through computational modeling of algorithmic traders and current financial market structure, my dissertation provides a characterization of the market conditions under which algorithmic trading activity can benefit or harm investors. This work also presents an analysis of the feasibility and efficacy of market-design interventions—such as periodic clearing at regular intervals, as in a frequent call market—to mitigate adverse effects of certain algorithmic trading strategies. Research of this nature is of potential importance to market participants, policymakers, and regulators, as algorithmic traders now operate at timescales faster than the speed of human response. Such trading activity can lead to significant economic consequences within a matter of seconds. The computational approach I employed in this thesis offers an effective framework for analyzing the social welfare implications of algorithmic traders in today's financial trading landscape. Although my thesis covers only three case studies, the methodology I have presented here has much wider scope, and can be readily applied to investigate a broad spectrum of market scenarios, both current and potential. The nature of trading has changed significantly within the last decade, and no doubt it will continue to evolve. This dissertation lays the groundwork for further study of the relationship between trader behavior and market structure, as both evolve in the years to come.

APPENDICES

APPENDIX A

Equilibria in Market Maker Games

A.1 Equilibria in Games without a Market Maker

Table A.1: Symmetric equilibria for games without market making, $N = 66$, calculated from the 6-player DPR approximation. The numeric column headings give $R_{\max}$ values for the ZI strategies. All employ $R_{\min} = 0$ with the exception of the double star and double dagger (‡) values which use $R_{\min} = \frac{1}{2}R_{\max}$. All employ $\eta = 1$, except for the single starred values which use $\eta = 0.8$, the dagger (†) value which uses $\eta = 0.6$, and the double dagger (‡) values which use $\eta = 0.4$. Each row of the table describes the mixture probabilities for strategies for one equilibrium, and corresponds to the matching row in Table 4.2.

Env	65*	125*	125	250*	250	500	500**	1000*	1000	1000‡	1500†	2000‡	2500
A1	0	0	0	0	0	0	0	0	0	1	0	0	0
A4	0	0	0	0.223	0	0.106	0.861	0	0	0.033	0	0	0
A4	0	0.113	0	0	0.209	0	0.568	0	0	0	0	0	0
A12	0	0	0	0	0	0	0.887	0	0	0	0	0	0
B1	0	0	0	0	0.029	0.115	0.856	0	0	0	0	0	0
B4	0	0	0	0	0	0	0.824	0	0	0.176	0	0	0
B12	0	0	0	0	0	0	0.683	0	0	0.317	0	0	0
C1	0	0	0	0	0	0	0.714	0.193	0.093	0	0	0	0
C1	0	0	0	0	0	1	0	0	0	0	0	0	0
C4	0	0	0	0	0	0	1	0	0	0	0.193	0	0
C4	0	0	0	0	0	0.011	0.796	1	0	0	0	0	0
C12	0	0	0	0	0	0	0.960	0	0	0.040	0	0	0

Table A.2: Symmetric equilibria for games without market making, $N = 25$, calculated from the 5-player DPR approximation. Data presented is as for Table A.1. Each row corresponds to the matching row in Table 4.3.

Env	65*	125*	125	250*	250	500	500**	1000*	1000	1000‡	1500†	2000‡	2500
A1	0	0	0	0	0	0	0	0	0	0.978	0.022	0	0
A1	0	0	0	0	0	0.035	0.965	0	0	0	0	0	0
A4	0	0	0	0	0	0.318	0.682	0	0	0	0	0	0
A12	0	0	0	0	0	0	1	0	0	0	0	0	0
A24	0	0	0	0.256	0	0.744	0	0	0	0	0	0	0
A24	0	0	0.132	0.868	0	0	0	0	0	0	0	0	0
A1	0.072	0	0	0	0	0	0.928	0	0	0	0	0	0
B1	0	0	0	0	0	0	0	0	0	1	0	0	0
B4	0	0	0	0	0	0	0.525	0	0	0.475	0	0	0
B12	0	0	0	0	0	0	0.491	0.039	0	0.470	0	0	0
B24	0	0	0	0	0	0	0.621	0	0	0.379	0	0	0
C1	0	0	0	0	0	0	1	0	0	0	0	0	0
C4	0	0	0	0	0	0	0.823	0	0	0.177	0	0	0
C12	0	0	0	0	0	0	0.694	0	0	0.306	0	0	0
C24	0	0	0	0	0	0	0.730	0	0	0.270	0	0	0

A.2 Equilibria in Games with a Market Maker

Table A.3: Role-symmetric equilibria for games with a market maker, $N = 66$, calculated from the $(6, 1)$-player DPR approximation. The numeric column headings give $R_{\max}$ values for the ZI strategies as in Table A.1, followed by ω values for the MM strategies. All MM strategies use $K = 100$ and $\xi = 50$ except those with subscripts indicating the ξ used. Each row of the table corresponds to the matching row in Table 4.5 and describes one equilibrium found. The columns for $R_{\max}$ values of 2000‡ and 2500, as well as MM ω values of 64, 128, and 256_{25}, are not listed in the table as these strategies are not played in any of the equilibria found.

