High Frequency Trading in Financial Markets: Information, Speed and Learning

Junqing Kang

A dissertation submitted in fulfilment of the requirements for the degree of Doctor of Philosophy (PhD)

Finance Discipline Group, UTS Business School
University of Technology Sydney

Principal Supervisor: Professor *Xue-Zhong (Tony) He*
Co-supervisor: Professor *F. Douglas Foster*
Co-supervisor: Dr *Lei Shi*
Co-supervisor: Associate Professor *Jianxin Wang*

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Junqing Kang, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy (PhD), in the Finance Discipline Group, UTS Business School at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

Signature: Signature removed prior to publication.

Date: 24 August 2020

Acknowledgements

Firstly, I would like to express my deep gratitude to my principal supervisor Professor Xue-Zhong (Tony) He. The completion of this thesis would not be possible without the encouragement, guidance and support from him. The discussions we had throughout my PhD candidature have proven to be most helpful. Not only guided me into the world of finance theory and taught me the way of conducting research, he also provided incredibly support and encouragement for my four years life in Sydney. His positive attitude towards research and life will continue to be a motivation for my future professional career. I am immensely fortunate to have him as my supervisor.

Besides, I would also thank my committee members Professor F. Douglas Foster, Dr. Lei Shi and Associate Professor Jianxin Wang. I learnt a lot from the discussions with them and benefited greatly from their knowledge. Furthermore, I greatly appreciate the discussions and supports from my co-authors Shen Lin and Xuan Zhou. I am eternally grateful for their advices about different aspects of research, accompany, and encourage through my PhD journey. They are always there to provide support for my studies, life and careers.

My grateful thanks are also extended to all the faculty members and my colleagues at the UTS business school. I thank my PhD coordinator Christina Nikitopoulos Sklibosios for her support, guidance and encouragement. I benefited a lot from the discussions with Dr. Adrian Lee, Dr. Kai Li, Dr. Le Ma, Professor Tails Putnins and Dr. Yajun Xiao. I thank my fellow Ph.D. colleague Keunbae Ahn, Nihad Aliyev, Marc Bohmann, Xikai Chen, Wei Hu, Jonathan Ravn Karlsen, Marta Khomyn, Man Nguyen, Thanh Huong Nguyen for their enormous help and support. Apart from these, I would like to especially express my appreciation to Liyao Zhu for her accompany. Her support has been indispensable for the completion of my doctoral studies.

Lastly, but not the least, I am extremely grateful and fortunate for the constant support from my loving mother and father.

Abstract

This thesis consists of three essays on high frequency trading in financial markets with respect to speed acquisition, information production, and learning. Development of financial markets is largely driven by technological innovations. Starting with the computerization of tasks on trading floors, through the introduction of completely electronic markets, to Algorithmic Trading (AT) and High Frequency Trading (HFT), trading becomes much faster. Whereases financial market participants have benefited enormously from these trading transformations, we have also experienced market instabilities or failures such as market breaks and flash crashes. Economists have expressed their concern about the potentially negative effects of "excessive speed", which brings some challenging issues for market regulators. Within rational expectations equilibrium framework, this thesis examines the impact of HFT on strategic trading behaviour among interacting traders, information efficiency, and market liquidity. The findings provide some insights into the underlying economic mechanisms and policy implications under the scope of the electronic evolution of modern trading environment.

The first phenomenon investigated in this thesis is the sudden liquidity deteriorations in financial markets, such as the so-called "Flash Crash" incident. Chapter 2 proposes a nonlinear rational expectations equilibrium model of high-frequency endogenous liquidity provision to explore fragile liquidity. The risk from endogenous liquidity provision of high frequency traders (HFTs), coupled with limits to participation by designated market makers (DMMs), intensifies the adverse selection faced by DMMs. This can generate a gap between liquidity supply from DMMs and liquidity demand by informed traders. As a result, endogenous liquidity provision produces fragile liquidity, with the possibility of market breaks when HFTs switch from liquidity provision to consumption on the basis of unexpected shocks. With "circuit breaker", the presence of HFTs is more likely to triggering mandated trading halts, though market liquidity can be improved. The competition among HFTs with overlapped information can ease a

systematic liquidity withdrew simultaneously but speed up liquidity dry-up, resulting market breaking down in alarming fashion during market turmoil.

Those who take the view that fast trading can be harmful for financial markets have proposed a broad array of policies to slow down trading in financial markets, e.g., “speed bumps” on IEX or Alpha. Chapter 3 theoretically evaluates how technological innovations drive fast trading investment for both speculators and exchanges and their impact on market. The negative externality of the speed acquisition from fast speculators can result in excessive investment, which is intensified as speculators’ speed technology advances. As exchange’s speed technology advances, faster exchange makes faster speculators more concentrated; that is, higher exchange speed shrinks market fraction of fast speculators but stimulates their optimal trading speed. As the result, market liquidity is improved but price discovery is reduced. Policy makers aiming to balance price discovery and deadweight loss from costly speed investment may lead to a mismatch between the desired exchange speed for policy makers and the optimal speed supplied by exchange, echoing the concerns of market regulations about market failure on speed arms race.

Finally, Chapter 4 explores how the evolution in speed technologies influences information production and strategic trading among different investors and hence price discovery. Speed hierarchy not only motivates fast trading competition on less precise information but also renders slower traders more informative. As a result, endogenous speed acquisition in equilibrium affects how information is produced and spread. When information diffusion is characterized by its disclosure level (measured by initial information precision) and dissemination rate (at which heterogeneous information disperses across investors), these factors can have opposite impacts on price discovery. High disclosure leads to more informative fast trading, while fast dissemination crowds out fast traders. Channelled by strategic complementarity of late to early trading, price discovery is improved with disclosure but reduced with dissemination over the short and long run.

Overall, this thesis sheds light into the influence of the most important technological innovation, namely HFT, on liquidity evaporation, speed acquisition, and information disclosure and provides some policy implications for market regulators on HFT.

Contents

List of Figures and Tables

Chapter 1

Introduction

This thesis consists of three essays on high frequency trading in financial markets with respect to speed acquisition, information production, and learning. Within rational expectations equilibrium framework, this thesis examines the impact of HFT on strategic trading behaviour among interacting traders, information efficiency, and market liquidity. The findings provide some insights into the underlying economic mechanisms and policy implications under the scope of the electronic evolution of modern trading environment. This chapter first briefly discusses the backgrounds and motivation for the thesis; then explores the main focus of each chapter of the thesis and outlines the research questions and the main findings in a big picture before going to the details of each chapter.

1.1 Background and Motivation

Development of financial markets is largely driven by technological innovations. Starting with the computerization of the tasks on trading floors, through the introduction of completely electronic markets, to algorithmic trading, trading becomes much faster (see Figure 1-1 cited from Pagnotta and Philippon (2018)). Recently, the rise of high-frequency traders (hereafter HFTs), those investors who rely on sophisticated computer programs to trade over extremely short time periods and gaining tremendous revenue, is

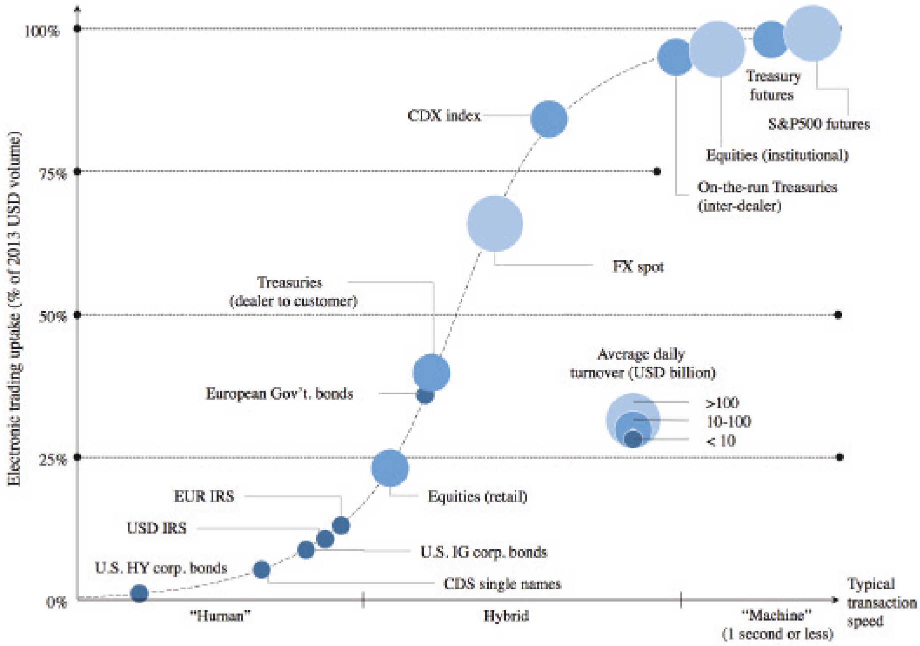


Figure 1-1

Assets classes and trading speeds. Source: Pagnotta and Philippon (2018)) and TABB Group and various sources.

one of the most important changes in financial markets.[1] Over the past two decades, financial markets have made costly investments in trading infrastructure to reduce order execution and communication latencies.[2] In the meantime, high-frequency institutions speed up execution through technology investment on computer hardware, algorithms, and fast connection to exchange servers. Considering information environment, the improvements of fast trading technology have changed how information is produced and spread. On the one hand, they enable some investors to trade in advance on less precise

[1] Ever since the SEC introduced regulations for alternative trading systems, fast trading technology started to take flight. Compare to the situation around 2000s when HFT represented less than 10% trading shares, they accounted for nearly 50% of all equity trades in the United State by 2012. According to Popper (2012), "*HFT is also rapidly growing in Europe and Asia, accounting for approximately 45% of stock trading volume in the European Union, 40% in Japan, and 12% in the rest of Asia as of late 2012*". Although massive investment in HFT has already been seen over the last decades, anecdotal evidence suggests that the profitability of high-frequency traders has decreased in recent years (Biais et al., 2015).

[2] For instance, the trading latency on the New York Stock Exchange (NYSE) dropped from 350 milliseconds in 2007 to 5 milliseconds in 2009. In October 2015, the Bombay Stock Exchange operated the fastest platform in the world, with a round-trip latency of 6 microseconds. Over the eight years (from 2007 to 2015), trading platforms experienced a 50,000-fold speed increase. This process has gone beyond equities and futures to options, bonds, and currency exchange markets.

information before news move stock price closer to its fundamental value.[3] On the other hand, some investors can extrapolate information from earlier trades and anticipate future price movements.

Some researchers view this trading speed evolution as beneficial as extreme fast trading speed reduces the cost for liquidity supplier through i) intensifying competition among liquidity providers, ii) lowering their inventory holding costs, and iii) expediting the search for trading counterparties. As the result, financial market participants have benefited enormously from these trading transformations.

However, economists and policy makers have recently expressed concerns about potentially negative effects of fast trading speed of both speculators and trading platforms. They include: i) *fragile liquidity*: high-frequency trading can increase the adverse selection costs for slow and liquidity traders (Biais et al., 2015, Hoffmann, 2014), ii) *excess speed*: the desire for speed can lead technology suppliers to provide traders faster access to market and enhance a perpetual arms race, increasing market volatility (Budish et al., 2015), and iii) *information production*: quick response to order flow information can reduce the profitability of producing fundamental information and thereby make asset prices less informative (Dugast and Foucault, 2018). Therefore, it is very important to understand how fast trading technologies affect HFT and other market participants and hence price efficiency and market liquidity.

1.1.1. Fragile Liquidity

Despite extensive studies in recent literature, the role of HFT is still not well understood; in particular, there is an increasing concern that HFT would lead to market instabilities or failures such as market breaks and flash crashes. Most recently, the trading week on March 9, 2020 for U.S. market started with loud thud, hitting limit down overnight on future market and triggering circuit breakers within moments of the opening bell on stock market. Within less than ten days, another three circuit breakers lead to vast waves of

[3] For instance, some investors can speed up execution through technology investment on computer hardware, algorithms, and fast connection to exchange servers (e.g., algorithmic and high-frequency trading technology).

liquidity breaks and vanishes in worldwide financial markets.[4] Concerned about the coronavirus outbreak disrupting global supply chains, fuelling fears of recession, and emptying supermarket shelves, financial market liquidity becomes the only precious commodity. With the echoes of the 2010 market instabilities getting louder and louder, extreme volatilities and liquidity breaks do not vanish quickly afterwards. Dubbed the "flash crash" liquidity break happened in 2010[5] puts a spotlight on the rise of high frequency trading, through endogenous liquidity provision, which continues to reverberate through financial markets.[6]

Financial markets are evolving and vulnerable to panicky plunges long before algorithms and fast trading technology emerged. However, high frequency traders do appear to contribute the increasing vulnerability of abrupt dislocation and be catalyst during these extreme price movements and the most recent liquidity breaks. HFTs dominate the market-making once done by humans in trading pits and the bowels of investment banks. HFTs are far more efficient market-makers than human pit traders. However, without market making obligation, HFTs can either compete with designated market makers by providing liquidity or attempt to profit from speculative trades to consume liquidity. This intensifies other liquidity traders' adverse selectin cost (such as

[4] Initially advocated by the Brady Commission following the Black Monday of 1987 and subsequently implemented in 1988, the U.S. circuit breaker was triggered only once in October 27, 1997 before the 2020 coronavirus pandemic. However, the recent history of financial markets, which are now undergoing a dramatic and an inescapable algorithmic trading evolution, witnesses more often episodes of extreme price movements and liquidity breaks within very short periods. Unsurprisingly, high frequency and algorithmic trading have been regarded by market participants and analysts alike as a main potential source of market fragility in the US and other major countries.

[5] The May 6, 2010 U.S. "Flash Crash" was a crash in the U.S. stock market which resulted in a rapid drop in major U.S. indices by between 5 and 6% and a similarly rapid recovery within half an hour. With this incident, the current market structure has revealed serious vulnerabilities that may be exacerbated by HFT.

[6] We have experienced several major crashes after 2010. The "flash crash" was followed by the October 15, 2014 Treasury Bond crash, where the yield on the benchmark 10-year U.S. government bond, dipped 33 basis points to yield 1.86% and then recovered to yield 2.13% by the end of the trading day. Another example is the August 25, 2015 ETF market freeze, when more than a fifth of all U.S. listed exchange traded funds and products were forced to stop trading. Concerned by the coronavirus outbreak, the US stock market triggered circuit breaks four times in two weeks. Besides, we also saw many mini crashes on both future and security markets. Brogaard et al. (2018) use three approaches to identify extreme price movements. The first simply labels all intervals that belong to the 99.9th percentile of 10-second absolute midpoint returns for each stock as extreme price movements. The second method is more sophisticated and accounts for predictable return correlations through time and across firms. The third method for identifying extreme price movements developed by Lee and Mykland (2012) accounts for contemporaneous (local) volatility. With this method Lee and Mykland (2012) identify 45,200 extreme price movements during 2008 and 2009.

designated market makers). They tend to adjust their bids aggressively when market mayhem breaks out. Under those circumstances, even a modest amount of selling or buying could have an outsized impact,[7] leading to fragile liquidity.

1.1.2. Excessive Speed

Speed was once a ticket to fortune in financial markets. Proprietary trading firms, such as Citadel, Virtu, Flow Traders or Jump Trading, bet their own capital and largely serve as market makers, facilitating the buying and selling of shares, currencies and other assets for investors. The extreme fast speed guarantees traders accumulated wealth through millions of times of razor-thin profit from capturing fleeting price differences across major markets. Thus, electronic trading groups have made fortunes by being first to pounce on favourable prices since decades ago. As a result, exchanges have also spent thousands of hundreds of dollars to build fast trading infrastructures, trying to cater to their quickest customers. Exchanges have supplied high-frequency trading shops with ever-improved connections, so that faster traders could profit from the difference between the most up-to-date trading price and older prices offered by slower traders. In doing so, slower traders have had no choice but to spend capital on fast trading technology as well. This process leads to a "latency war" in financial markets as both traders and exchanges speed up their trading speeds.

Concerning about HFT would lead to market instabilities or failures, those who take the view that fast trading is harmful for financial markets and economy have proposed a broad array of policies to slow down trading in financial markets. For instance, some trading platforms (e.g., IEX or Alpha) now use so called "speed bumps" to delay the time elapsed between the moment at which an order arrives to a trading platform and the

[7] Although this is common for both human traders and quants, quant strategies are programmed, quick and on autopilot. When quant strategies start pounding an increasingly thin market, it can cause significant dislocations between buy and sell orders and generate large gains or falls. Flash crashes may be initiated by malfunctioning algorithms or be reinforced by algorithmic trading and HFT, in which case speed and interconnectivity fuel market instability. Particularly during market stress, market-making algorithms may provide less liquidity than human traders. As the result, there is less liquidity when risk aversion spikes occur. This may give rise to order imbalances and sudden price drops.

Table 1-1
Design of Speed Bumps

Exchange	Date	Targets of delay	Length of delay
IEX	Oct-13	All	350 microseconds
NYSE American	Jul-17	All	350 microseconds
Thomson Reuters	Jun-16	All but cancel	0-3 milliseconds
Aequitas NEO	Mar-15	Liquidity takers	3-9 milliseconds
TSX Alpha	Sep-15	Liquidity takers	1-3 milliseconds
Eurex Exchange	Jun-19	Liquidity takers	1 or 3 milliseconds
Chicago Stock Exchange	Proposed	Liquidity takers	350 microseconds
NASDAQ OMX PHLX	Proposed	Liquidity takers	5 microseconds
ICE Futures	Proposed	Liquidity takers	3 microseconds
Interactive Brokers	Proposed	Liquidity takers	10-200 milliseconds

moment in which the order enters the matching engine of the trading platform (see Table 1-1 for exchanges applying speed bumps). To reduce market volatility, Budish et al. (2015) propose to switch from continuous electronic order books (the current market organization in many electronic markets) to frequent batch auctions, in which market clearings take place only at discrete points in time (say, every second). Implicitly, these proposals view that market forces alone are not able to coordinate fast trading and trading speed is excessive for social welfare. Hence, regulatory intervention is needed.