Env	Background-trader $R_{\max}$											Market maker ω			
	65*	125*	125	250*	250	500	500**	1000*	1000	1000‡	1500†	256	512	512_{100}	1024
A1	0	0	0	0.337	0.076	0.273	0.314	0	0	0	0	0	0.787	0.213	0
A1	0	0	0.116	0	0.185	0.699	0	0	0	0	0	0	0	1	0
A1	0.061	0	0	0.225	0	0	0.714	0	0	0	0	0	1	0	0
A4	0.034	0.072	0	0.724	0	0.171	0	0	0	0	0	1	0	0	0
A12	0	0.261	0	0	0.739	0	0	0	0	0	0	1	0	0	0
A12	0	0	0.144	0.254	0.602	0	0	0	0	0	0	1	0	0	0
B1	0	0	0	0	0	0	0.845	0	0	0.155	0	0	0	0.181	0.819
B4	0	0	0	0	0	0	0.843	0	0.012	0	0.145	0	0	1	0
B12	0	0	0	0	0	0	0.934	0	0	0	0.066	0	0	1	0
B12	0	0	0	0	0	0.032	0	0.968	0	0	0	0.082	0.918	0	0
C1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
C4	0	0	0	0	0	0	0.634	0.23	0.136	0	0	1	0	0	0
C4	0	0	0	0.081	0	0.919	0	0	0	0	0	1	0	0	0
C12	0	0	0	0	0	0	0	0.678	0.322	0	0	1	0	0	0
C12	0	0	0	0	0	0	0.554	0.446	0	0	0	1	0	0	0
C12	0	0	0	0.119	0	0	0	0	0.881	0	0	1	0	0	0

Table A.4: Role-symmetric equilibria for games with a market maker, $N = 25$, calculated from the $(5, 1)$-player DPR approximation. Each row of the table corresponds to the matching row in Table 4.6 and describes one equilibrium found. Data presented is as for Table A.3, but with columns for $R_{\max}$ values of 2000‡ and 2500 and MM ω values of 64, 128, and 1024 excluded from the table as these strategies are not played in any of the equilibria found.

Env	Background-trader $R_{\max}$											Market maker ω			
	65*	125*	125	250*	250	500	500**	1000*	1000	1000‡	1500†	256_{25}	256	512	512_{100}
A1	0	0	0	0.687	0	0.313	0	0	0	0	0	0	0	1	0
A1	0	0	0.156	0.300	0.544	0	0	0	0	0	0	0	0	0	1
A4	0	0	0	0.519	0	0.462	0.019	0	0	0	0	0	0	0	1
A4	0	0	0	0.221	0.779	0	0	0	0	0	0	0	0	0	1
A4	0	0	0	0.419	0.127	0.42	0.034	0	0	0	0	0	0	1	0
A4	0	0	0.130	0	0.870	0	0	0	0	0	0	0	0	0	1
A12	0	0.587	0	0	0.413	0	0	0	0	0	0	1	0	0	0
A12	0.149	0	0	0.835	0.016	0	0	0	0	0	0	1	0	0	0
A12	0	0	0.349	0	0.651	0	0	0	0	0	0	0.139	0.861	0	0
A24	0	0	0	0.856	0	0.112	0	0.32	0	0	0	0	0.321	0.679	0
A24	0.170	0.830	0	0	0	0	0	0	0	0	0	0	1	0	0
B1	0	0	0	0	0	0.367	0.633	0	0	0	0	0	0	0	1
B1	0	0	0	0	0	0	0.81	0	0	0.19	0	0	0	0	1
B4	0	0	0	0	0	0.088	0	0.912	0	0	0	0	0.852	0.148	0
B12	0	0	0	0	0	0	0.754	0.092	0	0.154	0	0	0	0.669	0.331
B24	0	0	0	0	0	0	0	0.256	0.537	0.208	0	0	1	0	0
B24	0	0	0	0	0	0	0.61	0.031	0.359	0	0	0	0	0	1
C1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1

Continued on next page

Table A.4 – *Continued from previous page*

Env	Background-trader $R_{\max}$											Market maker ω			
	65*	125*	125	250*	250	500	500**	1000*	1000	1000‡	1500†	256_{25}	256	512	512_{100}
C4	0	0	0	0	0	0.083	0.917	0	0	0	0	0	0	1	0
C12	0	0	0	0	0.315	0	0	0.685	0	0	0	1	0	0	0
C12	0	0	0	0	0.656	0	0	0.344	0	0	0	0	0	1	0
C12	0	0	0	0	0.239	0.742	0	0.019	0	0	0	0	1	0	0
C24	0	0	0	0	0	0	0.155	0.761	0	0	0.084	0	1	0	0
C24	0	0	0	0	0	0	0.897	0.103	0	0	0	0	0	1	0

APPENDIX B

Equilibria in Market Structure Games

Table B.1: Symmetric equilibria for market structure games for environment 1, $N = 24$, calculated from the 4-player DPR approximation. There is one game per latency $\delta \in \{0, 100, 200, 300, 400, 600, 700, 900\}$ per market configuration, which includes the two-market model (2M) both with and without LA, the centralized CDA, and the frequent call market (for which latency is equivalent to the length of the clearing interval). Each row of the table describes the mixture probabilities for strategies for one equilibrium, and corresponds to the matching row in Table 5.2. The numeric column headings give $R_{\max}$ values for the ZI strategies, for each role. All strategies employ $R_{\min} = 0$, with the exception of the double star and double dagger (‡) values which use $R_{\min} = \frac{1}{2}R_{\max}$. All strategies employ $\eta = 1$, except for the dagger (†) value which uses $\eta = 0.6$, and the circle (∘) and double dagger (‡) values which both use $\eta = 0.4$.

Model	Latency	125	250	500	500**	1000	1000‡	1000**	1500†	2000‡	2500∘	2500
CDA	–	0	0	0	0.096	0	0	0	0	0.528	0.376	0
CDA	–	0	0	0	0	0	0	0	0	0.507	0.493	0
2M	0	0	0	0	0	0	0	0	0	0	1	0
2M	0	0	0	0	0.177	0	0.123	0	0	0	0.7	0
Call	100	0.15	0.324	0	0	0	0	0	0.052	0	0.474	0
2M (no LA)	100	0	0	0	0	0	0	0	0	0.758	0.242	0
2M (no LA)	100	0	0	0	0	0	0.602	0	0	0.239	0.159	0
2M (LA)	100	0	0	0	0	0	0.237	0	0	0.537	0.226	0
Call	200	0.368	0	0.094	0	0.042	0	0	0	0	0.496	0
2M (no LA)	200	0	0	0	0	0	0.381	0	0	0.338	0.281	0
2M (LA)	200	0	0	0	0	0	0	0	0	0.679	0.321	0
Call	300	0.094	0.371	0	0	0	0	0	0	0	0.535	0
2M (no LA)	300	0	0	0	0	0	0	0	0	0.608	0.392	0
2M (no LA)	300	0	0	0	0	0	0.692	0	0	0.036	0.272	0
2M (LA)	300	0	0	0	0	0	0	0	0	0.655	0.345	0
Call	400	0	0	0.835	0.036	0	0	0	0	0	0	0.129
Call	400	0	0.416	0	0.163	0	0	0	0	0	0.421	0
Call	400	0	0.055	0	0.347	0.501	0	0	0.097	0	0	0
2M (no LA)	400	0	0	0	0	0	0	0	0	0.595	0.405	0

Continued on next page

Table B.1 – *Continued from previous page*

Model	Latency	125	250	500	500**	1000	1000‡	1000**	1500†	2000‡	2500°	2500
2M (LA)	400	0	0	0	0	0	0.47	0	0	0.258	0.272	0
Call	600	0	0.477	0	0	0.523	0	0	0	0	0	0
Call	600	0.22	0	0	0	0.379	0	0	0	0	0.401	0
Call	600	0.271	0.207	0	0.023	0	0	0	0	0	0.499	0
2M (no LA)	600	0	0	0	0	0	1	0	0	0	0	0
2M (no LA)	600	0	0	0	0	0	0.81	0	0	0	0.19	0
2M (LA)	600	0	0	0	0	0	0	0.029	0	0	0.971	0
2M (LA)	600	0	0	0	0	0	0.635	0	0	0	0.365	0
2M (LA)	600	0	0	0	0	0	0	0	0	0.643	0.308	0.049
Call	700	0.162	0.484	0.022	0	0.258	0	0	0.008	0	0.066	0
Call	700	0.185	0.471	0	0.059	0.216	0	0	0	0	0.069	0
2M (no LA)	700	0	0	0	0	0	0	0	0.209	0.791	0	0
2M (no LA)	700	0	0	0	0	0	0.739	0	0	0	0.261	0
2M (LA)	700	0	0	0	0	0	0.006	0	0	0.826	0.168	0
Call	900	0	0.246	0.498	0.256	0	0	0	0	0	0	0
2M (no LA)	900	0	0	0	0	0	0	0	0	0.836	0	0.164
2M (no LA)	900	0	0	0	0	0	0	0	0	0	1	0
2M (no LA)	900	0	0	0	0	0	0	0	0	0.537	0.463	0
2M (LA)	900	0	0	0	0	0	0	0.129	0.871	0	0	0
2M (LA)	900	0	0	0	0	0	0.131	0	0	0	0.869	0

Table B.2: Symmetric equilibria for market structure games for environment 2, $N = 238$, calculated from the 4-player DPR approximation. There is one game per latency $\delta \in \{0, 50, 100\}$ per market configuration, which includes the two-market model (2M) both with and without LA, the centralized CDA, and the frequent call market (for which latency is equivalent to the length of the clearing interval). Each row of the table describes the mixture probabilities for strategies for one equilibrium, and corresponds to the matching row in Table 5.3. Data presented is as for Table B.1.

Model	Latency	125	250	500	500**	1000	1000‡	1000**	1500†	2000‡	2500°	2500
CDA	–	0	0	0	0	0	0	0	0	0	0.726	0.274
CDA	–	0	0	0	0	0	0	0	0	0	0.659	0.341
2M	0	0.146	0	0	0	0	0	0	0	0	0.854	0
Call	50	0	0	0	0	0	0	0	0	0	1	0
2M (no LA)	50	0	0.162	0	0	0	0	0	0	0	0.838	0
2M (LA)	50	0	0	0.188	0	0	0	0	0	0	0.812	0
Call	100	0	0	0	0	0	0	0	0	0.1	0.739	0.161
2M (no LA)	100	0.051	0	0	0	0	0	0	0	0	0.76	0.189
2M (LA)	100	0	0	0	0	0	0	0	0	0.233	0.767	0

Table B.3: Symmetric equilibria for market structure games for environment 3, $N = 58$, calculated from the 4-player DPR approximation. There is one game per latency $\delta \in \{0, 25, 50, 75, 100\}$ per market configuration, which includes the two-market model (2M) both with and without LA, the centralized CDA, and the frequent call market (for which latency is equivalent to the length of the clearing interval). Each row of the table describes the mixture probabilities for strategies for one equilibrium, and corresponds to the matching row in Table 5.4. Data presented is as for Table B.1.