1.1.3. Information Production

Having the abilities to access and process information is the key to well-functioning securities market for promoting price efficiency and facilitating capital information. There is extensive evidence that decentralized and heterogeneous information diffusion influences investors' trading behaviour. [8] At the same time, trading technology improvements have, in turn, changed how information is produced and spread and therefore price discovery is affected. On the one hand, trading technology improvements enable some investors to trade in advance on less precise information before news moves stock price closer to its fundamental value. For instance, some investors can speed up execution through technology investment on computer hardware, algorithms, and fast connection to exchange servers (e.g., algorithmic and high-frequency trading technology). On the other hand, some investors can extrapolate information from earlier trades and anticipate future price movements.

Therefore, modern trading technology improvements can have two opposite effects on information efficiency. On the one hand, being able to trade earlier with less precise public information may crowd out private information production by market participants. This would reduce the profit for producing private information because fast traders have first mover advantage, while processing information takes time. On the other hand, price itself can serve as a coordinating device for investors' belief. Specifically, market equilibrium price can aggregate the heterogenous signals, releasing more information to slow traders. How information diffusion and speed competition affect price efficiency depends on the trade-off between these opposite effects. Understanding the effect of HFTs on market quality is important. More informative prices enhance the efficiency of capital allocation (see Bond et al. (2012) for an excellent survey).

From Chapter 2 to Chapter 4, within rational expectations equilibrium framework, this thesis examines strategic trading behaviour among interacting traders and the impact of HFT on information efficiency and market liquidity. The findings provide some insights into the underlying economic mechanisms and policy implications in modern trading environment. Overall, the goal of this thesis is to shed light into the influence of the most important technological innovation, namely HFT, on liquidity evaporation,

---

[8] Varian (1989) survey institutional investors in the NYSE and find that the majority attributes their most recent trades to discussions with peers or word of mouth communication. Similar evidence is also found for households (Ivković and Weisbenner, 2007), fund managers' portfolio choices (Hong et al., 2004), and on-line communication between retail foreign exchange traders (Heimer and Simon, 2012).

speed acquisition, and information disclosure and provides some policy implications for market regulators on HFT.

1.2 Research Questions and Contributions

The goal of this thesis is to shed light into the influence of a specific technological innovation, namely HFT, on the liquidity evaporation, fin-tech investment and information disclosure in the context of modern financial markets. The main context of this thesis is structured into three chapters. Chapter 2 proposes a nonlinear rational expectations equilibrium model of high-frequency endogenous liquidity provision to explore fragile liquidity. Chapter 3 focuses on the endogenous decisions about both demand for trading speed from investors and supply of trading speed from exchanges to justify the interventions of slowing down trading in financial markets. Chapter 4 considers how the changes in trading technology improvements affect the production and spread of information and studies their impacts on price discovery.

The first phenomenon investigated in this thesis is the sudden liquidity deteriorations in financial markets, such as the “Flash Crash” incident. Chapter 2 provides a conceptual framework to analyse how one of the typical high frequency traders, namely endogenous liquidity provider, influences the market liquidity. This chapter is based on the joint paper “The Microstructure of Endogenous Liquidity Provision” with F. Douglas Foster, Xue-Zhong He and Shen Lin. The May 2010 Flash Crash continues to reverberate through financial markets. It triggered heated debate about the relation between market crashes and high-frequency trading; in particular, through endogenous liquidity provision. Anand and Venkataraman (2016) provide evidence that, without explicit obligations, investors who supply liquidity simultaneously scale back when conditions are unfavourable. This contributes to the covariation in liquidity supply within and across stocks. Hence, a growing reliance on high frequency traders (HFTs) acting as liquidity providers has raised concerns with market regulators and investors. Chapter 2 thus focuses on the following research questions 1a-1b:

Research Question 1a: *Does a possible withdrawal from liquidity provision by HFTs result in illiquidity that destabilizes markets?*
Research Question 1b: *Does this liquidity withdrawal introduce and intensify market fragility during turbulent times?*

Chapter 2 finds that with fast trading speed and private information, high-frequency traders can either compete with designated market makers (DMMs) by providing liquidity or attempt to profit from speculative trades that consume liquidity. The risk from this endogenous liquidity provision, coupled with limits to participation by DMMs, intensifies the adverse selection faced by DMMs. This can generate a gap between liquidity supply from DMMs and liquidity demand by informed traders. As a result, endogenous liquidity provision produces fragile liquidity, with the possibility of market breaks when high-frequency traders switch from liquidity provision to liquidity consumption based on unexpected information signals.

Based on the microstructure of high frequency endogenous liquidity provision, Chapter 3 analyzes trading speed from more general perspective. Chapter 3 is based on the joint paper "The Fast and the Furious: Exchange Latency and Ever-fast Trading" with Xue-Zhong He and Xuan Zhou. Speed was once a ticket to fortune in financial markets. The extreme fast speed guarantees proprietary trading firms[9] to accumulate wealth through millions of razor-thin profits from capturing fleeting price differences across markets. Meanwhile, exchanges have spent hundreds of thousands dollars to build fast trading infrastructures to reduce order execution and communication latencies for

[9] Proprietary trading firms, such as Citadel, Virtu, Flow Traders or Jump Trading, bet their own capital and largely serve as market makers, facilitating the buying and selling of shares, currencies, and other assets for investors. These high-frequency institutions speed up execution through technology investment on computer hardware, algorithms, and fast connection to exchange servers. They have made fortunes by being first to pounce on favourable prices since decades ago.

catering to their quickest customers.[10] With ever-improved connections, exchanges have supplied high-frequency trading (HFT) to faster traders so that they can profit from price discrepancy between the most up-to-date (and more informative) prices and older (and less informative) prices offered by slower traders. In doing so, slower traders have had no choice but to invest in fast trading technology as well. This leads to a "latency war" in financial markets as both traders and exchanges speed up their trading. These magnificent growths in fast trading and fast exchange have raised concerns from researchers, market regulators and investors. However, the much focus of the existing literature is on either the influence of fast trading or impact of fast exchange, but not on the interaction between exchange latency and ever-fast trading speed. Chapter 3 thus focuses on the following research questions 2a-2e:

Research Question 2a: *Why should we expect trading speed to be excessive?*
Research Question 2b: *How much speed technology should investors and trading platform acquire?*
Research Question 2c: *From a social welfare viewpoint, why market forces fail to coordinate at an optimal speed level?*
Research Question 2d: *What are the implications for "price discovery" (information aggregation and price informativeness) and social cost?*
Research Question 2e: *Can regulatory interventions help to mitigate these market failures?*

Chapter 3 theoretically examines how technological innovations drive fast trading investment for both speculators and exchanges and their impact on market. The negative externality of the speed acquisition from fast speculators can result in excessive investment, which is intensified as speculators' speed technology advances. As

[10] Financial markets have made costly investments in trading infrastructure to reduce order execution and communication latencies. For instance, the trading latency on the NYSE dropped from 350 milliseconds in 2007 to 5 milliseconds in 2009. In October 2015, the Bombay Stock Exchange operated the fastest platform in the world, with a round-trip latency of 6 microseconds. Within a time window of eight years, trading platforms experienced a 50,000-fold speed increase. This process has gone beyond equities and futures markets to options, bonds, and currency exchange markets.

exchange's speed technology advances, faster exchange makes faster speculators more concentrated; that is, higher exchange speed shrinks market fraction of fast speculators but stimulates their optimal trading speed. As the result, market liquidity is improved but price discovery is reduced. Policy makers aiming to balance price discovery and deadweight loss from costly speed investment may lead to a mismatch between the desired exchange speed for policy makers and the optimal speed supplied by exchange, echoing the concerns of market regulations about market failure on speed arms race.

While the previous two chapters focus on the "matching speed" of high frequency traders, i.e., the speed at which orders are processed and filled by trading platforms, the "information speed", i.e., the speed at which traders receive information is also equivalently important to HFT. Chapter 4 presents the influence of information diffusion on price efficiency when information diffusion is characterized by its disclosure level (measured by initial information precision) and dissemination rate (at which heterogeneous information disperses across investors). This chapter is based on the joint paper "Information Diffusion and Speed Competition" with Xue-Zhong He. Modern trading technologies have fundamentally changed how information is disseminated in financial markets. Motivated by the changes in information structure, Chapter 4 studies the following research questions 3a-3c:

Research Question 3a: *How does the evolution in speed technologies influence information production and strategic trading among different investors?*
Research Question 3b: *How does the dramatic changes in information diffusion process affect price discovery?*
Research Question 3c: *What are the novel implications for information management?*

Chapter 4 examines information diffusion and speed competition. Speed hierarchy not only motivates fast trading competition on less precise information but also renders slower traders more informative. As a result, endogenous speed acquisition in equilibrium affects how information is produced and spread. When information diffusion is characterized by its disclosure level and dissemination rate, these factors can have opposite impacts on price discovery. An increase in disclosure leads to more informative fast trading, while fast dissemination crowds out fast traders. Channeled by strategic complementarity of late to early trading, price discovery is improved with disclosure but reduced with dissemination over the short and long run.

Finally, Chapter 5 summarizes the findings of the thesis and discusses implications for regulators and policy makers, market operators and investors. It further outlines the interface with other areas in finance and machine learning research and provides a guideline for future research. All the proofs and further discussions of the results in Chapters 2-4 are collected in the Appendix.

In sum, this thesis sheds light into the influence of the most important technological innovation, namely HFT, on liquidity evaporation, speed acquisition, and information disclosure and provides some policy implications for market regulators. The main contributions of the thesis are threefold as detailed in the following three chapters. Firstly, we have proposed several theoretical frameworks to examine the most concerned research issues, such as liquidity breakers, HFT arms race, speed bumps, information diffusion, in the fields of high frequency trading. Secondly, our models generate several empirical and policy implications for policy maker, investors and institutions in financial markets. These implications involves, among many others, how to regulate trading platform trading speed, how company should manage its information under high frequency trading environment, what is the trade-off between benefit from HFTs and possible market instabilities. Lastly, this thesis also novelty applies reinforcement learning to solve a nonlinear Bayesian updating problems, which contributes theoretical market microstructure literature from methodology perspective and provides potential solution for exploring nonlinear rational expectations equilibrium.

Chapter 2

The Microstructure of Endogenous Liquidity Provision

On March 9, 2020 U.S. financial markets opened with a thud, hitting limit down trading halts overnight on future markets and triggering circuit breakers within moments of the stock market opening bell. Investors were concerned with the coronavirus outbreak, disruption to global supply chains, fears of a recession, and empty supermarket shelves making liquidity a precious commodity in financial markets. Extreme volatility and liquidity breaks did not vanish; this was followed by another three circuit breakers triggered in less than ten days. This experience illustrates how liquidity breaks are buffeting financial markets, with the echoes of the "flash crash" of 2010.[11] The flash crash experience focus attention on high frequency trading; in particular through endogenous liquidity provision. This experience continues to influence market commentary.

Initially advocated by the Brady Commission following the Black Monday of 1987 and subsequently implemented in 1988, the U.S. circuit breaker system was triggered only once (October 27, 1997) before the 2020 coronavirus pandemic. However, we have

[11] At 2:32pm on May 6, 2010, US equities suddenly and mysteriously careened lower. In just 36 minutes the S&P 500 fell more than 8% before a powerful rebounding.

recently seen more episodes of extreme price movements and liquidity breaks within very short periods. Unsurprisingly, high frequency and algorithmic trading have been regarded by market participants and analysts alike as a significant source of fragility in electronic markets, both in the US and other major markets.

Financial markets are evolving and have been vulnerable to plunges long before algorithmic trading and fast trading technologies. Hence, there area there are many other potential explanations for the increased the frequency of liquidity breaks. It difficult to diagnose fully the myriad possible triggers and drivers. High frequency traders (HFTs afterwards) are consistently thought to contribute to increased vulnerability of abrupt dislocation and therefore a catalyst for extreme price movements and liquidity breaks.

HFTs dominate the market-making once done by humans on floors and in trading pits. HFTs have brought a lot of efficiencies in facilitating the exchange and clearly of financial products. However, without market making obligations, HFTs can either compete with designated market makers (DMMs afterwards) by providing (making) liquidity or attempt to profit from speculative trades, thereby consuming (making) liquidity. Anand and Venkataraman (2016) provide evidence that, without explicit obligations, investors who supply liquidity simultaneously scale back when market conditions are unfavourable.[12] Van Kervel and Menkveld (2019) also show that HFTs initially lean against large institutional orders that execute through a series of smaller orders but eventually change direction and take position in the same direction of the most informed institutional orders. HFT's tend to adjust their bids more aggressively when

[12] With highly profitable market making and no obligation to provide liquidity, HFTs have incentive to either supply or demand liquidity, depending on their beliefs. While the literature generally finds that HFTs act as liquidity providers, there is evidence that HFTs reduce their liquidity commitments when market conditions become unfavourable (Anand and Venkataraman, 2016, Bessembinder et al., 2016, Korajczyk and Murphy, 2019, Raman et al., 2014).

markets are volatile. In turbulent markets this means that even a modest amount of selling or buying might have an outsized impact.[13]

These events have created and revived a series of debates about the relation between market breaks and HFTs, in particular through HFTs endogenous liquidity provision. A growing reliance on HFTs acting as liquidity providers has raised concerns with market regulators and investors.[14] Market regulators have paid great attention to liquidity breaks; brought on by the greater incidences of sudden, unexpected, and significant price drops in one or more instruments or markets. For example, will a possible withdrawal from liquidity provision by HFTs result in illiquidity that may destabilize markets? Will this liquidity withdrawal introduce and intensify market fragility during turbulent times? This chapter proposes a rational expectations equilibrium model of high-frequency endogenous liquidity provision to explore these questions on fragile liquidity.

To explore these issues more formally, we introduce a two-dimensional model of liquidity provision based on orders from a DMM and orders from a high-frequency endogenous liquidity provider (ELP afterwards). In our setting the ELP is also an HFT; they have a technological advantage that allows them to choose to either provide or consume liquidity based on their private information. In addition to the ELP and DMM, there is an informed trader. To characterize information asymmetry and learning, we consider multiple sources of randomness in the payoff to the risky asset. Specifically, the informed trader and ELP have both distinct and common views about the payoff of the risky asset, though there is a noise shock to ELP's private information. This captures HFTs' information speed. It also allows for both fundamental (such as coronavirus outbreak), and algorithmic-driven trade to influence HFTs' liquidity provision.

---

[13] This is a common issue for both human traders and quants, but quant strategies are programmed, quick and on autopilot. When quants start pounding an increasingly thin market, they can cause dislocations between buy and sell orders and produce big gains or falls. Therefore, flash crashes may be initiated by malfunctioning algorithms or be reinforced by algorithmic trading and HFTs, in which case speed and interconnectivity fuel market instability. Particularly during times of market stress, market-making algorithms may provide less liquidity than human traders who are aimed to maintain a certain level of market liquidity. As the result, there is less liquidity during market stress when risk aversion spikes. This may give rise to order imbalances and sudden price drops.

[14] "The issue is whether the firms that effectively act as market makers during normal times should have obligations to support the market in reasonable way in tough times", Mary L. Shapiro, then SEC Chairperson, in a speech to the Economic Club of New York on September 7, 2010.

Based on this information structure, we build a rational expectations equilibrium model of fragile liquidity with a high-frequency ELP. Because of HFTs' endogenous switching between liquidity provision and consumption, the resulting equilibrium is nonlinear. Due to the nonlinearity and complexity in the DMM's posterior beliefs, we introduce a reinforcement learning approach to obtain nonlinear demand and price functions. This approach brings new evidence and insights from reinforcement learning techniques to theoretical microstructure research, thereby synthesizing these two streams of literature. Our nonlinear equilibrium delivers several main results. The price function when the ELP consumes liquidity (acts as a liquidity taker) is concave with respect to aggregate order flow. It has a much steeper slope than under the regime where the ELP provides liquidity (acts as a liquidity maker). This implies that the liquidity under the taker regime is more fragile (can change more dramatically) than under the maker regime.

As a result, HFTs' endogenous liquidity provision and DMMs' limited market participation produces fragile liquidity, with the possibility of market breaks when HFTs switch from liquidity provision to liquidity consumption based on unexpected information signals. In practice, the financial media may report some of these signals by alluding to unusual trading patterns, with the culprits typically being either erratic algorithms or fickle trades; examples include the 2010 "Flash Crash" and the most recent 2020 "Associated British Foods" trading error.[15] However, these signals can be also from more fundamental events, such as the 2015 crashes induced by concerns about the slowing Chinese economic development and the 2020 coronavirus pandemic that has led to widespread health, social, and economic disruption.

Absent an obligation to supply liquidity, HFTs can and will switch between supplying and consuming liquidity. E.g. HFTs may make liquidity when their private information innovation is relatively small (within a threshold range) but take liquidity as speculators when their private information shock is relatively large (outside of a threshold range). The resulting equilibrium is characterized by an endogenous cut-off strategy with two

[15] On 16 March 2020, the London stock exchange announced that in view of the erroneous price executed in the ABF at 09:07:11, under Rule 2120 the Exchange has cancelled the execution . the security is open for regular trading as standard. During that day, around 09:07 am, there are total 10 trades with price 0.01. This trading error reduce the price from 1500 to just 0.01 in only one second.

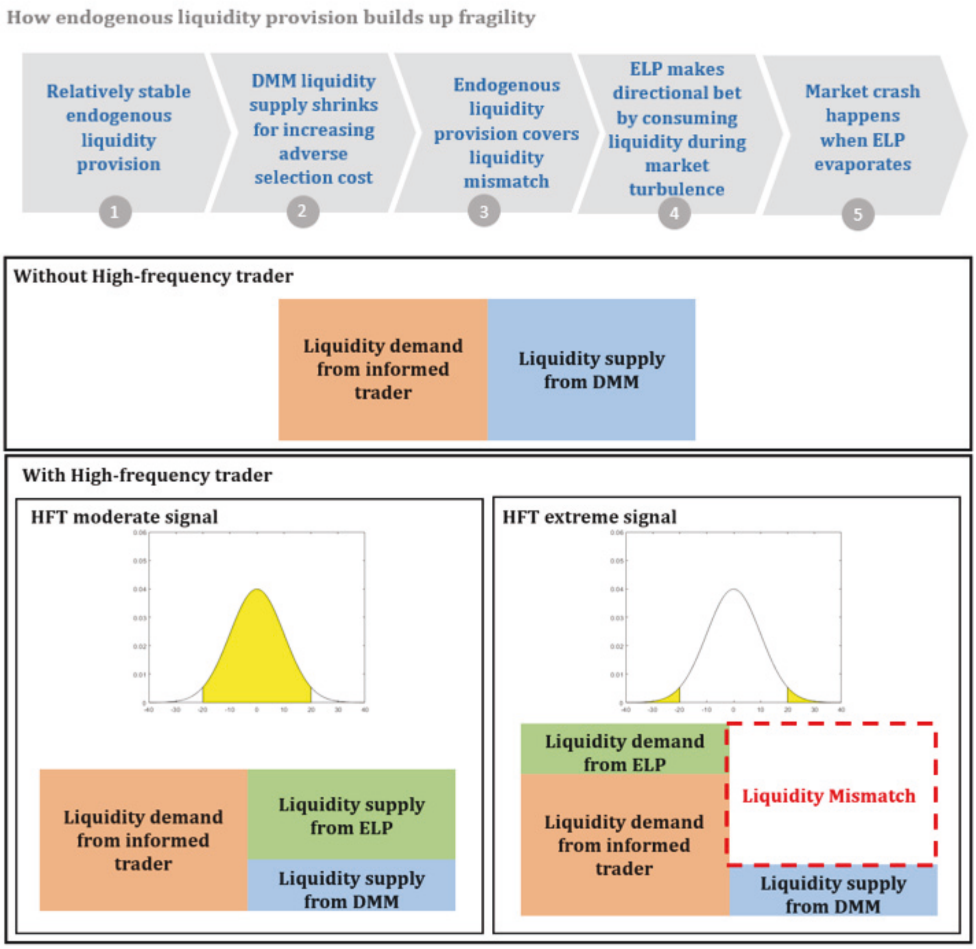


Figure 2-1

The mechanism of how endogenous liquidity provision builds up fragility. The yellow shaded areas in the probability densities represent possible realizations of small (on the left) and large (on the right) noisy information about the asset payoff for the ELP in normal and turbulent times, respectively. We depict instances of liquidity demand (from the informed), liquidity supply (from the DMM) and both liquidity demand and supply (from the ELP). The top panel illustrates the setting without a high-frequency ELP, the lower left panels are for normal times, and the lower right panels are for turbulent times. Without the ELP, the liquidity demand from the informed trading matches the liquidity supply from the DMM in both normal and turbulent times. The two lower panels illustrate the effect of high-frequency ELP on liquidity. During normal times (the lower left panel), the ELP reduces the liquidity supply from the DMM but covers the liquidity mismatch between the DMM's liquidity supply and informed trader's liquidity demand. During turbulent times (the lower right panel), the ELP takes liquidity, thereby exacerbating the liquidity mismatch.

regimes: making or taking liquidity. Apart from uncertain asset values, this additional risk of HFTs' switching roles, coupled with limited participation of DMMs (resulting from trading technology and trading automation), intensifies the adverse selection faced by DMMs. Therefore, this structure shrinks significantly the liquidity supply from the DMM, generating a gap between DMMs' liquidity supply and informed traders' liquidity demand,. In addition, this structure expands aggregate liquidity demand during market turmoil (i.e. occasions when the signal received by the ELP is much different than they expect). This induces a strategic complementarity in liquidity mismatches: traders demand more liquidity when the aggregate liquidity supply falls. Instead of providing liquidity normally, the ELP now consumes liquidity, which can intensify the adverse selection faced by DMMs. This then serves as a conduit for a loss of liquidity and possible market break when private signals are much different than expected. That is, the ELP propagates initial shocks, which then induce a liquidity mismatch and lead to an intensified loss of liquidity. This mechanism is showed in Figure 2-1.

Our model delivers several implications. First, price changes becomes more sensitive to small trading volume, but less so to larger trading volume, a nonlinear price impact documented in the empirical literature. Second, a relatively small shock in one of the state variables (fundamental or non-fundamental) can lead to a disproportionately large change in liquidity, generating fragile liquidity. Third, fragile liquidity could further lead to market breaks (so that equilibrium price function is discontinuous in one of its continuous arguments). Fourth, the interdependence of the uncertainty and liquidity provision decision generates a feedback loop and creates a multiplier that amplifies the initial effect of an exogenous parameter change in equilibrium, an amplification effect. Lastly, a linear approximation might capture the same underlying mechanism but fail to capture the consequences of nonlinearity and overstate liquidity when HFTs take liquidity.

Our economy of endogenous liquidity provision also generates several implications for market regulators. First, as the information for the informed trader and the ELP becomes more similar, liquidity under the taker regime decreases while liquidity under the maker regime increases, intensifying liquidity fragility. This should be of concern for market regulators and investors, especially as markets become more integrated. Second, decreasing in ELP's information speed or increasing in information latency in a segmented market can improve market liquidity in both taker and maker regimes. This

Table 2-1

Summary for main results

		Probability		CB probability	DMM adverse selection
		Taking liquidity	Systematic liquidity withdrew		
Section 2.2	Limited participation $\tau = 1 \rightarrow 0$	↘			↗
	Information similarity between ELP and informed trader	↘			↗
	ELP information noise	↘			↘
Section 2.3	Discretionary liquidity trader split → concentrate	↘		↗	↘
Section 2.4	Information similarity between different ELPs	↘	↘		↗

suggests that there may be benefits to slow ELPs, echoing recent concern from market regulator about the superior trading capabilities of HFTs.

To examine further the effects of circuit breakers and competition among HFTs, we extend the baseline model in two directions. In the first extension, circuit breakers may affect the equilibrium through the strategic trading decisions of a discretionary liquidity trader who either trades in both periods or concentrate their trades in the early period. We show that the presence of ELP, encourages the discretionary liquidity trader to concentrate trading in the early period. This increases both the probability of trigging a circuit breaker and trading cost of the discretionary trader, which decreases market liquidity. Hence, in this setting algorithmic trading can lead to more volatile market outcomes and impose adverse selection on other market participants. However, with the an active ELP a circuit breaker can reduce the probability of liquidity breaks and thereby improve market liquidity. I.e. when introducing circuit breaker into financial market, the discretionary liquidity trader might concentrate her order to the early period to mitigate the cost of not being able to trade in late period. Hence, the aggregate liquidity trading in the first period rises, which in turn lowers the ELP's chance to take liquidity. This leads to improved market liquidity.

In the second extension, we apply reinforcement learning methods to explore the influence of competition among different types of ELPs with correlated information. We show that an increase in information correlation among ELPs reduces the risk of systematic liquidity withdrew but intensifies DMM's adverse selection cost, accelerating the loss of liquidity. We have seen increasing presence of "new" algorithmic market-makers, competing for liquidity provision with not only traditional DMMs but also high-frequency trading institutions. Most of the time this competition benefits financial markets through faster and better liquidity provision. However, when markets are turbulent, not only high-frequency traders but also these "new" algorithmic market-makers automatically limit their quoted prices and order size. Therefore, increasing endogenous liquidity provision may exacerbate liquidity fragility. We summarize the above discussions and main results in Table 2-1.

Related Literature

Financial market development is driven by technological innovation. This includes examples such as the computerization of tasks on trading floors, the introduction of completely electronic markets, and algorithmic trading (AT); all are examples of innovation that allows trading to become much faster. The recent rise of HFTs, that some investors rely on sophisticated computer programs for trading over extremely short time periods, is a major change in financial market (see Biais and Foucault (2014), Foucault and Moinas (2018), Menkveld (2016) for excellent surveys on both theoretical and empirical research about HFTs). Speed of trading is important for HFTs to profit from their strategies, such as directional trading (through either high-frequency arbitrage or front- and back-running) on very short-lived signals and high-frequency market making.

Prior theoretical work on HFTs mainly focuses on their role in either liquidity consuming perspective (i.e., directional trading) or liquidity providing perspective (i.e., high-frequency market making), but not both. For example, Aït-Sahalia and Saglam (2013) analyze theoretically the optimal pricing behaviour of a high frequency market maker (see also Jovanovic and Menkveld (2016)). They characterize the market maker's optimal pricing policy and show that his bid-ask spread is smaller on average when he is faster. Intuitively, trading faster enables the market maker to revise his quotes more frequently and thereby to better manage his inventory (e.g., to shade his bid more quickly after accumulating a long position). However, several theoretical papers have analyzed how the acceleration of market access and role of HFTs as speculators to news (broadly defined, i.e., including quote updates for instance) affect adverse selection costs in financial markets (see, for instance, Biais et al. (2015) and Foucault et al. (2016)).

While these studies are important in providing our basic understanding of how traders' speed can affect costs of liquidity provision (from perspectives of (i) search, (ii) adverse selection, (iii) inventory risk, and (iv) competition among liquidity providers), focusing solely on only one of the roles understates the fundamental merits of a high frequency trading mechanism itself. In reality, HFTs do not necessarily specialize in using one trading strategy (market making, directional trading, or arbitrage) and might opportunistically use insights from all perspectives (although empirical evidence suggests that there is some specialization, see Boehmer et al. (2018)). At the same time, recent empirical evidence (Anand and Venkataraman (2016)) shows that HFTs may switch between providing and consuming liquidity, depending on market conditions. We fill these gaps by characterizing both market-making and directional trading through high-frequency endogenous liquidity provision.

By developing a model to capture HFTs' endogenous liquidity provision through a cut-off strategy based on an endogenously determined threshold value, our work complements these studies. Specifically, it provides a better understanding the effects of endogenous liquidity provision on overall market quality from theoretical perspective (see Christensen et al. (2014); Kirilenko et al. (2017) and Brogaard et al. (2018) for related empirical evidence). Budish et al. (2015), Menkveld and Zoican (2017) and Haas et al. (2019) also consider both high-frequency market makers and liquidity consuming high-frequency speculators in financial markets. However, in these environments, a high-frequency trader that succeeds in posting his quotes initially becomes a high-frequency market-maker; while other HFTs act as quote snipers, poised to trade on stale quotes. There is no switching between being a high-frequency maker or taker of liquidity. Incorporation of the endogenously switching behaviour differentiates our chapter from these studies.

In this spirit, our model is closely related to the empirical analysis from Van Kervel and Menkveld (2019), who examine whether HFTs lean against large institutional orders that execute through a series of child orders. The possibility of liquidity breaks when HFTs switch from liquidity provision to liquidity consumption on the basis of unexpected information signals, as we see in our chapter, is qualitatively similar to what they find empirically. Their results show that HFTs initially lean against these orders but eventually change direction and take position in the same direction of the most informed institutional orders. Connecting to our mechanism, initially, few and small institutional orders imply weak signals for HFTs so that they provide liquidity. However, seeing more institutional orders indicates an extreme signal and HFTs start to speculate on this insight. When deciding trade intensity, HFTs seem to trade-off between higher speculative profit and higher risk of being detected and losing the advantage of their insight into the order flow.

Our model is also closely related to three additional theoretical literature on high-frequency trading (Bongaerts and Van Achter, 2020, Hoffmann, 2014, Li et al., 2019) but differs from them in the following perspectives. First, Bongaerts and Van Achter (2020) also find that, with low levels of informed trading, the speed technology of HFTs improves market liquidity and welfare, reflected by lower transaction cost for end-users. However, an asymmetric information problem arises when HFTs retract in anticipation of toxic order flow. Yet, the model mechanism differs at several important points. In the Bongaerts and Van Achter (2020) model, liquidity providers make an endogenous choice in technology adoption to become either fast, informed or both before trading rounds, there is no switching behaviour based on signals during trading rounds. Second, similar to Hoffmann (2014), HFTs in Bongaerts and Van Achter (2020) can revise the order after observing the signals. More specifically, fast traders (FTs) are able to revise their limit orders after news arrivals, but only in case the next trader is a slow trader (ST). Different from Hoffmann (2014), we focus on different sources of adverse selection cost. In Hoffmann (2014), the presence of FTs induces slow traders to strategically submit limit orders with a lower execution probability, thereby reducing trade. In our model, it is the HFT's switching behaviour that indicates the extreme signals and significantly intensifies DMM's adverse selection cost. Last, Li et al. (2019) provide the first model representing the behaviour of algorithmic traders that are slower than HFTs.[16] They focus on different liquidity provider in different situations; while we focus on HFT's endogenous liquidity provision decision and its implication on market dynamics.

While our model focuses on endogenous liquidity provision, its impact on fragile liquidity and market breaks can be more broadly applicable. Besides endogenous liquidity withdrawals, other types of market friction such as market-wide trading halts, funding liquidity constrains and price limits all have similar effects on the illusion of liquidity. This in turn results in liquidity breaks, extreme price movement, and amplified volatility. Hence, we extend our framework to examine the impact of market interruptions, such as the circuit breakers. Prior theoretical work circuit breakers focuses on their role in reducing excess volatility and restore orderly trading in the presence of market imperfections such as limited participation, information asymmetry and market power(Greenwald and Stein, 1991, Subrahmanyam, 1994). Recently, Chen et al. (2017b) developing a model to capture investors' first-order trading needs, reflecting the cost of circuit breakers, in welfare, price level and volatility, in a benchmark setting without imperfections. Our chapter complements these studies by taking the endogenous liquidity provision into consideration.

There are extensive empirical studies on the impact of high frequency trading in general. The growing literature focuses on three aspects of these technological changes; (i) what information high-frequency traders use to trade (Hu et al., 2016, Van Kervel and Menkveld, 2019); (ii) how high-frequency trading incorporates information (Brogaard et al., 2014, O'hara et al., 2014); and (iii) what impact of high-frequency trading on market quality (Budish et al., 2015, Weller, 2016, Yang and Zhu, 2020). In summary, our chapter provides a novel theoretical framework to examine the potential influence of high frequency endogenous liquidity provision on market liquidity.

Lastly, this chapter also contributes to the recent applications of machine learning in economics and finance. Varian (2014), Abadie and Kasy (2017), and Mullainathan and Spiess (2017) provide excellent discussions of how machine learning can be applied to

[16] The results show that under continuous pricing, algorithmic traders dominate liquidity provision by using aggressive limit orders to stimulate HFTs' market orders. Under discrete pricing, HFTs dominate liquidity provision when the bid-ask spread is binding at one tick.

analyze economic problems involving big data. Recent empirical studies further apply machine learning to finance literature.[17] We complement this empirical literature by applying machine learning in solving nonlinear equilibrium models theoretically. When models have increasingly diverse agents and more complex market structures, such as in this chapter, market equilibrium becomes nonlinear and hence difficult to solve. While being able to characterize precisely how innovative forms of trading affects price and liquidity dynamics is important for investors and regulators, this need for richer models poses significant technical challenges. We demonstrate how reinforcement learning techniques, combined with model structures from microstructure theory, can play an important role in more intricate model structures and provide insights to important market microstructure questions. Unlike prior research (e.g. Bernardo and Judd (2000)), we do not rely on parameterizations of price adjustment rules, asset demand, or use of the projection method. We provide a solution technique for more complicated interactions between traders in continuous state space (instead of discrete state space, e.g. Breugem and Buss (2019)), which can be generalized to study rational expectations models with other market frictions.

The chapter is organized as follows. Section 2.1 describes the economy model with endogenous liquidity provision and illustrates the reinforcement learning approaches intuitively. Section 2.2 presents the main results of a baseline model on limited market participation, endogenous liquidity provision and fragile liquidity, resulting from the nonlinear rational expectation equilibrium, and provides some policy implications. Section 2.3 extends to an intertemporal model to examine the joint impact of endogenous liquidity provision and circuit breaker. Section 2.4 extends the model to examine the effect of the competition of multiple ELPs. Section 2.5 concludes. All the proofs and additional discussions are provided in the Appendix.

[17] Chinco et al. (2019) apply LASSO techniques to make 1-minute ahead equity return forecasts. Rossi (2018) uses boosted regression trees to forecast stock returns and volatility. Krauss et al. (2017) use machine learning for statistical arbitrage with the S&P 500. De Prado (2018) provide extensive analyses of financial machine learning techniques and applications. Gu et al. (2018) apply multiple machine learning regression algorithms in asset pricing and find that non-linear regression methods can give rise to better R-squares than standard econometric models.