Model	Latency	125	250	500	500**	1000	1000‡	1000**	1500†	2000‡	2500°	2500
CDA	–	0	0	0	0	0	0	0	0	0.248	0.752	0
2M	0	0	0	0.017	0	0	0	0	0	0.004	0.979	0

Continued on next page

Table B.3 – *Continued from previous page*

Model	Latency	125	250	500	500**	1000	1000‡	1000**	1500†	2000‡	2500°	2500
Call	25	0	0	0.06	0	0	0	0	0	0	0.735	0.205
2M (no LA)	25	0	0	0	0	0	0	0	0	0	0.854	0.146
2M (LA)	25	0	0	0	0	0	0.117	0	0	0.883	0	0
2M (LA)	25	0	0	0	0	0	0	0	0	0.21	0.79	0
Call	50	0	0	0	0	0	0	0	0	0	1	0
2M (no LA)	50	0	0	0	0	0	0	0	0	0.782	0	0.218
2M (no LA)	50	0	0	0	0	0	0	0	0	0	0.948	0.052
2M (LA)	50	0	0	0	0	0	0	0.142	0	0.793	0	0.065
2M (LA)	50	0	0	0	0	0	0	0	0.065	0.043	0.892	0
Call	75	0	0.12	0	0	0	0	0	0	0	0.88	0
2M (no LA)	75	0	0	0	0	0	0	0	0	0	0.823	0.177
2M (LA)	75	0	0	0	0	0	0	0	0	0.142	0.858	0
Call	100	0	0	0	0	0	0	0	0	0.18	0.82	0
2M (no LA)	100	0	0	0	0	0	0	0	0	0.722	0.002	0.276
2M (no LA)	100	0	0	0	0	0	0	0	0	0	0.839	0.161
2M (LA)	100	0	0	0.082	0	0	0	0	0	0.918	0	0
2M (LA)	100	0	0	0.015	0	0	0	0	0	0.231	0.754	0

APPENDIX C

Equilibria in Strategic Market Choice Games

Table C.1: Role-symmetric equilibria for the first four strategic market choice games (one each for environments I–IV), $N_{\text{FAST}} = N_{\text{SLOW}} = 21$, calculated from the $(3, 3)$-player DPR approximation. Each row of the table describes one equilibrium found, the selected market mechanism (CALL or CDA), the welfare (total surplus of all players), and the mixture probabilities of strategies for each role in the RSNE. The numeric column headings give $R_{\max}$ values for the ZI strategies, for each role. All strategies employ $R_{\min} = 0$, with the exception of the double star and double dagger (‡) values which use $R_{\min} = \frac{1}{2}R_{\max} = 500$. All strategies employ $\eta = 1$, except for the double dagger (‡) values which use $\eta = 0.4$. A dash indicates that a strategy is not available for the specified market type. There is at least one all-CALL RSNE in each environment; all but environment IV have at least one all-CDA equilibrium.

Env	Market	Welfare	FAST							SLOW						
			125	250	500	1000	1000‡	1000**	2500	125	250	500	1000	1000‡	1000**	2500
I	CALL	27288	0	.965	.035	0	–	0	0	.106	0	.894	0	–	0	0
I	CALL	26697	.212	0	.788	0	–	0	0	0	.037	0	.963	–	0	0
I	CDA	27261	0	0	1	0	0	–	0	.277	0	.723	0	0	–	0
I	CDA	26785	0	.064	.551	0	.385	–	0	0	0	1	0	0	–	0
I	CDA	25321	0	0	.422	.483	.095	–	0	0	0	0	0	1	–	0
I	CDA	26133	0	0	.383	0	.617	–	0	0	0	0	1	0	–	0
II	CALL	21050	.347	.120	0	0	–	0	.533	0	0	0	0	–	0	1
II	CDA	21242	0	0	.080	0	.920	–	0	0	0	0	0	0	–	1
III	CALL	19992	.510	0	0	0	–	0	.490	0	0	0	0	–	0	1
III	CALL	20441	0	0	0	1	–	0	0	.117	0	0	0	–	0	.883
III	CDA	19734	0	0	0	0	1	–	0	0	0	0	0	0	–	1
IV	CALL	18067	.236	0	0	0	–	0	.764	0	0	0	0	–	0	1

Table C.2: Role-symmetric equilibria for the three strategic market choice games without mean reversion (one each for environments V–VII), $N_{\text{FAST}} = N_{\text{SLOW}} = 21$, calculated from the $(3, 3)$-player DPR approximation. Data presented is as for Table C.1. There is at least one all-CDA RSNE in each environment. Environment V has one equilibrium in which the FAST traders are in both the CALL and the CDA, and all SLOW traders are in the CALL; all reported mixture probabilities for this RSNE are for the CALL, with the exception of column $R_{\max} = 1000^{\ddagger}$ (which is the only strategy in the CDA in this equilibrium).

Env	Market	Welfare	FAST							SLOW						
			125	250	500	1000	$1000^{\ddagger}$	1000**	2500	125	250	500	1000	$1000^{\ddagger}$	1000**	2500
V	Both	24038	0	.243	.506	0	.251	0	0	0	0	0	0	–	1	0
V	CDA	26457	0	0	0	0	1	–	0	0	0	0	0	1	–	0
VI	CDA	28681	0	0	0	0	1	–	0	0	0	0	0	1	–	0
VII	CDA	29412	0	0	0	0	1	–	0	0	0	0	0	1	–	0

BIBLIOGRAPHY

BIBLIOGRAPHY

ABERNETHY, J., CHEN, Y., AND WORTMAN VAUGHAN, J. 2011. An optimization-based framework for automated market-making. In *12th ACM Conference on Electronic Commerce*. 297–306.

ABERNETHY, J. AND KALE, S. 2013. Adaptive market making via online learning. In *Advances in Neural Information Processing Systems*. 2058–2066.

ADLER, J. 2012. Raging bulls: How Wall Street got addicted to light-speed trading. *Wired Magazine 20,* 9.