2.1 An Economy with Endogenous Liquidity Provision

This section describes a rational expectations equilibrium model of liquidity with a high-frequency endogenous liquidity provider (ELP). A distinct feature of this model is that high-frequency trader's endogenous liquidity provision operates through a cut-off strategy where a threshold of liquidity provision is determined endogenously. We show that the withdrawal of the ELP's liquidity can have significant effects on market clearing. This possible loss of liquidity from the ELP (replaced with a concomitant use of liquidity by the ELP) results in fragile liquidity.

2.1.1. Agents, Assets, and Information

There are three risk-neutral traders in the economy: the ELP who can choose between market orders (demanding or taking liquidity) and limit orders (providing or making liquidity); an informed trader who submits market orders to consume liquidity; and a designated market maker (DMM) who submits limit orders to provide liquidity. Unlike the ELP, the DMM is required to post limit orders, hence their designation. Additionally, there are liquidity traders who randomly submit market orders and thereby take liquidity. The existence of liquidity traders prevents both the ELP and informed trader's private information from being revealed fully in equilibrium.

We consider a one-period model with two dates ($t = 1, 2$), which is descried in Table 2-2. At date 1, all traders submit orders based on their information and trades are cleared. At date 2, assets pay off and all traders consume. There is a risky asset, which is in zero net supply. Let $\tilde{p}$ denote the price of the risky security.[18] The payoff or value of $\tilde{V}$ of the risky asset has multiple random components,

$$\tilde{V} = \underbrace{\tilde{\delta}_f + \tilde{\delta}_d}_{\substack{\text{distinct views between} \\ \text{ELP and informed trader}}} + \underbrace{\tilde{\delta}_0}_{\substack{\text{common views between} \\ \text{ELP and informed trader}}}, \tag{2-1}$$

where $\tilde{\delta}_i \sim N(0, \sigma_i^2)$ are mutually independent draws from normal distributions. Specifically, at date 1 before trading, the informed trader knows the realization of $\tilde{s} = \tilde{\delta}_0 + \tilde{\delta}_d$, while

[18] Throughout the paper, a tilde signifies a random variable.

Table 2-2

The timeline and market participants in Section 3 and Section 4

<table>
<tr><td colspan="3" rowspan="2"></td><td rowspan="2">$t = 0$</td><td colspan="2">$t = 1$</td><td rowspan="2">$t = 1'$</td><td rowspan="2">$t = 2$</td></tr>
<tr><td>$\lvert\tilde{\theta}\rvert > \theta^*$</td><td>$\lvert\tilde{\theta}\rvert \leq \theta^*$</td></tr>
<tr><td rowspan="4">Section 2.2 & Section 2.3</td><td colspan="2">Informed trader</td><td>Observe $\tilde{s}$</td><td colspan="2">Submit $x_I(\tilde{s})$</td><td>…</td><td>Receive π_I</td></tr>
<tr><td colspan="2">Noise trader</td><td>…</td><td colspan="2">Submit $\tilde{z}_1$</td><td>…</td><td>…</td></tr>
<tr><td colspan="2">ELP</td><td>Observe $\tilde{\theta}$</td><td>Submit
$x_T(\tilde{\theta})$</td><td>Submit
$x_M(\tilde{\theta};\, \tilde{u}_1, \tilde{p})$</td><td>…</td><td>Receive π_E</td></tr>
<tr><td colspan="2">DMM
(labelled as DMM_e in Section 2.3)</td><td>…</td><td colspan="2">Set $\tilde{p}_1$</td><td>…</td><td>Observe $\tilde{v}$</td></tr>
<tr><td colspan="3"></td><td colspan="5"></td></tr>
<tr><td rowspan="4">Only in Section 2.3</td><td colspan="2">Noise trader at date 1′</td><td>…</td><td colspan="2">…</td><td>Submit $\tilde{z}_{1'}$</td><td>…</td></tr>
<tr><td colspan="2">DMM at late period
(labelled as DMM_l in Section 2.3)</td><td>…</td><td colspan="2">…</td><td>Set $\tilde{p}_{1'}$</td><td>Observe $\tilde{v}$</td></tr>
<tr><td rowspan="2">Discretionary
liquidity trader</td><td>Split</td><td rowspan="2">…</td><td colspan="2">$\tilde{l}$</td><td>$\tilde{l}$</td><td rowspan="2">…</td></tr>
<tr><td>Concentrate</td><td colspan="2">$2\tilde{l}$</td><td>…</td></tr>
</table>

the ELP observes private information described by the signal $\tilde{\theta} = \tilde{\delta}_0 + \tilde{\delta}_f + \tilde{\varepsilon}_\theta$. The innovation $\tilde{\varepsilon}_\theta \sim N(0, \sigma^2_{\varepsilon,\theta})$ measures the ELP's information precision or speed, which can be influenced by the degree of technology investment.[19]

Given this structure, the signal $\tilde{\delta}_0$ received by both the informed trader and ELP represents a common information component and describes information similarity; whereas $\tilde{\delta}_f$ and $\tilde{\delta}_d$ become trader-specific information and characterize heterogeneous information between the informed trader and ELP.[20] With the increasingly relative importance of common shock (i.e., higher σ^2_0), the high-frequency ELP might observe a signal that is closely related to the signal seen by the informed trader. Hence, this information structure allows us to explore the impact of increasing information commonality (for example, due to highly correlated market[21]) between high-frequency institutions and investors such as mutual funds.

2.1.2. Preference and Endogenous Liquidity Provision

Using the structure outlined above risk-neutral traders trade at date 1 to maximize their payoffs subject to quadratic trading costs, as in Banerjee et al. (2018),

$$\max_{x_i} E\left[x_i(\tilde{V} - \tilde{p})|I_i\right] - \frac{1}{2\gamma}x_i^2, \tag{2-2}$$

[19] Although we do not explicitly provide micro-foundation for ELP's private information, it fits to a cross-learning framework, consistent with Cespa and Foucault (2014) and Goldstein et al. (2014). When the same or similar assets trade in multiple venues, this information structure allows high-frequency ELP to "back-run" (inferring fundamental information from order flows) in faster markets and to "front-run" (anticipating the order flows and fundamental value) in slower markets. Therefore, the noise $\tilde{\varepsilon}_\theta \sim N(0, \sigma^2_{\varepsilon,\theta})$ in the ELP's private signal measures how fast the ELP can process order flow information from faster markets (we refer to this as the *information speed*).

[20] Empirical literature and market practice suggest HFTs (such as ELPs) and active mutual fund investors (e.g. the informed trader in our model) focus on different types of information. HFTs rely mainly on order flow information through back running, front running and cross-market inference. Active mutual fund investors rely on investment processes that consider fundamental and other research as well as due diligence. Hence, the information components $\tilde{\delta}_f$ and $\tilde{\delta}_d$ in equation (2-1) can be viewed as different dimensions of the information about the risky asset payoff that characterize heterogeneous information between the informed trader and ELP.

[21] When markets become more related, information traded and disseminated on different markets might also become more similar. Therefore, the high-frequency ELP, who usually trades on order flow information through back running, front running, and cross-market inference, might observe information that is highly related to signals known to informed traders.

where γ measures the reciprocal of trading cost, I_i is the information set for trader i and x_i is the demand of trader $i\epsilon\{E,I,D\}$, representing ELP, informed trader, and DMM, respectively. There is also a liquidity (noise) trader with supply $\tilde{z}\sim N(0,\sigma_z^2)$ for the risky asset, which is independent of all other variables in the economy.

Based on any relevant signal, a liquidity maker submits a price-contingent demand schedule, whereas a liquidity taker submits a market order. Specifically, the informed trader submits a market order $x_I(\tilde{s})$ after obtaining private signal $\tilde{s}$ and DMM submits a limit order $x_D(\tilde{u},\tilde{p})$ conditional on price $\tilde{p}$ after observing overall market order flow $\tilde{u}$.

The ELP can choose to make liquidity by submitting limit orders when the magnitude of their signal $\tilde{\theta}$ is small (given the signal has a zero expected value we refer to these outcomes as typical market conditions). However, the ELP will take liquidity by submitting market orders in turbulent market conditions, i.e. when $\tilde{\theta}$ is large in magnitude.[22] This cut-off strategy, x_E, can be written as

$$x_E = x_T(\tilde{\theta})\,\mathbf{1}_{|\tilde{\theta}|>\theta^*} + x_M(\tilde{\theta};\,\tilde{u},\tilde{p}\,)\mathbf{1}_{|\tilde{\theta}|\leq\theta^*}, \tag{2-3}$$

where $\mathbf{1}_{|\tilde{\theta}|>\theta^*}$ is the indicator function, x_T and x_M are the orders when the ELP chooses to be liquidity taker and maker, respectively. Parameter $\theta^* > 0$ represents the cut-off level of the information component $\tilde{\theta}$ for the ELP to switch between liquidity taker and maker, which is to be determined endogenously in equilibrium.

2.1.3. Market Clearance and Limited Market Participation

In this subsection, we show that the liquidity withdrawal of the ELP can have significant effects on market clearing, which is the key ingredient to produce fragile liquidity. Based on the switching behaviour of ELP outlined above, there are two equilibrium regimes: a

---

[22] The current theoretical literature assumes that HFTs can be either liquidity makers or liquidity takers, which is given exogenously. For example, Yang and Zhu (2020) model the strategic interaction between fundamental investors and high-frequency liquidity takers or 'back-runners' (whose only information is about history order flow of fundamental investors; back-runners partly infer fundamental investors' information from their order flow and exploit it in subsequent trading). Menkveld and Zoican (2017) show that speeding up the processing of orders by the exchange does not necessarily improve liquidity, given their modelling of a high-frequency market maker. The innovative feature of the model in this chapter is that the ELP applies a cut-off strategy to either profit from making liquidity or from taking liquidity and speculating on the size of information shock based on their signal. Empirically, HFTs typically act as liquidity makers; however, they can shift to taking liquidity as warranted by market conditions (Anand and Venkataraman, 2016, Korajczyk and Murphy, 2019, Raman et al., 2014).

'taker regime' where the ELP consumes liquidity; and a 'maker regime' where the ELP provides liquidity. In the taker regime all three traders (informed, ELP, and liquidity) take liquidity and their orders are absorbed by the DMM, who is the only liquidity provider in this setting. In equilibrium, the market clearing condition in the taker regime is given by

$$\tilde{u}_T + x_D(\tilde{u}_T, \tilde{p}_T) = 0 \quad \text{with} \quad \tilde{u}_T = x_I(\tilde{s}) + x_T(\tilde{\theta}) + \tilde{z}. \tag{2-4}$$

In the maker regime, the informed and liquidity traders take liquidity, while the ELP and DMM *compete* for providing liquidity. Developments in trading technology and trading automation (such as supercomputers and co-location with the exchange) have intensified the competition among HFTs and led to a concentration of liquidity provision by a small number of high-frequency firms. This concentration seems to limit DMM liquidity supply when high-frequency firms also provide liquidity. With the speed advantage, ELP's limit orders have a higher priority than the DMM's limit order.[23] This trading speed advantage can be characterized by parameter τ in the following market clearing condition in equilibrium,

$$\tilde{u}_M + x_M(\tilde{\theta}; \tilde{u}_M, \tilde{p}_M) + \tau x_D(\tilde{u}_M, \tilde{p}_M) = 0 \quad \text{with} \quad \tilde{u}_M = x_I(\tilde{s}) + \tilde{z}. \tag{2-5}$$

Here $\tau = \{0,1\}$ measures the *limited market participation* of the DMM. For $\tau = 0$, the ELP completely crowds out the DMM, who encounters limit market participation, and is the only liquidity provider. For $\tau = 1$, the ELP and the DMM share the market orders in financial market.

Definition 2.1: *The rational expectations cut-off equilibrium is characterized by asset price* $\tilde{p}$, *demand functions* $(x_I^*, x_M^*, x_T^*, x_D^*)$ *and optimal cut-off threshold value* θ^* *such that*

a) *given the cut-off threshold value* θ^*, *a demand function* x_i *maximizes trader's expected payoff as described in* (2-2);
b) *the risky asset price* $\tilde{p}$ *clears the market as described in equations* (2-4) *and* (2-5);

[23] This is consistent with empirical observations of the concentration of liquidity provision and reduced participation of retail investors in financial markets. HFT's take a large proportion of market-making trading volume, resulting in limited participation of DMM whose limit orders can only be executed partially when competing with HFTs.

c) *given the price and demand functions, the endogenous threshold value θ^* for the endogenous liquidity provision* (2-3) *maximizes ELP's expected payoff;*

d) *for every realization of the signals $\tilde{\theta}$ and $\tilde{s}$, the beliefs of all traders are consistent with a jointly conditional probability distribution.*

Due to the switching strategy of the ELP, the equilibrium becomes nonlinear. Specifically, with endogenous liquidity provision as outlined in (2-3), the ELP either submits market orders or limit orders by applying the cut-off strategy, depending on his private information. Hence, the ELP's trading behaviour provides additional uncertainty and information to the DMM and informed traders about risk asset's value. This leads to a nonlinear inference problem for the DMM (conditional on a truncated normal distribution). As a result, the equilibrium price function is nonlinear, and its specific functional form is unknown. Hence, the liquidity taker's trading strategies, which also determine the aggregate order flow, are also nonlinear.

2.1.4. Nonlinear Equilibrium with Endogenous Liquidity Provision

In this subsection, we characterize the nonlinear equilibrium with respect to the equilibrium demand and price functions and the endogenous threshold value of the ELP. After considering a linear approximation, we then introduce reinforcement learning to deliver the numerical results in Section 2.2. Firstly, with the exogenously given cut-off threshold value and price functions, we derive the demand functions for informed traders, ELP, and DMM.

Lemma 2.1 (Demand functions): *Given the threshold value θ^*, the demand function of the informed trader is given by*

$$x_I(\tilde{s}) = \gamma\tilde{s} - \gamma f_{p,I}(\tilde{s});$$

the demand functions of the ELP when consuming or suppling liquidity are given by, respectively,

$$x_T(\tilde{\theta}) = \gamma \frac{\sigma_f^2 + \sigma_0^2}{\sigma_f^2 + \sigma_0^2 + \sigma_{\varepsilon,\theta}^2} \tilde{\theta} - \gamma f_{p,E}(\tilde{\theta}), \qquad x_M(\tilde{\theta};\, \tilde{u}, \tilde{p}\,) = \gamma f_{V,E}(\tilde{\theta}, \tilde{u}) - \gamma \tilde{p};$$

and the demand function of the DMM is given by

$$x_D(\tilde{u}, \tilde{p}) = \gamma\big((1-\tau) + \tau \mathbf{1}_{\psi=0}\big) f_{V,D}(\tilde{u}) + \gamma \tau \mathbf{1}_{\psi \neq 0} f_{V,E}(\tilde{\theta}, \tilde{u}) - \gamma \tilde{p};$$

where

$$f_{p,E}(\tilde{\theta}) = E\big(\tilde{p}_T \big| \tilde{\theta} = \tilde{\delta}_0 + \tilde{\delta}_f + \tilde{\varepsilon}_\theta\big); \qquad f_{V,E}(\tilde{\theta}, \tilde{u}) = E\big(\tilde{V} \big| \tilde{\theta} = \tilde{\delta}_0 + \tilde{\delta}_f + \tilde{\varepsilon}_\theta, \tilde{u}\big);$$

$$f_{p,I}(\tilde{s}) = Pr\big(|\tilde{\theta}| > \theta^* | \tilde{s}\big) E\big(\tilde{p}_T \big| \tilde{s} = \tilde{\delta}_0 + \tilde{\delta}_d, |\tilde{\theta}| > \theta^*\big) + Pr\big(|\tilde{\theta}| \leq \theta^* | \tilde{s}\big) E\big(\tilde{p}_M \big| \tilde{s} = \tilde{\delta}_0 + \tilde{\delta}_d, |\tilde{\theta}| \leq \theta^*\big);$$

$$f_{V,D}(\tilde{u}) = E\big(\tilde{V} \big| \tilde{u}_T, |\tilde{\theta}| > \theta^*\big);$$

are, respectively, the rational expectations functions of the ELP about the equilibrium price and fundamental value, the informed about the equilibrium price and the DMM about the fundamental value conditional on their information. While the parameter ψ in DMM's demand is defined as

$$\psi = \tilde{p} - \frac{\tilde{u}}{\gamma} - E\big(\tilde{V} \big| \tilde{u}, |\tilde{\theta}| > \theta^*\big).$$

Lemma 2.1 leads to two key observations. First, the ELP's expectation function $f_{V,E}(\tilde{\theta}, \tilde{u})$ about the fundamental does not depend on the equilibrium price $\tilde{p}$. This is because DMM has no private information other than order flow; hence, the equilibrium price only reflects ELP's private information $\tilde{\theta}$ and the aggregate order flow $\tilde{u}$. Second, when $\tau = 1$, the DMM can compete with the ELP to provide liquidity. Although DMM cannot observe ELP's endogenous liquidity provision decision, DMM can infer the equilibrium regime from the equilibrium price and aggregate order flow, i.e., either the DMM is the only market maker (then the equilibrium price will be $\tilde{p}_T(\tilde{u})$) or ELP also makes market (then the equilibrium price will be $\tilde{p}_M(\tilde{\theta}, \tilde{u})$). Hence, if $\psi \neq 0$, DMM could be sure that ELP providing liquidity and would learn the ELP's information from the equilibrium price. Applying Lemma 2.1 to market clearing conditions (2-4) and (2-5), we obtain the following equilibrium price functions for given threshold value.