AMIHUD, Y., LAUTERBACH, B., AND MENDELSON, H. 2003. The value of trading consolidation: Evidence from the exercise of warrants. *Journal of Financial and Quantitative Analysis 38,* 4, 829–846.

AMIHUD, Y. AND MENDELSON, H. 1980. Dealership market: Market-making with inventory. *Journal of Financial Economics 8,* 1, 31–53.

AMIHUD, Y., MENDELSON, H., AND LAUTERBACH, B. 1997. Market microstructure and securities values: Evidence from the Tel Aviv Stock Exchange. *Journal of Financial Economics 45,* 3, 365–390.

ANGEL, J. J., HARRIS, L. E., AND SPATT, C. S. 2011. Equity trading in the 21st century. *Quarterly Journal of Finance 1,* 1, 1–53.

ARNUK, S. L. AND SALUZZI, J. C. 2012. *Broken Markets: How High Frequency Trading and Predatory Practices on Wall Street are Destroying Investor Confidence and Your Portfolio*. FT Press.

AVELLANEDA, M. AND STOIKOV, S. 2008. High-frequency trading in a limit order book. *Quantitative Finance 8,* 3, 217–224.

BALDAUF, M. AND MOLLNER, J. 2014. High-frequency trade and market performance. Tech. rep., Stanford University Economics Department. December.

BALDAUF, M. AND MOLLNER, J. 2015. Trading in fragmented markets. Tech. Rep. No. 15-018, Stanford Institute for Economic Policy Research. May.

BANKS, J., CARSON II, J. S., NELSON, B. L., AND NICOL, D. M. 2005. *Discrete-Event System Simulation* Fourth Ed. Prentice Hall.

BARON, M., BROGAARD, J., AND KIRILENKO, A. 2012. The trading profits of high frequency traders. Tech. rep., Commodity Futures Trading Commission.

BENNETT, P. AND WEI, L. 2006. Market structure, fragmentation, and market quality. *Journal of Financial Markets 9,* 1, 49–78.

BESSEMBINDER, H. 2003. Issues in assessing trade execution costs. *Journal of Financial Markets 6,* 1, 233–257.

BESSEMBINDER, H., HAO, J., AND LEMMON, M. L. 2011. Why designate market makers? Affirmative obligations and market quality. *SSRN Electronic Journal*, 1–66.

BESSEMBINDER, H., HAO, J., AND ZHENG, K. 2015. Market making contracts, firm value, and the IPO decision. *Journal of Finance 70,* 5, 1997–2028.

BIAIS, B., GLOSTEN, L., AND SPATT, C. 2005. Market microstructure: A survey of microfoundations, empirical results, and policy implications. *Journal of Financial Markets 8,* 2, 217–264.

BLUME, M. E. 2007. Competition and fragmentation in the equity markets: The effects of Regulation NMS. Tech. Rep. 02-07, The Rodney L. White Center for Financial Research, The Wharton School, University of Pennsylvania.

BOWLEY, G. 2010. U.S. markets plunge, then stage a rebound. *The New York Times*.

BROGAARD, J. 2010. High frequency trading and its impact on market quality. *Northwestern University Kellogg School of Management Working Paper*.

BRUSCO, S. AND JACKSON, M. O. 1999. The optimal design of a market. *Journal of Economic Theory 88,* 1, 1–39.

BUCHANAN, M. 2009. Meltdown modelling. *Nature 460,* August, 680–682.

BUDIMIR, M. 2014. The Xetra intraday auction: Growing potential for strong price discovery. *Best Execution Winter 2014/2015*.

BUDISH, E., CRAMTON, P., AND SHIM, J. 2014. Implementation details for frequent batch auctions: Slowing down markets to the blink of an eye. *American Economic Review 104,* 5, 418–424.

BUDISH, E., CRAMTON, P., AND SHIM, J. 2015. The high-frequency trading arms race: Frequent batch auctions as a market design response. *Quarterly Journal of Economics 130,* 4, 1547–1621.

CARDELLA, L., HAO, J., KALCHEVA, I., AND MA, Y.-Y. 2014. Computerization of the equity, foreign exchange, derivatives, and fixed-income markets. *Financial Review 49,* 2, 231–243.

CASSELL, B.-A. AND WELLMAN, M. P. 2013. EGTAOnline: An experiment manager for simulation-based game studies. In *Multi-Agent-Based Simulation XIII*. Lecture Notes in Artificial Intelligence Series, vol. 7838. Springer.

CHAKRABORTY, T. AND KEARNS, M. 2011. Market making and mean reversion. In *11th ACM Conference on Electronic Commerce*. 307–314.

CHAN, N. T. AND SHELTON, C. 2001. An electronic market-maker. Tech. Rep. AI Memo 2001-005, Massachusetts Institute of Technology.

CHEN, Y. AND PENNOCK, D. M. 2007. A utility framework for bounded-loss market makers. In *23rd Conference on Uncertainty in Artificial Intelligence*. 49–56.

CHOWHDRY, B. AND NANDA, V. 1991. Multimarket trading and market liquidity. *Review of Financial Studies 4,* 3, 483–511.

CLARK, J. 2014. Thomson Reuters to trial randomization on FX matching platform. *Euromoney*.

CLIFF, D. 2009. ZIP60: Further explorations in the evolutionary design of online auction market mechanisms. *IEEE Transactions on Evolutionary Computation 13*, 3–18.

COHEN, S. N. AND SZPRUCH, L. 2012. A limit order book model for latency arbitrage. *Mathematics and Financial Economics 6*, 211–227.