Proposition 2.1 (Price functions): *Given the threshold value θ^*, the equilibrium price functions under taker and maker regimes can be represented, respectively, by*

$$\tilde{p}_T(\tilde{u}) = \frac{1}{\gamma} \tilde{u} + f_{V,D}(\tilde{u}), \qquad \tilde{p}_M(\tilde{\theta}, \tilde{u}) = \frac{1}{\gamma(1+\tau)} \tilde{u} + f_{V,E}(\tilde{\theta}, \tilde{u}).$$

From Proposition 2.1, increasing ELP's speed advantage directly decreases liquidity under maker regime, which further influences the liquidity under taker regime through changing other participants' equilibrium demand. We next characterize the endogenous equilibrium threshold value. Unlike the DMM, the ELP has no obligation to supply liquidity. Thus, the ELP endogenously provides liquidity, depending on his information and market conditions. The equilibrium threshold signal value is defined such that the ELP's expected payoffs under both taker and maker regimes are the same. Proposition 2.2 gives the ELP threshold value that fulfils this requirement.

Proposition 2.2 (Cut-off strategy): *The endogenous threshold value θ^* of the ELP's cut-off strategy that determines endogenous liquidity provision is determined by*,

$$\theta^* = \left\{ \theta \,\middle|\, \underbrace{\left(\frac{\sigma_f^2 + \sigma_0^2}{\sigma_f^2 + \sigma_0^2 + \sigma_{\varepsilon,\theta}^2} \theta - f_{p,E}(\theta) \right)^2}_{taker\ regime} = \underbrace{E\left[\left(\frac{1}{\gamma(1+\tau)} \tilde{u} \right)^2 \middle| \tilde{\theta} = \theta \right]}_{maker\ regime} \right\}.$$

Proposition 2.1 and Proposition 2.2 show that the equilibrium is determined by the conditional expectation functions of the informed, ELP and DMM. Specifically, with fast trading speed and private information, HFTs can either compete with designated market makers (DMMs) by providing liquidity or attempt to profit from speculative trades to consume liquidity. When consuming liquidity, ELP bets on his extreme signals; when providing liquidity, HFTs accumulate wealth through millions of razor-thin profits from capturing fleeting price differences across major markets. In the following we focus on the filtering problem of traders' expectation functions (in Lemma 2.1), instead of dealing with the equilibrium demand and price functions directly.

Following the rational expectations in Lemma 2.1, the beliefs of all traders should be consistent with the jointly conditional probability distribution in equilibrium. This means that, in equilibrium,

$$f_{p,E}(\tilde{\theta}) = E\left(\frac{1}{\gamma}\tilde{u} + f_{V,D}(\tilde{u}) \middle| \tilde{\theta}, \tilde{u} = \gamma\tilde{s} - \gamma f_{p,I}(\tilde{s}) + \gamma \frac{\sigma_f^2 + \sigma_0^2}{\sigma_f^2 + \sigma_0^2 + \sigma_{\varepsilon,\theta}^2} \tilde{\theta} - \gamma f_{p,E}(\tilde{\theta}) + \tilde{z} \right). \tag{2-6}$$

Similar arguments are applied to other expectation functions in Lemma 2.1 (these are given in Appendix A2). Therefore, the equilibrium is characterised by a *'fixed-function'* problem of the expectation functions. In the traditional linear equilibrium framework, this is reduced to a 'fixed-point problem'. There are two key requirements that allow this reduction to work for traditional methods: the specific equilibrium structure needs to be known a priori; and the filter problem can be solved. However, in our model, due to the ELP's endogenous switching and the DMM's inference problem, the equilibrium becomes nonlinear. The equilibrium structure is unknown, which makes the equilibrium inference problem difficult via traditional methods.

Before formally introducing reinforcement learning methodology to solve the nonlinear equilibrium, we first consider a linear approximation to understand how the ELP's endogenous order choice affects the DMM's adverse selection. This assumption allows us to obtain a closed-form linear equilibrium and helps to identify the mechanism for liquidity fragility.

Assumption 2.1 (Linear approximation): *In calculating the posterior expectation about the fundamental payoff when the ELP takes liquidity, the DMM applies the projection theorem by using the posterior covariance and variance,*

$$E\left[\tilde{V}\middle|\ \tilde{u}_T = x_I(\tilde{s}) + x_{E,T}(\tilde{\theta}) + \tilde{z}, \left|\tilde{\theta}\right| > \theta^*\right] \approx \frac{COV\left(\tilde{V}, \tilde{u}_T\middle|\ \tilde{u}_T, \left|\tilde{\theta}\right| > \theta^*\right)}{VAR\left(\tilde{u}_T\middle|\ \tilde{u}_T,\ \left|\tilde{\theta}\right| > \theta^*\right)}\tilde{u}. \tag{2-7}$$

Proposition 2.3 (Linear approximation): *Given the optimal threshold value for ELP's cut-off strategy θ^*, under Assumption 2.1, there exists a linear equilibrium, where the price and demand functions are described by*

$$\tilde{p} = \left(\hat{\lambda}_T\tilde{u}_T\right)\mathbf{1}_{|\tilde{\theta}|>\theta^*} + \left(\hat{\lambda}_{\theta,M}\tilde{\theta} + \hat{\lambda}_M\tilde{u}_M\right)\mathbf{1}_{|\tilde{\theta}|\leq\theta^*};$$

$$x_I = \hat{\beta}_I\tilde{\delta}_d; \qquad x_T = \hat{\beta}_E\tilde{\theta}; \qquad x_M = \gamma\left(\hat{\phi}_{E,\theta}\tilde{\theta} + \hat{\phi}_E\tilde{u} - \tilde{p}\right);$$

$$x_D = \gamma\left(\left((1-\tau) + \tau\mathbf{1}_{\psi=0}\right)\hat{\phi}_D\tilde{u} + \tau\mathbf{1}_{\psi\neq0}\left(\hat{\phi}_{E,\theta}\tilde{\theta} + \hat{\phi}_E\tilde{u}\right) - \hat{\phi}_{D,P}\tilde{p}\right);$$

coefficients ($\hat{\beta}_I, \hat{\beta}_E, \hat{\phi}_{E,\theta}, \hat{\phi}_E, \hat{\phi}_D, \hat{\phi}_{D,P}, \hat{\lambda}_T, \hat{\lambda}_{\theta,M}, \hat{\lambda}_M$) are the functions of parameters ($\sigma_f^2, \sigma_{\varepsilon,\theta}^2, \sigma_d^2, \sigma_z^2$) given by (A-14), (A-16), (A-19), (A-20), (A-23) and (A-24) in Appendix A5. Furthermore, according to Lemma 2.1 *and applying the expression of* $E\left(\tilde{V}\middle|\tilde{u}, \left|\tilde{\theta}\right| > \theta^*\right)$ *from linear approximation results, parameter ψ satisfies*

$$\psi = \tilde{p} - \hat{\lambda}_T\tilde{u}_T;$$

the optimal threshold value for ELP's cut-off strategy are determined by

$$\theta^* = \left(\hat{\lambda}_M - \hat{\phi}_E\right)\sqrt{\frac{\left(\hat{\beta}_I^2\sigma_d^2 + \sigma_z^2\right)}{\beta_E^2 - \left(\phi_{E,\theta} - \lambda_{\theta,M}\right)^2}}.$$

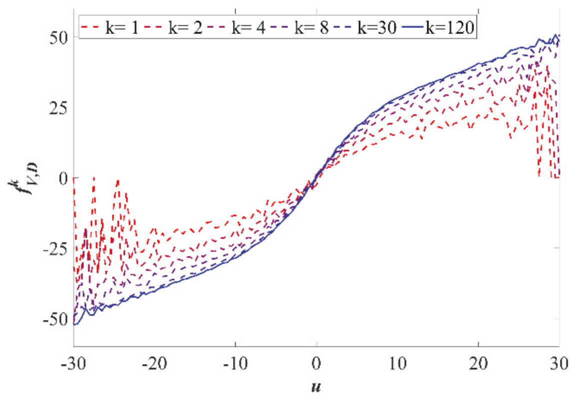
With the closed-form linear equilibrium in Proposition 2.3, we can examine the effect of the ELP's endogenous liquidity choice on the DMM's adverse selection, market liquidity, and the amplification of adverse selection from the ELP when switching from making to taking liquidity. As detailed later, linear frameworks can be helpful in capturing general mechanisms, they however may fail to deliver precise insights.

2.1.5. Reinforcement Learning

In this subsection we introduce the reinforcement learning algorithm and provide intuition about how to compute equilibrium expectation functions (which depicts the financial market equilibrium) and an equilibrium cut-off strategy (which describes ELP's endogenous liquidity provision decision), while leaving the details to Appendix B. Essentially, highly non-parametric reinforcement learning can be used to extract the conditional expectation functions without pre-specifying functional forms. The general idea of reinforcement learning technique is to delivering stable and consistent outcomes through gradually adjusting inferences and equilibrium parameters with simulations. By bootstrapping from experimental observations, it can capture the characteristics that parametric models may not recognize.

Based on Subsection 2.1.4, the nonlinear equilibrium is characterized by two parts. First, the financial market equilibrium is fully depicted by traders' expectation functions. For instance, in the benchmark model, there are four expectation functions, which are f_h with $h \in \{\{p, I\}; \{p, E\}; \{V, E\}; \{V, D\}\}$. Second, the ELP's liquidity provision decision is determined by the cut-off strategy. Specifically, the optimal threshold value θ^* for the cut-off strategy is determined through comparing ELP's realized profit of taking liquidity, $U_{\{E,T\}}$, and making liquidity, $U_{\{E,M\}}$. Denote F the function set for describing the expectation and utility functions and Θ the cut-off strategy set for governing ELP's endogenous liquidity provision decision. In the benchmark model, $F = \{f_{\{p,I\}}, f_{\{V,E\}}, f_{\{p,E\}}, f_{\{V,D\}}, U_{\{E,M\}}, U_{\{E,T\}}\}$ and $\Theta = \{\hat{\theta}\}$.

(a)



(b)

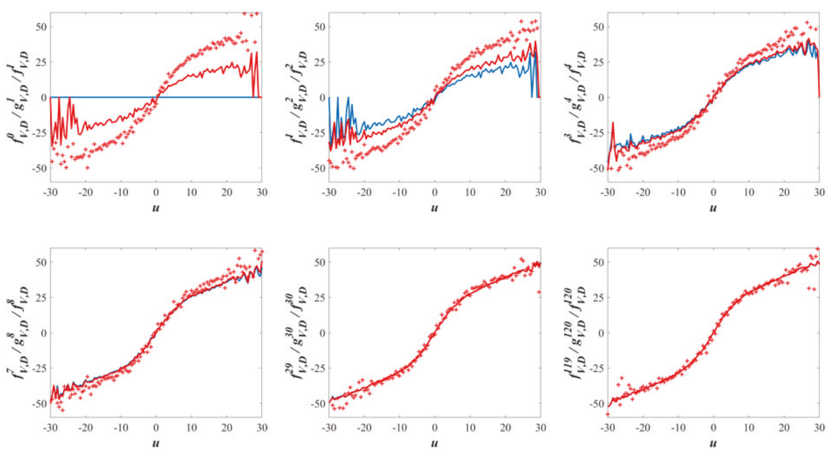


Figure 2-2

Reinforcement learning process. Panel (a) describes the evolution of the DMM's expectation function with reinforcement learning. The curves represent the evolution process of DMM's expectation functions $f_{V,D}^{k_m}$ about the fundamental value. At the end of $k_m - th$ run, we plot the updated expectation $f_{V,D}^{k_m}$ in dash lines. The finally converged $f_{V,D}^{k_m^*}$ is plotted as a solid line. The sequence numbers of the runs are shown in the legend. Panel (b) describes updating process for the DMM's fundamental value expectation functions with reinforcement learning during different stages of the simulations. At each panel, the plotted (blue and red) lines represent the expectation functions before and after the updating of the given run. The (red) dots represent the average of the observed fundamental values (as the input for the updating) during the given run. The six panels for the snapshots of updating are, from left to right and top to bottom: 1st, 2nd, 32nd and 120th. Parameters are $\sigma_f = 10$, $\sigma_d = 10$, $\sigma_{\varepsilon,\theta} = 8$, $\sigma_Z = 5$, $\gamma = 2$, $\tau = 0$.

In brief, we start with zero initial expectation functions and threshold value $\hat{\theta} = 2\sqrt{\sigma_0^2 + \sigma_d^2 + \sigma_{\varepsilon,\theta}^2}$, and use reinforcement learning algorithms to gradually update the expectation functions and cut-off strategy based on simulation realizations. Let m denote the m^{th} round and $\Theta^m = \{\hat{\theta}^m\}$ the m^{th} updating of the threshold value. In each round, traders' expectation functions are updated to reach financial market equilibrium. The resulting utility functions are then used to update the threshold value. Donate k_m as the sequential updating of function set in round m for the k_m-th updating and F^{k_m} the k_m-th updating function set. When the learning converges, the expectation functions satisfy equation (2-6). So that we reach the equilibrium F^* and Θ^*.

To generate experimental observations for $(k_m + 1) - th$ updating, we draw n realization values of randomly input variables and denote $(\tilde{\delta}_0^{k_m,i}, \tilde{\delta}_f^{k_m,i}, \tilde{\delta}_d^{k_m,i}, \tilde{z}^{k_m,i}, \tilde{\varepsilon}_\theta^{k_m,i})$ as the $i - th$ observation in this round. Based on these values and the expectation functions F^{k_m}, we calculate the demand functions of the informed, ELP and DMM based on Lemma 2.1. We then use the market clearing condition and Proposition 2.1 to obtain the equilibrium price and consequently the order flows and the ELP's realized profit of taking and making liquidity as $(\tilde{p}^{k_m,i}, \tilde{u}^{k_m,i}, \tilde{U}_{E,T}^{k_m,i}, \tilde{U}_{E,M}^{k_m,i})$. With n observations, the input-output pairs form a distribution for the four intermediate expectation functions, denoted G^{k_m}. The expectation functions are then updated according to $F^{k_m+1} = \beta^{k_m} G^{k_m} + (1 - \beta^{k_m}) F^{k_m}$, a weighted average of the prior expectation F^{k_m} in round k_m and the realized expectation G^{k_m}.[24] Figure 2-2 (a) shows the updating evolution process for DMM's expectation function about the fundamental value $f_{V,D}$.

We jointly apply this learning to all the four expectation functions in Lemma 2.1. We then introduce a convergence test (see details in Appendix B1) to determine the number of experiments. The convergence is tested for both determining financial market (given the cut-off strategy for liquidity provision) and governing endogenous liquidity provision. Firstly, we show the convergence of the learning to the expectation functions, resulting the *fixed-function* for equilibrium expectation functions. As an example, Figure 2-2 (b) illustrates the convergence to the solution of the 'fixed-function' problem. It shows

[24] We provide the details of how to calculate the realized expectation function G_h^k and the weighting function in Appendix B. Intuitively, we put more weight on the realized expectation function in the earlier experiments and less weight on the realized expectation function in the later experiments (when the simulation is stabilizing). This ensures a fast learning at the beginning of the process (depending more on the current outcomes due to the scarce historical data), with a constant learning speed in the latter experiments.

snapshots of the 1st, 2nd, 4th, 8th, 32nd and 120th updating. At the earlier rounds, the difference between the prior expectation (blue solid line) and updated expectation (red solid line) is significant. As the number of experiment increases, the difference between the prior and updated expectations vanishes so that the expectation functions converge. Secondly, after reaching the financial market equilibrium $F^{k_m^*}$ given the cut-off strategy set Θ^m, we can calculate the value Θ^{m+1} that makes the ELP indifferent between making and taking liquidity. In the equilibrium, these should be equal, i.e., $\Theta^m = \Theta^{m+1}$.

2.2 Limited Market Participation, Endogenous Liquidity Provision, and Fragile Liquidity

This section delivers equilibrium results and implications. The results show that fragile liquidity is a consequence of endogenous liquidity provision and limited market participation. This fragility, or possibility of liquidity loss, raises the prospect of market breaks when ELP's switch from liquidity making to liquidity taking in atypical times (i.e. when the ELP's signal suggests that the risky asset value is much different than expected).