DARLEY, V., OUTKIN, A., PLATE, T., AND GAO, F. 2000. Sixteenths or pennies? Observations from a simulation of the NASDAQ stock market. In *IEEE/IAFE/INFORMS Conference on Computational Intelligence for Financial Engineering*. 151–154.

DAS, R., HANSON, J. E., KEPHART, J. O., AND TESAURO, G. 2001. Agent-human interactions in the continuous double auction. In *17th International Joint Conference on Artificial Intelligence*. 1169–1176.

DAS, S. 2005. A learning market-maker in the Glosten–Milgrom model. *Quantitative Finance 5,* 2, 169–180.

DAS, S. 2008. The effects of market-making on price dynamics. In *7th International Joint Conference on Autonomous Agents and Multiagent Systems*. 887–894.

DAS, S. AND MAGDON-ISMAIL, M. 2008. Adapting to a market shock: Optimal sequential market-making. In *Advances in Neural Information Processing Systems*. 361–368.

DE LA MERCED, M. J. 2013. Shutdown at Nasdaq is traced to software. *The New York Times*.

DING, S., HANNAH, J., AND HENDERSHOTT, T. 2014. How slow is the NBBO? A comparison with direct exchange feeds. *Financial Review 49,* 2, 313–332.

DU, S. AND ZHU, H. 2014. Welfare and optimal trading frequency in dynamic double auctions. Tech. rep., National Bureau of Economic Research.

ECONOMIDES, N. AND SCHWARTZ, R. A. 1995. Electronic call market trading. *Journal of Portfolio Management 21,* 3, 10–18.

FARMER, J. D. AND FOLEY, D. 2009. The economy needs agent-based modelling. *Nature 460,* 7256, 685–686.

FARMER, J. D. AND SKOURAS, S. 2012. Review of the benefits of a continuous market vs. randomised stop auctions and of alternative priority rules (policy options 7 and 12).

FEATHERSTONE, D. 2014. Frequent batch auctions. Tech. rep., Optiver. December.

FENG, Y., YU, R., AND STONE, P. 2004. Two stock-trading agents: Market making and technical analysis. In *Agent-Mediated Electronic Commerce V. Designing Mechanisms and Systems*. Springer, 18–36.

FOUCAULT, T., KOZHAN, R., AND THAM, W. W. 2015. Toxic arbitrage. Tech. Rep. No. FIN-2014-1040, HEC Paris.

FREY, S. AND GRAMMIG, J. 2006. Liquidity supply and adverse selection in a pure limit order book market. *Empirical Economics 30,* 4, 1007–1033.

FRICKE, D. AND GERIG, A. 2015. Too fast or too slow? Determining the optimal speed of financial markets.

FRIEDMAN, D. 1993. The double auction market institution: A survey. In *The Double Auction Market: Institutions, Theories, and Evidence*, D. Friedman and J. Rust, Eds. Addison-Wesley, 3–25.

GAMMELTOFT, N. AND GRIFFIN, D. 2013. Goldman Sachs said to send stock-option orders by mistake. *Bloomberg*.

GARMAN, M. B. 1976. Market microstructure. *Journal of Financial Economics 3,* 3, 257–275.

GARVEY, R. AND WU, F. 2010. Speed, distance, and electronic trading: New evidence on why location matters. *Journal of Financial Markets 13,* 4, 367–396.

GLOSTEN, L. R. 1994. Is the electronic open limit order book inevitable? *Journal of Finance 49,* 4, 1127–1161.

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

GODE, D. K. AND SUNDER, S. 1993. Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. *Journal of Political Economy 101,* 1, 119–137.

GODE, D. K. AND SUNDER, S. 1997. What makes markets allocationally efficient? *Quarterly Journal of Economics 112,* 2, 603–630.

GOETTLER, R. L., PARLOUR, C. A., AND RAJAN, U. 2009. Informed traders and limit order markets. *Journal of Financial Economics 93,* 1, 67–87.

GOLDSTEIN, M. A., KUMAR, P., AND GRAVES, F. C. 2014. Computerized and high-frequency trading. *The Financial Review 49,* 2, 177–202.

GOVERNMENT OFFICE FOR SCIENCE, LONDON. 2012. Foresight: The future of computer trading in financial markets. Tech. rep.

GROSSMAN, S. J. AND MILLER, M. H. 1988. Liquidity and market structure. *Journal of Finance 43,* 3, 617–633.

HANSON, R. 2007. Logarithmic market scoring rules for modular combinatorial information aggregation. *Journal of Prediction Markets 1*, 3–15.

HANSON, T. A. 2012. The effects of high frequency traders in a simulated market. In *Midwest Finance Association Annual Meeting*.

HASBROUCK, J. AND SAAR, G. 2013. Low-latency trading. *Journal of Financial Markets 16,* 4, 646–679.

HASBROUCK, J. AND SOFIANOS, G. 1993. The trades of market makers: An empirical analysis of NYSE specialists. *Journal of Finance 48,* 5, 1565–1593.

HENDERSHOTT, T., JONES, C. M., AND MENKVELD, A. J. 2011. Does algorithmic trading improve liquidity? *Journal of Finance 66,* 1, 1–33.

HOPE, B. 2015. NYSE Group planning midday auction for its stock markets. *The Wall Street Journal*.

HUANG, J. AND WANG, J. 2010. Market liquidity, asset prices, and welfare. *Journal of Financial Economics 95*, 107–127.

HUANG, R. D. AND STOLL, H. R. 1996. Dealer versus auction markets: A paired comparison of execution costs on NASDAQ and the NYSE. *Journal of Financial Economics 41*, 313–357.