2.2.1. Fragile Liquidity

We report our main numerical results of the nonlinear endogenous liquidity provision equilibrium in Figure 2-3. Figure 2-3 (a) reports the nonlinear equilibrium price functions under maker and taker regimes. The slope of the price function reflects the sensitivity of equilibrium prices to order flows, measuring liquidity (as in Kyle (1985)). Panel (a) shows that the price function under the taker regime is concave (in the absolute order flow) and has much higher slope than under the maker regime. This implies that liquidity under the taker regime is more fragile than under the maker regime. Besides, the price under the taker regime is solely determined by the DMM. This means that the nonlinear price function under the taker regime is mainly due to DMM's nonlinear inference. Therefore,

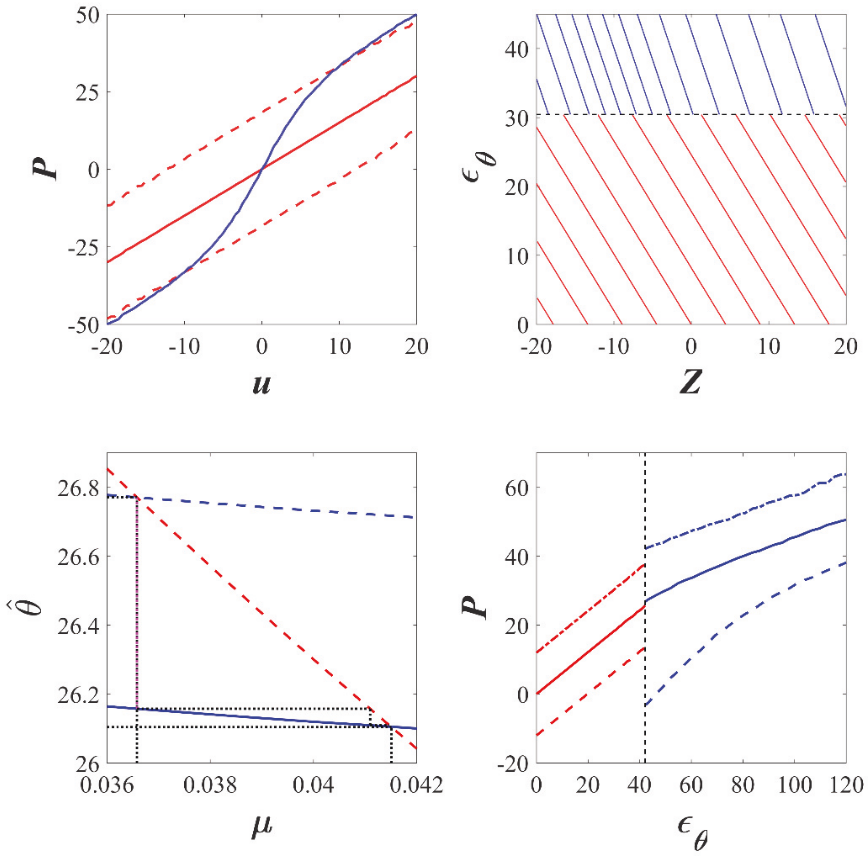


Figure 2-3

Equilibrium results. Top left panel (a) describes the price function under taker (blue) and maker (red) regime, while the solid and upper and bottom dashed lines represent realization value of $\tilde{\theta} = 0, 30, -30$, respectively. Top right panel (b) describes the iso-price lines in the information noise shock $\tilde{\varepsilon}_\theta$ and noise trading shock $\tilde{z}$. The (red) and (blue) solid lines represent the maker and taker regimes, respectively. The dashed line represents the threshold boundary between ELP's liquidity taker and maker regions, with the realization values of the systematic and idiosyncratic risk components $\tilde{\delta}_f = 0$ and $\tilde{\delta}_d = 0$. Bottom left panel (c) describes the best response functions of the ELP for a trading cost reduction from $\gamma = 2$ to $\gamma = 2.2$ using linear approximation. These two cases are represented by the dashed and solid lines, respectively. The red (blue) line represents the best response function for μ ($\hat{\theta}$) the with respect to $\hat{\theta}$ (μ), while the (pink) dotted line represents the initially direct effect of the reduction in trading cost. Bottom right panel (d) describes the equilibrium price with respect to information noise shocks for three realizations of liquidity trading $\tilde{z} = -8, 0, 8$, representing by the dotted, dashed and solid lines, respectively; here the realization values of the systematic and idiosyncratic risk components are $\tilde{\delta}_f = 0$ and $\tilde{\delta}_d = 0$. Other parameters are the same as in Figure 2-2.

the DMM's adverse selection is the source of fragile liquidity. These results lead to the following five important implications.

First, in contrast to the existing market microstructure models with constant price impact, empirical studies (by examining the impact of individual transactions in limit order book market) find that price change is more sensitive to small trading volume but less so to larger trading volume. These results (Hasbrouck, 1991, Hausman and a Simon, 1992, Keim and Madhavan, 1996, Lillo et al., 2003) are consistent with our concave price functions. More explicitly, the DMM's expectation function is S-shaped, with a higher slope when the order flow is relatively balanced between the informed and the ELP. Intuitively, a balanced aggregate order flow indicates that the orders from the informed trader and liquidity trader are opposite to the orders from the ELP. Therefore, the balanced aggregate order flow is less informative for the DMM, which increases the DMM's adverse selection and hence sensitivity to aggregate order flow.

Second, the price sensitivity to the changes in order flow can generate fragile liquidity. That is, a relatively small shock in one of the state variables (either fundamental or non-fundamental) can lead to a disproportionately large change in liquidity, as shown in Figure 2-3 (b). For example, liquidity can fall markedly with a small shock to ELP's private signal near the threshold value. This could trigger the ELP to switch from being a liquidity maker to a liquidity taker.[25] We conduct comparative statics analyses by comparing equilibrium prices across small variations in both parameters and realized state variables.[26] Given the parameters and realizations of information noisy shocks and noise trading shocks, the kinked contour lines of different pairs in Panel (b) represent equilibrium iso-price lines. It shows that the contours consist of two subspaces: taker regime and maker regime. Panel (b) shows that the contours have a steeper slope and are more densely distributed in the taker regime than in the maker regime, indicating low liquidity under taker regime. Furthermore, under the taker regime, the distances between the iso-price lines are tighter for balanced orders while wider for extreme orders, which is consistent with the nonlinear price impact function under the taker regime.

---

[25] For example, in a high-frequency trading setting, increased liquidity trading in another market might lead the ELP to infer aggressive aggregate trading volume that could push the realized signal above the threshold value and trigger the ELP to switch to a taker role, with a concomitant break in liquidity.

[26] Other static models of market breaks (Barlevy and Veronesi, 2003, Cespa and Foucault, 2014, Gennotte and Leland, 1990, Yuan, 2005) also use this approach.

Third, fragile liquidity could further lead to a market break, which is depicted in Figure 2-3 (d). Following Gennotte and Leland (1990), we define a market break as a scenario in which the equilibrium price function is discontinuous in any of its continuous arguments.[27] This definition captures the essential element of a break that a small change in a state variable is associated with a disproportionately large change in asset prices. Figure 2-3 (d) shows that for some state variable values, the equilibrium price function is discontinuous with respect to the ELP's signal. The price jumps can be either positive or negative when the ELP switches from being a liquidity maker to a liquidity taker. The results in Figure 2-3 (b) further show that market demand elasticity can drop precipitously due to the ELP's switching roles (more densely distributed price lines under the maker regime than the taker regime). As a result, the equilibrium price becomes very sensitive to shocks, leading to dramatical changes in prices. Therefore, the endogenous switch behaviour of high-frequency ELP can lead to sudden price jumps. Furthermore, Figure 2-3 (d) demonstrates that, even without changes in the fundamental value component, shocks to ELP's noise signal can trigger price jumps. Therefore, market breaks can occur regardless of the value of the underlying asset.

Fourth, the endogenous liquidity provision produces an amplification effect to the fragility. In the equilibrium, the probability μ for the ELP to take liquidity (referred to 'uncertainty' in the following) and the threshold value θ^* for the ELP's cut-off strategy (referred to 'liquidity provision decision' in the following) are linked through two best response functions.[28] Thus, an exogenous change influences the liquidity provision

[27] We are not the first to model market breaks. However, the existing literature of market breaks focuses on very different mechanisms. For example, through uninformed traders' inference problem, Barlevy and Veronesi (2003) and Yuan (2005) derive an upward-sloping demand functions that lead to market crashes. We instead show market breaks can happen through the ELP switching from liquidity maker to liquidity taker. This provides liquidity fragility as a novel mechanism for stock market breaks.

[28] Given the ELP's endogenous liquidity provision decision, the DMM forms beliefs about the ELP's trading behaviour. However, the DMM's pricing rule will affect the ELP's order type (making or taking) based on their expected profits. Therefore, the uncertainty and liquidity provision decisions are determined endogenously.

decision θ^*, which in turn affects the uncertainty μ that feeds back to the threshold value θ^*, as follows.[29]

Corollary 2.1 (Amplification effect): *Let Q be one of the exogenous parameters* $(\sigma_f, \sigma_d, \sigma_{\varepsilon,\theta}, \sigma_z, \gamma)$ *that determines the probability μ that ELP takes liquidity and the threshold θ* of ELP's cut-off strategy. The effect of Q on the uncertainty or liquidity provision decision is determined by*

$$\underbrace{\frac{di}{dQ}}_{total\ effect} = \underbrace{\kappa}_{Amplify\ effet} \times \underbrace{\left(\frac{\partial f_i}{\partial Q} + \frac{\partial f_i}{\partial j}\frac{\partial f_j}{\partial Q}\right)}_{direct\ effect} \tag{2-8}$$

for i, j=θ^*, μ*, in which* f_μ (f_{θ^*}) *are the best response function for μ* ($\hat{\theta}$) *with respect to* $\hat{\theta}$ (μ)*; and κ is an instability multiplier defined by*

$$\kappa = \left(1 - \frac{\partial f_\mu(\theta^*)}{\partial \theta^*}\frac{\partial f_\theta(\mu)}{\partial \mu}\right)^{-1}, \tag{2-9}$$

which is larger than one, $\kappa > 1$.

Corollary 2.1 illustrates the consequences of the feedback mechanism. Due to the substitution effect, the direct effect of a small shock to a fundamental parameter can be amplified through a feedback.[30] In summary, the interdependence of the uncertainty μ and liquidity provision decision θ^* forms a feedback loop and creates a multiplier (κ) that amplifies the initial effect of an exogenous parameter change in equilibrium. Figure 2-3 (c) illustrates this mechanism by changing the trading cost which then triggers feedback effects.

---

[29] An exogenous change in the liquidity provision decision can spill over to the uncertainty when $\partial f_\mu(\theta^*)/\partial\theta^* \neq 0$, while the change of the uncertainty can also feed back to the liquidity provision decision when $\partial f_\theta(\mu)/\partial\mu \neq 0$. We can show that both $\partial f_\mu(\theta^*)/\partial\theta^*$ and $\partial f_\theta(\mu)/\partial\mu$ are negative. Therefore, both the uncertainty and liquidity provision decision are substitutes; an increase in μ (θ^*) leads to a decline in $\theta^*(\mu)$, an amplify effect (so that $\kappa > 1$).

[30] Consider, for instance, the effects of a small reduction in the trading cost, denoted by $\Delta\gamma > 0$. The direct effect of this reduction lowers ELP's liquidity provision (by $(\partial f_\theta/\partial\gamma)\Delta\gamma$), increases the uncertainty (due to the negative spill-over effect $\partial f_\mu(\theta^*)/\partial\theta^* < 0$) by $(\partial f_\mu(\theta^*)/\partial\theta^*)(\partial f_\theta/\partial\gamma)\Delta\gamma$. This can occur even though all the fundamental parameters (fundamental value volatility σ_f^2 and signal noise volatility $\sigma_{\varepsilon,,\theta}^2$) are unchanged. Due to the amplifying effect, ELP's liquidity provision reduces further, triggering another feedback loop: the decrease in θ^* leads to an increase in μ, which further decrease θ^*, etc. Therefore, the total effect of an initial decrease in the trading costs is an order of magnitude larger than its initial, a direct effect.

Table 2-3

Model comparison in Section 3

This table compares the market dynamics of the equilibrium result of (i) the Kyle model (no HFT), (ii) FLT (HFT submits market orders to consume liquidity only), (iii) NLMP (nonlinear equilibrium without DMM's limited market participation), (iv) linear (approximation) equilibrium, and (v) nonlinear equilibrium. The making intensity is measured by the sensitivity of liquidity maker's demand to the aggregate market order flows, while liquidity is measured by the price sensitivity to the liquidity trader demand. We report market conditions, including liquidity when the ELP submits a market order (in the taker regime) and limit orders (in the maker regime), together with the threshold value and the probability of taking liquidity. We also report the taking intensity of the informed and ELP traders and the making intensity of the ELP and the DMM. The result for a linear equilibrium is based on the linear approximation, while the results for nonlinear equilibrium are based on the reinforcement learning (described in Appendix B) with ten-million simulated data (after the convergence). Liquidity is based on the regression of price on the noise demand. Trading intensity is based on regressing the expected price or fundamental value on observed signal for taking or making intensity, respectively. The parameter values are σ_f = 10, σ_d = 10, $\sigma_{\varepsilon,\theta}$ = 8, σ_Z = 5 and γ = 2.

	Threshold value	Probability	Liquidity under taker regime	Liquidity under maker regime	Informed taking intensity	ELP making intensity	ELP taking intensity	DMM making intensity
Kyle			1.5		0.5			1.00
FLT			1.76		0.44		0.27	1.26
NLMP	15.1	23.84%	2.03	1.25	0.52	1.00	0.24	1.54
Linear	26.77	3.66%	1.82	1.50	0.50	1.00	0.26	1.32
Nonlinear	42.10	0.10%	2.79	1.50	0.50	1.00	0.17	2.77

Lastly, we compare further the linear approximation and the nonlinear equilibrium. Table 2-3 shows that the linear model overestimates liquidity in the taker regime.[31] Therefore, the linear approximation might capture the same underlying mechanism, but fail to capture the consequences of nonlinearity and overstate liquidity under the taker regime. Further, with a higher threshold signal value for the DMM in the nonlinear

[31] In the nonlinear equilibrium, liquidity, measured by the price impact of orders, is nonlinear, depending on trade size. This price change is sensitive to small trades, but proportionally less sensitive to larger trades. To compare this nonlinear price impact with a constant price impact under the linear approximation, we use simulations to regress price to trades. The resulting regression coefficients are reported in the last row of Table 2-3. Specifically, after N-th updating, the results converge. Therefore, observations$\left(\tilde{\delta}_0^{N,i}, \tilde{\delta}_f^{N,i}, \tilde{\delta}_d^{N,i}, \tilde{z}^{N.i}, \tilde{\varepsilon}_\theta^{N,i}\right)$ and$\left(\tilde{p}^{N,i}, \tilde{u}^{N,i}, \tilde{U}_{E,T}^{N,i}, \tilde{U}_{E,M}^{N,i}\right)$ follow the equilibrium distribution. To measure the liquidity, we run the linear regression, $\tilde{p}^{N,i} = a + b\tilde{z}^{N.i}$, for all the observations i, and then the coefficient b measures the liquidity.

equilibrium (by about 50%), liquidity is more fragile than would be computed with the linear approximation. This implies that there is a less chance for liquidity to break down in the nonlinear equilibrium (comparing with the linear model), but when this occurs, the impact becomes more significant. This comparison has important implications for empirical studies. Most empirical framework focuses on linear analysis. However, our results show that this might be incomplete or misleading. When market participants are more diverse and market structures are more complex, being able to characterize how trading behaviour affects price and liquidity dynamics is important for investors and regulators. While linear frameworks can be helpful in capturing general mechanisms, they may fail to deliver precise insights.

Our overall results show that it is the DMM's inference problem that leads to a nonlinear price function, fragile liquidity, and a market break. We next provide an analysis of the underlying mechanism through which *HFTs' endogenous liquidity provision* and *DMM's limited market participation* (e.g., due to frictions induced by trading automation) contribute to liquidity fragility. Both mechanisms are integrated and affect jointly the adverse selection faced by the DMM.

2.2.2. Endogenous Liquidity Provision

The first channel for ELP to affect DMM's adverse selection is the ELP's endogenous liquidity provision. The ELP endogenously switches between providing and consuming liquidity based on an endogenous cut-off threshold value on his information. Without the liquidity supply obligation, the ELP withdraws liquidity supply when receiving extreme signals. Consequently, under the taker regime the DMM's demand function can be described by

$$x_D = \gamma E\left[\tilde{V}\middle|\ \tilde{u}_T = x_I(\tilde{s}) + x_{E,T}(\tilde{\theta}) + \tilde{z}, \left|\tilde{\boldsymbol{\theta}}\right| > \boldsymbol{\theta}^*\right] - \gamma\tilde{p}. \tag{2-10}$$

Equation (2-10) means that, under the taker regime, the DMM receives orders and knows that the ELP consumes liquidity. However, the ELP only consumes liquidity and submits a speculative market order when his signal is much different from what was expected. Thus, the DMM takes the ELP's selective liquidity into consideration, which means there is increasing toxicity from the aggregate market order flow (when $|\tilde{\theta}| > \theta^*$). In general,

(a) (b)

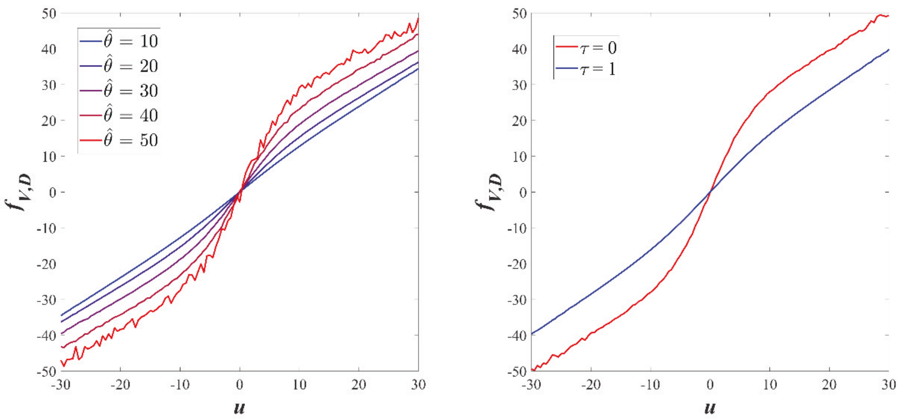


Figure 2-4

The inference of the DMM. Panel (a) shows DMM's expectation functions with different exogenous given threshold value $\hat{\theta}$ based on the rational learning for all types of traders. Characterized by the making intensity, it shows that the DMM's adverse selection increases with the ELP's optimal threshold signal value. Specifically, the DMM's expectation function about the fundamental payoff becomes more nonlinear and much steeper for moderate aggregate order flow. This intensifies DMM's adverse selection, generating a significant S-shaped price impact when the ELP takes liquidity. Panel (b) shows the DMM's expectation function for different values of τ. It shows that DMM's expectation function becomes more concave with respect to the (absolute) aggregate order flow as DMM's participation becomes limited (as τ change from 1 to 0). This makes the liquidity under the taker regime worse. Parameter values are $\sigma_f = 10$, $\sigma_d = 10$, $\sigma_{\varepsilon,\theta} = 8$, $\sigma_Z = 5$, $\gamma = 2$. In addition $\tau = 0$ in Panel (a).

the DMM knows that under these situations market conditions are unfavourable, which significantly intensifies the adverse selection, as shown next.