INDUSTRY SUPER NETWORK. 2013. Toward a fairer and more efficient share market: Frequent sealed bid call auctions with random durations. Tech. rep. February. ISN Research Report.

JACOBS, B. I., LEVY, K. N., AND MARKOWITZ, H. M. 2004. Financial market simulation. *Journal of Portfolio Management 30,* 5, 142–152.

JARROW, R. A. AND PROTTER, P. 2012. A dysfunctional role of high frequency trading in electronic markets. *International Journal of Theoretical and Applied Finance 15,* 3, 1250022–1–1250022–15.

JORDAN, P. R., WELLMAN, M. P., AND BALAKRISHNAN, G. 2010. Strategy and mechanism lessons from the first Ad Auctions Trading Agent Competition. In *11th ACM Conference on Electronic Commerce*. 287–296.

JUMADINOVA, J. AND DASGUPTA, P. 2010. A comparison of different automated market-maker strategies. In *12th Workshop on Agent-Mediated Electronic Commerce*. 141–154.

KALAY, A., WEI, L., AND WOHL, A. 2002. Continuous trading or call auctions: Revealed preferences of investors at the Tel Aviv Stock Exchange. *Journal of Finance 57,* 1, 523–542.

KEARNS, M., KULESZA, A., AND NEVMYVAKA, Y. 2010. Empirical limitations on high-frequency trading profitability. *Journal of Trading 5,* 4, 50–62.

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

LANG, L. H. P. AND LEE, Y. T. 1999. Performance of various transaction frequencies under call markets: The case of Taiwan. *Pacific-Basin Finance Journal 7,* 1, 23–39.

LEACH, J. C. AND MADHAVAN, A. N. 1992. Intertemporal price discovery by market makers: Active versus passive learning. *Journal of Financial Intermediation 2,* 2, 207–235.

LEBARON, B. 2004. Building the Santa Fe artificial stock market. In *Agent-Based Economics: Theory, Languages and Experiments*, F. Luna, P. Tierna, and A. Perrone, Eds. Routledge Publishing.

LEBARON, B. 2006. Agent-based computational finance. In *Handbook of Agent-Based Computational Economics*, L. Tesfatsion and K. L. Judd, Eds. Elsevier, 1187–1233.

LEBARON, B., ARTHUR, W. B., AND PALMER, R. 1999. Time series properties of an artificial stock market. *Journal of Economic Dynamics & Control 23,* 1, 1487–1516.

LEE, W. B., CHENG, S.-F., AND KOH, A. 2011. Would price limits have made any difference to the 'Flash Crash' on May 6, 2010? *Review of Futures Markets 9*, 55–93.

LEE, Y.-T., LIU, Y.-J., ROLL, R., AND SUBRAHMANYAM, A. 2004. Order imbalances and market efficiency: Evidence from the Taiwan Stock Exchange. *Journal of Financial and Quantitative Analysis 39,* 2, 327–341.

LEWIS, M. 2014. *Flash Boys: A Wall Street Revolt*. W. W. Norton & Company.

LI, Z. AND DAS, S. 2016. An agent-based model of competition between financial exchanges: Can frequent call mechanisms drive trade away from CDAs? In *15th International Conference on Autonomous Agents and Multiagent Systems*. To appear.

MADHAVAN, A. 1992. Trading mechanisms in securities markets. *Journal of Finance 47,* 2, 607–641.

MADHAVAN, A. 2000. Market microstructure: A survey. *Journal of Financial Markets 3,* 3, 205–258.

MADHAVAN, A., MING, K., STRASER, V., AND WANG, Y. 2002. How effective are effective spreads? an evaluation of trade side classification algorithms. Tech. rep., ITG, Inc.

MANASTER, S. AND MANN, S. C. 1996. Life in the pits: Competitive market making and inventory control. *Review of Financial Studies 9,* 3, 953–975.

MCPARTLAND, J. 2013. Recommendations for equitable allocation of trades in high frequency trading environments. Tech. rep., Federal Reserve Bank of Chicago.

MEHTA, N. 2012. Nasdaq chief blames software for delay in Facebook debut. *Bloomberg*.

MENDELSON, H. 1987. Consolidation, fragmentation, and market performance. *Journal of Financial and Quantitative Analysis 22,* 2, 189–207.

MENKVELD, A. J. 2013. High frequency trading and the new market makers. *Journal of Financial Markets 16,* 4, 712–740.

O'HARA, M. 1995. *Market Microstructure Theory*. Vol. 108. Blackwell Cambridge.

O'HARA, M. AND OLDFIELD, G. 1986. The microeconomics of market making. *Journal of Financial and Quantitative Analysis 21,* 4, 361–376.

O'HARA, M. AND YE, M. 2011. Is market fragmentation harming market quality? *Journal of Financial Economics 100,* 3, 459–474.

PALMER, R. G., ARTHUR, W. B., HOLLAND, J. H., LEBARON, B., AND TAYLER, P. 1994. Artifical economic life: A simple model of a stock market. *Physica D: Nonlinear Phenomena 75,* 1, 264–274.

PANCS, R. 2013. Comparing market structures: Allocative and informational efficiencies of continuous trading, periodic auctions, and dark pools. Tech. rep., Department of Economics, University of Rochester. February.

PATTERSON, S. 2014. High-speed stock traders turn to laser beams. *The Wall Street Journal*.

PELLIZZARI, P. AND DAL FORNO, A. 2007. A comparison of different trading protocols in an agent-based market. *Journal of Economic Interaction and Coordination 2,* 1, 27–43.

POPPER, N. 2012. Flood of errant trades is a black eye for Wall Street. *The New York Times*.

RABERTO, M. AND CINCOTTI, S. 2005. Modeling and simulation of a double auction artificial financial market. *Physica A: Statistical Mechanics and its Applications 355,* 1, 34–45.

RIORDAN, R. AND STORKENMAIER, A. 2012. Latency, liquidity and price discovery. *Journal of Financial Markets 15,* 4, 416–437.

ROSOV, S. 2014. Are frequent batch auctions a solution to HFT latency arbitrage? http://blogs.cfainstitute.org/marketintegrity/2014/11/10/are-frequent-batch-auctions-a-solution-to-hft-latency-arbitrage.

ROSS, D. K. 2014. Synchronized frequent batch auctions: A rebuttal. http://tabbforum.com/opinions/synchronized-frequent-batch-auctions-a-rebuttal.

SAAR, G. 2010. Specialist markets. In *Encyclopedia of Quantitative Finance*. Wiley Online Library.

SANDÅS, P. 2001. Adverse selection and competitive market making: Empirical evidence from a limit order market. *Review of Financial Studies 14,* 3, 705–734.

SCHNEIDER, D. 2012. The microsecond market. *IEEE Spectrum 49,* 6, 66–81.

SCHNEIDERMAN, E. 2014. Remarks on high-frequency trading & insider trading 2.0. New York Law School Panel on "Insider Trading 2.0—A New Initiative to Crack Down on Predatory Practices".

SCHWARTZ, R. A. AND PENG, L. 2013. Market makers. In *Encyclopedia of Finance*. Springer, 487–489.

SCHWARTZ, R. A. AND WU, L. 2013. Equity trading in the fast lane: The staccato alternative. *Journal of Portfolio Management 39,* 3, 3–6.

SECURITIES AND EXCHANGE COMMISSION. 2005. Regulation NMS. 17 CFR Parts 200, 201, 230, 240, 242, 249, 270.

SELLBERG, L.-I. 2010. Using adaptive micro auctions to provide efficient price discovery when access in terms of latency is differentiated among market participants. Tech. rep., Cinnober Financial Technology AB.

SEPPI, D. J. 1997. Liquidity provision with limit orders and a strategic specialist. *Review of Financial Studies 10,* 1, 103–150.

SPARROW, C. 2012. The failure of continuous markets. *Journal of Trading 7,* 2, 44–47.

STAFFORD, P. 2014. LSE to launch midday stock auction. *Financial Times*.

THURNER, S., FARMER, J. D., AND GEANAKOPLOS, J. 2012. Leverage causes fat tails and clustered volatility. *Quantitative Finance 12*, 695–707.

VIVES, X. 2010. *Information and Learning in Markets: The Impact of Market Microstructure*. Princeton University Press.

VYTELINGUM, P., CLIFF, D., AND JENNINGS, N. R. 2008. Strategic bidding in continuous double auctions. *Artificial Intelligence 172,* 14, 1700–1729.

WAH, E. 2016. How prevalent and profitable are latency arbitrage opportunities on U.S. stock exchanges? *SSRN Electronic Journal*.

WAH, E., BARR, M., RAJAN, U., AND WELLMAN, M. P. 2013. Public Comment in response to the U.S. Commodity Futures Trading Commission Concept Release on Risk Controls and System Safeguards for Automated Trading Environments. `http://comments.cftc.gov/PublicComments/ViewComment.aspx?id=59450`.

WAH, E., HURD, D. R., AND WELLMAN, M. P. 2015. Strategic market choice: Frequent call markets vs. continuous double auctions for fast and slow traders. In *3rd EAI Conference on Auctions, Market Mechanisms, and their Applications*.

WAH, E. AND WELLMAN, M. P. 2013. Latency arbitrage, market fragmentation, and efficiency: A two-market model. In *14th ACM Conference on Electronic Commerce*. 855–872.

WAH, E. AND WELLMAN, M. P. 2015. Welfare effects of market making in continuous double auctions. In *14th International Conference on Autonomous Agents and Multiagent Systems*. 57–66.

WAH, E., WRIGHT, M., AND WELLMAN, M. P. 2016. Welfare effects of market making in continuous double auctions. Tech. rep., University of Michigan.

WEBB, R. I., MUTHUSWAMY, J., AND SEGARA, R. 2007. Market microstructure effects on volatility at the TAIFEX. *Journal of Futures Markets 27,* 12, 1219–1243.

WELLMAN, M. P. 2006. Methods for empirical game-theoretic analysis (extended abstract). In *21st National Conference on Artificial Intelligence*. 1552–1555.

WELLMAN, M. P. 2009. Countering high-frequency trading. `http://ai.eecs.umich.edu/people/wellman/?p=40`.

WELLMAN, M. P. 2011. *Trading Agents*. Morgan & Claypool.

WELLMAN, M. P. AND WAH, E. 2016. Strategic agent-based modeling of financial markets. *Russell Sage Foundation Journal of the Social Sciences*. To appear.

WHEATLEY, M. 2010. We need rules to limit the risks of superfast trades. *Financial Times*.

WHITE, M. J. 2014. Enhancing our equity market structure. Sandler O'Neill & Partners, L.P. Global Exchange and Brokerage Conference.

WIEDENBECK, B. AND WELLMAN, M. P. 2012. Scaling simulation-based game analysis through deviation-preserving reduction. In *11th International Conference on Autonomous Agents and Multiagent Systems*. 931–938.

ZHAN, W. AND FRIEDMAN, D. 2007. Markups in double auction markets. *Journal of Economic Dynamics and Control 31,* 9, 2984–3005.