We examine the effect of endogenous liquidity provision on the DMM's making intensity and hence adverse selection. For the nonlinear equilibrium threshold value changes have two effects,

$$\frac{df_{V,D}(\tilde{u};\theta^*)}{d\theta^*} = \underbrace{\frac{\partial f_{V,D}(\tilde{u};\theta^*)}{\partial \tilde{u}}\frac{\partial \tilde{u}}{\partial \theta^*}}_{indirect\ effect} + \underbrace{\frac{\partial f_{V,D}(\tilde{u};\theta^*)}{\partial \theta^*}}_{direct\ effect}. \tag{2-11}$$

The direct effect in (2-11) describes the changes in adverse selection as the threshold value increases. We illustrate this effect with the results from reinforcement learning in Figure 2-4 (a). Characterized by the making intensity, it shows that the DMM's adverse selection increases with the ELP's optimal threshold signal value. Specifically, the DMM's expectation function about the fundamental payoff becomes more nonlinear and much steeper for moderate aggregate order flow. This intensifies DMM's adverse selection, generating a significant S-shaped price impact when the ELP takes liquidity (as illustrated in Figure 2-3 (a)).

To understand better the role of asymmetric information through the ELP and the uncertainty of the ELP's switching faced by other market participants, we compare the nonlinear equilibrium to two other linear equilibrium scenarios. The first scenario is the standard Kyle model without an HFT. The second scenario considers an HFT who only demands liquidity by submitting market orders, which is referred as FLT (fast liquidity taker) for convenience of discussion. Comparing the Kyle model with FLT shows the effect of asymmetry information, while comparing FLT with ELP shows the effect of the ELP's endogenous switch behaviour. The results are reported in Table 2-3. It shows that, compared to the Kyle model, the information asymmetry resulting from the presence of informed high-frequency liquidity taker reduces liquidity (by about 10%) and increases the adverse selection (by about 25%). Compared to FLT, the ELP's endogenous switching improves liquidity in the maker regime (by about 10%) but worsens the liquidity in the taking regime (by about 10%). Therefore, the ELP increases the adverse selection.

2.2.3. Limited Market Participation

The second channel for ELP to affect the DMM's adverse selection is that DMM's limited market participation. It is possible that developments in trading technology and trading automation (such as supercomputers and co-location with the exchange) have intensified competition among HFTs and led to a concentration of liquidity provision by a small number of HFT firms. HFT's speed and information advantages limit DMM's market participation. Table 2-3 and Figure 2-4 (b) show the influence of limited market participation.

Table 2-3 illustrates that the liquidity (overall, taker and maker regimes) is worse as the DMM's participation becomes limited, but probability for liquidity crashes also decreases. The main mechanism behind these results is that limited market participation intensifies the DMM's adverse selection cost. Figure 2-4 (b) depicts the influence of limited participation on the DMM's expectation function (hence price impact). It shows that the DMM's expectation function becomes more concave with respect to the (absolute) aggregate order flow as the DMM's participation becomes limited (as τ changes from 1 to 0). This makes the liquidity under the taker regime worse. When $\tau = 1$, two market makers (ELP and DMM) are competing for providing liquidity. Hence, the liquidity under the maker regime also becomes worse with limited DMM market participation.

The intuition behind these insights into the DMM's adverse selection is as follows. For $\tau = 1$, both DMM and ELP compete for providing liquidity. As the result, the payoffs of ELP for providing liquidity is relatively low comparing to the case when the ELP is the only market maker under maker regime (that is, when $\tau = 0$). Therefore, the ELP more prefers to provide liquidity. This leads to a large threshold value for the cut-off strategy. Second, and more importantly, Subsection 3.2. shows that order toxicity is higher with larger threshold value for the cut-off strategy. The ELP has no obligation to supply liquidity; and endogenously provides liquidity depending on his information and market conditions. When conditions become unfavourable, the ELP stops to making liquidity but takes liquidity instead. A larger threshold value means that the DMM faces higher adverse selection under the taker regime. Thus, limited market participation of DMM is consistent with a higher probability that market becomes extreme volatile, which intensifies the adverse selection.

2.2.4. Policy Implications

In this subsection, we explore several policy implications based on the above analysis.

Information Differentiation: Although we do not explicitly provide a micro-foundation for the information structure of ELP's private information, it is consistent with Cespa and Foucault (2014), Goldstein et al. (2014) and Yang and Zhu (2020) work on cross-learning. When the same or similar assets trade in multiple venues, this information structure allows high-frequency ELP to "back-run" (inferring fundamental information from order flows) in faster markets and to "front-run" (anticipating the order flows and

Table 2-4

Market condition and trading intensity

Market dynamics of the nonlinear equilibrium (based on machine learning) with respect to changes in information similarity (in the upper panel) and noise level in the ELP's signal (in the lower panel). In the upper panel, the information similarity between the informed and ELP is measured by the ratio of the common part (σ_0^2) to the total volatility of fundamentals for the informed ($\sigma_0^2 + \sigma_f^2$) and ELP ($\sigma_0^2 + \sigma_d^2$), respectively. The standard deviation (of the total volatility) is selected as 10. For example, σ_0^2 and σ_d^2 are 10 and 90 when the information similarity is 10%. Other parameter values are $\sigma_{\varepsilon,\theta} = 8, \sigma_Z = 5$ and $\gamma = 5$. In the lower panel, the noise level of the ELP's signal, measured by $\sigma_{\varepsilon,\theta}$, is given in the panel heading. Other parameter values are $\sigma_0 = 0, \sigma_f = 10, \sigma_d = 10, \sigma_Z = 5$ and $\gamma = 5$. The values in the parenthesis are the ratio of the threshold value to the standard deviation of θ. In both panels, market conditions and trading intensity are defined and estimated as in Table 2-3.

Information similarity	0%	2%	4%	6%	8%	10%
Threshold value	30.4	32.3	33.0	35.2	37.8	39.4
Liquidity under taker regime	1.659	1.679	1.685	1.666	1.718	1.723
Liquidity under maker regime	1.122	1.111	1.102	1.093	1.084	1.075
Informed taking intensity	0.751	0.749	0.747	0.745	0.742	0.739
ELP taking intensity	0.300	0.286	0.276	0.260	0.243	0.231
ELP making intensity	0.922	0.912	0.902	0.893	0.884	0.875
DMM making intensity	1.733	1.791	1.817	1.895	1.991	2.052

Noise of ELP	4	6	8	10	12	14
Threshold value	22.4 (2.08)	25.6 (2.20)	30.4 (2.37)	36.5 (2.58)	44.1 (2.82)	51.9 (3.02)
Liquidity under taker regime	1.732	1.690	1.659	1.649	1.610	1.577
Liquidity under maker regime	1.128	1.125	1.122	1.120	1.120	1.119
Informed taking intensity	0.741	0.746	0.751	0.754	0.756	0.757
ELP taking intensity	0.409	0.356	0.300	0.250	0.207	0.171
ELP making intensity	0.927	0.925	0.922	0.921	0.920	0.919
DMM making intensity	1.824	1.771	1.733	1.689	1.659	1.619

fundamental value) in slower markets. Therefore, when the venues become more integrated, the information impounded into different venues also becomes more similar. This makes information between ELP and informed trader more correlated. In our setup, the similarity between informed and ELP signals is measured by the ratio of the common part (σ_0^2) to the total volatility of the fundamentals for informed trader ($\sigma_0^2 + \sigma_f^2$) and ELP ($\sigma_0^2 + \sigma_d^2$).

The results on the impact of this information similarity are reported in the top panel of Table 2-4. It shows that, as the information of informed trader and ELP becomes more similar, the liquidity under the taker regime decreases while the liquidity under the maker regime increases. This implies that breaks in liquidity are more severe when their information signals are more similar. When markets become more related, the information traded and disseminated on different markets might also become more similar. Therefore, the high-frequency ELP, who usually trades on order flow information through back running, front running, and cross-market inference, might observe information that highly related to information signals of informed traders. Our analysis suggests that, even though the probability for breaks falls with increasing threshold values, the information similarity can make liquidity more fragile, potentially resulting in more significant liquidity breaks.

Information Quality: The ELP has private, but noisy, information about the fundamentals. Intuitively, high-frequency ELP may obtain information through "back-running" from a faster trading venue. Geographic distances can separate trading venues and cause information to travel between them with a delay. Therefore, the high-frequency ELP may not be able to obtain and process information immediately. The volatility ($\sigma_{\varepsilon,\theta}^2$) of the noise in ELP's private signal, hence the information quality, affects how fast the ELP can process order flow information from faster markets. With increasing information speed (or decreasing information latency), the volatility $\sigma_{\varepsilon,\theta}$ falls and ELP can obtain better information. In practice, this is influenced by institutional investments, such as communication technologies and server's colocation, and regulator innovations, such as speed bumps and message fees.[32]

To explore these changes in market structure, we conduct a comparative analysis on $\sigma_{\varepsilon,\theta}$. The results are reported in the low panel of Table 2-4. We find that, as the ELP's signal becomes nosier, liquidity under both taker and maker regimes improves. This suggests that liquidity traders benefit from low ELP's information speed or high latency in segmented markets. This is opposite to Lee (2019) who finds that, as cross-venue latency decreases (instead of increasing as in our model), liquidity and price discovery improve while the expected profits of informed traders decline. In Lee (2019), it is the designated market makers who observe cross market information. Therefore, a decline in cross-venue latency makes price at each venue more sensitive to cross venue information but less so to its own order flow information. In contrast, in our model, it is the high-frequency ELP who observes cross market information. Absent the obligation of making markets, the ELP can endogenously supply or consume liquidity based on their private information. Thus, an increase in information quality may impede market liquidity when ELPs are consuming liquidity.

[32] For example, orders and real time data are transmitted mostly via fibre optic cables, whose length adds to cross-venue latency. Thus, the rise of technology investment on computer hardware and connection to exchange servers has sped up information acquisition and dissemination. For example, Spread Networks' new fibre optic cable between Chicago and New York was 175 miles shorter than the existing cable. The new cable could allow traders in Chicago to learn of trades in New York earlier and update limit orders faster. These technological advances improve an ELP's information speed and reduce information latency in segmented markets, which improves ELP's information quality. *Colocation*: most exchanges allow traders to set up their servers beside the exchange, reducing the time for trader's order to reach the exchange. *Speed bumps*: "…exchange operators have responded with innovations such as co-location, microwave towers and latency upgrades as communications technology approaches the speed of light. Given these speeds, it has become easier to slow participants in order to speed others up. Several competing trading venues have proposed a number of speed bumps with a variety of features. Such delays may apply to all orders equally (symmetrically), or, as is the case for IEX, NYSE American, and CHX in the United States and TSX Alpha in Canada, almost all incoming messages are slowed – except for a specific subset of resting limit orders (asymmetric) which pay additional fees. Further, the duration of such speed bumps can be either deterministic or random…". More details of institutional backgrounds about speed bumps can be found in Chen et al. (2017a). *Message fees*: Malinova et al. (2018) examine an exchange that applies a message fee for cancellations and updates. They find that limit orders are updated less frequently after the fee is introduced and suggest that limit orders became less responsive to new information.

2.3 Circuit Breakers and Endogenous Liquidity Provision: An Intertemporal Model

This section considers the interaction between circuit breakers and endogenous liquidity provision. Following Subrahmanyam (1994), we extend our benchmark model of endogenous liquidity provision to a two-period model to incorporate circuit breakers. Circuit breakers affect the equilibrium through the strategic trading decisions of discretionary liquidity traders (i.e. with exogenous, or non-information based, needs to trade). In this section, we first introduce discretionary liquidity trading into a two-period model and analyze the effect of circuit breakers taking the endogenous liquidity provision into consideration. We then analyze the equilibrium and show that dramatic and inescapable development in algorithmic trading can lead to extreme volatile episodes.

2.3.1. The Intertemporal Two-Period Model

To incorporate the circuit breaker mechanism, we introduce discretionary liquidity trading to a two-period intertemporal model. To focus on the interaction of ELP and circuit breaker, we assume limited DMM participation (i.e. $\tau = 0$).

Two trading periods: As described in Table 2-2, the security is traded in two periods ($t = 1, 1'$) and its liquidation payoff, which is described by (2-1), is paid out at date 2. For tractability, following Subrahmanyam (1994), we assume that both informed trader and ELP could trade only in the early period ($t = 1$) and that their information sets are the same as in Subsection 2.1. In the later period ($t = 1'$), there is no further informed trading or endogenous liquidity provision. Apart from the DMM that was introduced before (denoted as DMM_e who bids for an order flow in the early period), there is another DMM, denoted as DMM_l who bids for an order flow in the later period. This allows us to tractably focus on the intertemporal problem of discretionary liquidity traders.

Discretionary liquidity trading: We model a single discretionary liquidity trader having an exogenous demand $2\tilde{l}$, who may choose to either split her trade equally across two periods or concentrate her trading in the early period. The random variable $\tilde{l}$ is normally distributed with mean zero and variance σ_l^2 and uncorrelated with other random variables. There is also a certain amount of nondiscretionary liquidity trading in each period ($t = 1, 1'$). The net nondiscretionary trade in period t is denoted by z_t, which is identically and independently normally distributed with mean zero and variance σ_z^2 and uncorrelated with other random variables.

Note that under our model structure, all agents except the discretionary trader have a one-period strategy. In deriving equilibrium, we assume that agents change their strategies in a consistent fashion. That is, when the discretionary liquidity trader changes her strategy, informed trader, ELP and DMMs respond by changing their strategies to conform to the new variance of liquidity trading in the trading periods. In equilibrium, the discretionary trader chooses the strategy with the lowest expected cost, while the market liquidity parameters reflect the total (discretionary and nondiscretionary) variances of liquidity trading in different periods under the strategy.

Circuit breaker: A circuit breaker is a rule that stipulates that if the price in the early period falls or raises outside specific price bounds, trading in the market will be halted in the later period (while trading in early period remains). This captures the feature that circuit breakers are not activated until the price (following the execution of an extreme order flow) crosses certain bounds. In the presence of the circuit breaker, the discretionary liquidity trader needs to decide whether to split her trade equally across two periods or concentrate her trading in the early period after factoring in the probability that trading in later period may not go through. We parameterize the cost of not being able to trade in later period by an exogenous parameter K^*; which is typically expected to be much larger than the unconditional expected losses suffered from trading with informed traders.

2.3.2. The Circuit Breaker and Endogenous Liquidity Provision

Considering the circuit breaker, the discretionary liquidity trader has two options: concentrating orders in early period or splitting orders into both periods.[33] When the discretionary liquidity trader chooses to concentrate her trading in early period, the aggregate market orders from liquidity trading in the early and later periods are given by $\tilde{n}_1^c = \tilde{z}_1 + 2\tilde{l}$ and $\tilde{n}_{1'}^c = \tilde{z}_{1'}$ (we use subscript *c* to represent the concentration of liquidity trading). Alternatively, when the discretionary liquidity trader chooses to split her trading into two periods, the aggregate market orders from liquidity trading in the early and later periods are given by $\tilde{n}_1^s = \tilde{z}_1 + \tilde{l}$ and $\tilde{n}_{1'}^s = \tilde{z}_{1'} + \tilde{l}$ (we use subscript *s* to represent splitting liquidity trade), respectively. Given that the trading in the early period always occurs no matter whether the price in early period moves outside the bounds or not, the equilibrium in the early period remains the same as given by Lemma 2.1, Proposition 2.1 and Proposition 2.2 through changing the market liquidity parameters to reflect the total variances of liquidity trading (from σ_z^2 to $Var(\tilde{n}_1^c) = \sigma_z^2 + 4\sigma_l^2$ or $Var(\tilde{n}_1^s) = \sigma_z^2 + \sigma_l^2$, depending on the discretionary liquidity trader's endogenous decision).

Lemma 2.2 (First period price functions with circuit breaker): *Given the discretionary liquidity trader's decision $k \in \{c, s\}$ of whether to split $(k = s)$ the order into both periods or concentrate $(k = c)$ the order in the early period, the aggregate liquidity trading in the early period is given by $\tilde{n}_1 = \{\tilde{n}_1^c, \tilde{n}_1^s\}$, with $\tilde{n}_1^c = \tilde{z}_1 + 2\tilde{l}$ and $\tilde{n}_1^s = \tilde{z}_1 + \tilde{l}$. Then the equilibrium prices $\tilde{p}_1^k$ for $k \in \{c, s\}$ in the early period is characterized by*

$$\tilde{p}_{1,T}^k(\tilde{u}_1) = \frac{1}{\gamma}\tilde{u}_1 + f_{V,D}^k(\tilde{u}_1), \qquad \tilde{p}_{1,M}^k(\tilde{\theta}, \tilde{u}_1) = \frac{1}{\gamma}\tilde{u}_1 + f_{V,E}^k(\tilde{\theta}, \tilde{u}_1).$$

The expectation functions of DMM_e $f_{V,D}^k$ and ELP $f_{V,E}^k$ depend on market liquidity parameters that reflect the total variances of liquidity trading $\sigma_{n,1}^c = \sqrt{\sigma_z^2 + 4\sigma_l^2}$ and $\sigma_{n,1}^s = \sqrt{\sigma_z^2 + \sigma_l^2}$.

[33] Discretionary liquidity trader will not concentrate trades in the second period. It becomes clear later that, without circuit breaker, discretionary liquidity trader always prefers to split the order in order to reduce the price impact by trading in a dispersed fashion across periods. Hence, once taking the circuit breaker (thus the cost of not being able to trade in the second period) into consideration, splitting orders is strictly better than concentrating orders in the later period.

On the equilibrium in the later period, by observing the equilibrium order imbalance $\tilde{u}_1$ and price $\tilde{p}_1^k$ in the early period, DMM_l knows which equilibrium regime in the early period (it must be in the taker regime when $\tilde{p}_1^k - \tilde{u}_1/\gamma - f_{V,D}^k(\tilde{u}_1) = 0$ or the maker regime otherwise) and bids based on the order flow $\tilde{u}_2$ in the later period and the equilibrium information in early period. We highlight two insights. First, when the ELP makes liquidity, observing equilibrium order imbalance $\tilde{u}_1$ and price $\tilde{p}_{1,M}^k$ means that DMM_l can extract the value of $\tilde{\theta}$ according to Lemma 2.2. Second, the order flow $\tilde{u}_2$ can only be useful when the discretionary liquidity trader chooses to split her trading across the two periods.[34] In this case, $\tilde{u}_1$, can provide information to predict the demand of the discretionary liquidity trader in the early period.

For example, assuming ELP takes liquidity, then the equilibrium order imbalance and price are equivalent in this case since the equilibrium order imbalance $\tilde{u}_1$ contains information about signals $\tilde{s}$ and $\tilde{\theta}$. In addition, DMM_l can use the order flow $\tilde{u}_2$ to predict the noise order flow (i.e., the discretionary liquidity trading $\tilde{l}$) in the early period. By Bayes' rule the two signals combine to generate the following signal in predicting the fundamental $\tilde{V}$:

$$\tilde{y} \equiv \tilde{u}_1 - \frac{\sigma_l^2}{\sigma_z^2 + \sigma_l^2}\tilde{u}_2 = x_I(\tilde{s}) + x_T(\tilde{\theta}) + \tilde{z}_1 + \tilde{l} - E(\tilde{z}_1 | \tilde{u}_2 = \tilde{z}_2 + \tilde{l})$$

$$= x_I(\tilde{s}) + x_T(\tilde{\theta}) + \tilde{z}_1 + \frac{\sigma_z^2}{\sigma_z^2 + \sigma_l^2}\tilde{l} - \frac{\sigma_l^2}{\sigma_z^2 + \sigma_l^2}\tilde{z}_2 = x_I(\tilde{s}) + x_T(\tilde{\theta}) + \tilde{n}_y.$$

The signal $\tilde{y}$ summarizes the overall information that DMM_l can extract after observing the aggregate order imbalances in both periods, in which the noise term $\tilde{n}_y$ follows

$$\tilde{n}_y \sim N\left(0, \sigma_z^2 + \frac{\sigma_z^2\sigma_l^2}{\sigma_z^2 + \sigma_l^2}\right).$$

When ELP makes liquidity, Bayes' rule delivers a similar result about the combination of the two signals, leading to the following equilibrium in the later period.

Proposition 2.4 (Second period price functions with circuit breaker): *Given the discretionary liquidity trader's decision $k \in \{c, s\}$ of whether to split the order or*

[34] When the discretionary liquidity trader chooses to concentrate trading in the early period, the order flow in the later period becomes a pure noise. Therefore, DMM_l has exactly the same expectation (as either ELP or DMM_e, depending on the equilibrium regime) as in the early period.

concentrate in the early period, the aggregate liquidity trading in the later period is given by $\tilde{n}_{1'} = \{\tilde{n}^c_{1'}, \tilde{n}^s_{1'}\}$, *where* $\tilde{n}^c_{1'} = \tilde{z}_{1'}, \tilde{n}^s_{1'} = \tilde{z}_{1'} + \tilde{l}$. *Then the equilibrium price* $\tilde{p}^c_{1'}$ *in the later period is characterized by*

$$\tilde{p}^c_{1',T}(\tilde{u}_1, \tilde{u}_{1'}) = \frac{1}{\gamma}\tilde{u}_{1'} + f^c_{V,D}(\tilde{u}_1), \qquad \tilde{p}^c_{1',M}(\tilde{\theta}, \tilde{u}_1, \tilde{u}_{1'}) = \frac{1}{\gamma}\tilde{u}_{1'} + f^c_{V,E}(\tilde{\theta}, \tilde{u}_1),$$

while the equilibrium prices $\tilde{p}^s_2$ *in the later period is characterized by*

$$\tilde{p}^s_{1',T}(\tilde{u}_1, \tilde{u}_{1'}) = \frac{1}{\gamma}\tilde{u}_{1'} + f^y_{V,D}(\tilde{y}), \qquad \tilde{p}^s_{1',M}(\tilde{\theta}, \tilde{u}_1, \tilde{u}_{1'}) = \frac{1}{\gamma}\tilde{u}_{1'} + f^y_{V,E}(\tilde{\theta}, \tilde{y}).$$

The expectation functions of DMM_l $f^y_{V,D}$ *when ELP takes liquidity and* $f^y_{V,E}$ *when ELP makes liquidity depend on market liquidity parameters that reflect the total variances of liquidity trading* $\sigma^s_{n,y} = \sqrt{\sigma^2_z + \sigma^2_z\sigma^2_l/(\sigma^2_z + \sigma^2_l)}$.

Finally, we examine the decision of the discretionary liquidity traders on whether to split the order into two periods or concentrate in the early period. Specifically, the discretionary liquidity trader takes the price function and the decisions of other market participants as given and compares her own utilities of splitting the order versus concentrating the order. Without circuit breakers, the discretionary liquidity trader always prefers to split the order in order to reduce the price impact. This is consistent with Chowdhry and Nanda (1991) and Subrahmanyam (1994) that it is optimal for discretionary traders to split their trades across markets or across periods for the same security rather than concentrating their trading in a particular market or in a particular period.

With a circuit breaker, the optimal strategy of the discretionary trader is then determined by comparing her expected trading costs among the two strategies. By concentrating the order into the early period, the discretionary liquidity trader would not bear the cost of not being able to trade in later period. In this case, holding $2\tilde{l}$ units of the risky asset yields an expected trading cost C^c.[35] Based on the price function in Lemma 2.2

[35] If $\tilde{l} > 0$, the discretionary liquidity trader buys $\tilde{l}$ units for $\tilde{p}_1$ in the early period and sells in the later period, receiving a payoff of $\tilde{V}$; if $\tilde{l} < 0$, she first shorts $\tilde{l}$ units of the risky asset, which then generates a payoff of $\tilde{p}_1$ per unit in the early period and an obligation of $\tilde{V}$ per unit in the later period. Since liquidity traders are trading against speculators who have superior information, the expected revenue $E[(\tilde{V} - \tilde{p}_1)2\tilde{l}]$ is negative, representing the expected cost for liquidity trader to meet her liquidity need. Therefore, we use the negative expected revenue to measure the expected trading loss directly.

and the order size $2\tilde{l}$, the expected trading cost for the discretionary liquidity trader is given by

$$C^c = E\left[\left(\tilde{p}_1^C - \tilde{V}\right)2\tilde{l}\right] = Pr\left(\left|\tilde{\theta}\right| > \theta^*\right) E\left[\left(\tilde{p}_{1,T}^C - \tilde{V}\right)2\tilde{l}\right] + Pr\left(\left|\tilde{\theta}\right| \leq \theta^*\right) E\left[\left(\tilde{p}_{1,M}^C - \tilde{V}\right)2\tilde{l}\right].$$

This cost reflects the uncertainty of the discretionary liquidity trader due to the endogenous liquidity provision of the ELP.

Alternatively, in the presence of the circuit breaker, to calculate the expected loss for splitting orders, the discretionary trader factors in the probability that the trading in the later period may not be possible. In this case, trading might be halted in the later period when the price in the early period moves out the price range $|\tilde{p}_1| > p^*$ with a probability of $P(|\tilde{p}_1^s| > p^*)$ for given $p^* > 0$. Based on the price function in Proposition 2.4 and the order size $\tilde{l}$ for the discretionary liquidity trader in each period, the expected trading cost of splitting trades is given by

$$C^s = E\left[\left(\tilde{p}_1^s - \tilde{V}\right)\tilde{l}\right] + P(|\tilde{p}_1^s| \leq p^*)E\left[\left(\tilde{p}_{1'}^s - \tilde{V}\right)\tilde{l}\right] + P(|\tilde{p}_1^s| > p^*)K^*.$$

Here $E\left[\left(\tilde{p}_1^s - \tilde{V}\right)\tilde{l}\right]$ and $E\left[\left(\tilde{p}_{1'}^s - \tilde{V}\right)\tilde{l}\right]$ are the expected trading loss of splitting trades in the early and later periods, respectively. The discretionary liquidity trader also faces the uncertainty about ELP trading behaviour, which changes the trading size.

Proposition 2.5 (Discretionary liquidity trader decision): *In the presence of a circuit breaker, there is a unique equilibrium and the discretionary trader concentrates her trades in the early period if and only if* $C^c < C^s$.

When the circuit breaker bound is crossed, the discretionary liquidity trader chooses to concentrate her trades in the early period. In this instance the expected cost of concentrated trading in the early period is less than that of splitting trades over two periods. When K^* is large and the price limit value p^* is relatively small, the total trading loss of splitting order C^s can be high for a wide range of parameter values.

2.3.3. Results and Implications

To understand better the underlying mechanism of the interaction between endogenous liquidity provision and circuit breaker in the full model (FM) outlined above, we begin by illustrating the main intuition for the case of exogenous liquidity provision; for convenience of discussion, we refer this to XLP . When the ELP's liquidity provision is given exogenously, the ELP randomly chooses to either consume liquidity with probability μ or supply liquidity with probability $(1-\mu)$, that is,

$$x_E = x_T(\tilde{\theta})\, \mathbf{1}_{\varphi=1} + x_M(\tilde{\theta};\, \tilde{u}_1, \tilde{p}\,)\mathbf{1}_{\varphi=0},$$

where φ is a binomial random variable, independent of others, satisfying $P(\,\varphi = 1) = \mu$ and $P(\,\varphi = 0) = 1-\mu$. Hence, the ELP's liquidity provision decision contains no information about the fundamental payoff. In this case, we have the following Lemma 2.3 from Proposition 2.4 and Lemma 2.2.

Lemma 2.3: *Given the discretionary liquidity trader's decision $k \in \{c, s\}$ and the exogenous probability that ELP takes liquidity μ^c and μ^s under these two scenarios, (i) the equilibrium price $\tilde{p}_1^k$ for $k \in \{c, s\}$ in the early period is given by*

$$\tilde{p}_{1,T}^k(\tilde{u}_1) = \left(\frac{1}{\gamma} + \phi_D^k\right)\tilde{u}_1, \qquad \tilde{p}_{1,M}^k(\tilde{\theta}, \tilde{u}_1) = \left(\frac{1}{\gamma} + \phi_E^k\right)\tilde{u}_1 + \phi_{E,\theta}^k\tilde{\theta};$$

(ii) the equilibrium prices $\tilde{p}_{1'}^c$ in the later period is given by

$$\tilde{p}_{1',T}^c(\tilde{u}_1, \tilde{u}_{1'}) = \frac{1}{\gamma}\tilde{u}_{1'} + \phi_D^c\tilde{u}_1, \qquad \tilde{p}_{1',M}^c(\tilde{\theta}, \tilde{u}_1, \tilde{u}_{1'}) = \frac{1}{\gamma}\tilde{u}_{1'} + \phi_E^c\tilde{u}_1 + \phi_{E,\theta}^c\tilde{\theta};$$

(iii) the equilibrium prices $\tilde{p}_{1'}^s$ in the later period is given by

$$\tilde{p}_{1',T}^s(\tilde{u}_1, \tilde{u}_{1'}) = \phi_D^y\tilde{u}_1 + \left(\frac{1}{\gamma} - \frac{\sigma_l^2\phi_D^y}{\sigma_z^2 + \sigma_l^2}\right)\tilde{u}_{1'}, \qquad \tilde{p}_{1',M}^s(\tilde{\theta}, \tilde{u}_1, \tilde{u}_{1'})$$
$$= \phi_E^y\tilde{u}_1 + \left(\frac{1}{\gamma} - \frac{\sigma_l^2\phi_E^y}{\sigma_z^2 + \sigma_l^2}\right)\tilde{u}_{1'} + \phi_{E,\theta}^y\tilde{\theta};$$

in which the parameter $\{\phi_D^c, \phi_D^s, \phi_D^y, \phi_{E,\theta}^c, \phi_{E,\theta}^s, \phi_{E,\theta}^y\}$ are depicted in Appendix A4. In the presence of a circuit breaker, there is a unique equilibrium and the discretionary liquidity trader concentrates trades in the early period if and only if

$$\sigma_l^2\left[\frac{1}{\gamma}+\mu^s\,\phi_D^s+(1-\mu^s\,)\phi_E^s\right]$$
$$+\sigma_l^2\left[\frac{1}{\gamma}+\frac{\sigma_Z^2}{\sigma_Z^2+\sigma_l^2}\left(\mu^s\,\phi_D^y+(1-\mu^s\,)\phi_E^y\right)\right]\left[\mu^s\,\left(1-2N\left(\frac{-p^*}{std\left(\tilde{p}_{1,T}^s\right)}\right)\right)\right.$$
$$\left.+(1-\mu^s\,)\left(1-2N\left(\frac{-p^*}{std\left(\tilde{p}_{1,M}^s\right)}\right)\right)\right]$$
$$+2\left[\mu^s\,N\left(\frac{-p^*}{std\left(\tilde{p}_{1,T}^s\right)}\right)+(1-\mu^s\,)N\left(\frac{-p^*}{std\left(\tilde{p}_{1,M}^s\right)}\right)\right]C^*$$
$$>4\sigma_l^2\left[\frac{1}{\gamma}+\mu^c\,\phi_D^c+(1-\mu^c\,)\phi_E^c\right],$$

where $N(\cdot)$ *represents the cumulative standard normal distribution function.*

To explore the role of the uncertainty in FM and XLP, we also consider a scenario where a high-frequency market maker only provides liquidity by submitting limit orders, denoted FLM (fast liquidity maker) for convenience of discussion. Comparing FLM with XLP shows the effect of the (exogenous) uncertainty about liquidity provision, while comparing XLP with FM shows the additional effect of endogenous liquidity provision. We choose additional parameter values: $\sigma_0=0$, $\sigma_f=10$, $\sigma_d=10$, $\sigma_{\varepsilon,\theta}=8$, $\sigma_Z=2$, $\sigma_l=2$, $\gamma=2$, $\tau=0$. For these parameters, we calculate the equilibrium for FLM when the discretionary liquidity trader splits trades. We chose price bound p^* =42.4451 for the circuit breaker so that, with the probability of 0.1%, the price in the early period moves outside the bounds. With this price limits, we calculate the equilibrium for the FM with endogenous liquidity provision and show that the probabilities that ELP takes liquidity when the discretionary liquidity trader splits and concentrates trades are 0.0088 and 0.0021, respectively. We then choose $\mu^s=0.0088$ and $\mu^c=0.0021$ and calculate the equilibrium for the XLP. The results are reported in Table 2-5. Comparing results of concentrating and splitting illustrates how the discretionary liquidity trader's decision affects market quality and trading equilibrium, resulting in four observations.

First, with the introduction of HFT endogenous liquidity provision, the probability of trigging a circuit breaker increases dramatically, no matter how the discretionary liquidity trader splits or concentrates trades. Specifically, in the FM, the circuit breaker probability increases by 8.99% (from 0.11988% to 0.13%) and 15.3% (from 0.1% to 0.1153%)

Table 2-5

The effect of circuit breaker and liquidity provision

This table compares the market dynamics of the equilibrium results among three scenarios, (i) FLM (HFT submits limit orders to provide liquidity only), (ii) XLP (HFT provides liquidity with exogenous probability), and (iii) FM (the nonlinear equilibrium with endogenous liquidity provision), when the discretionary liquidity trader with an exogenous demand $2\tilde{l}$ either concentrating in the early period or splitting equally across two periods. Panels (a), (b), (c) and (d) report the probability of triggering a circuit breaker (CB), the unconditional trading loss, the threshold cost of not being able to trade (that let discretionary liquidity trader indifference between split or concentrate orders), and price impact, respectively. The results show that, as the introduction of the exogenous and then endogenous uncertainty on liquidity provision (from HFT1 to XLP and then to FM), Panel (a) shows that the probability of circuit break (CB) increases (though the probability is higher when the discretionary trader chooses to split trades); Panel (b) shows that the trading cost increases; Panel (c) shows that the threshold cost of not being able to trade reduces; while Panel (d) shows that market liquidity is reduced. The parameter values are σ_f = 10, σ_d = 10, $\sigma_{\varepsilon,\theta}$ = 8, σ_Z = 5 and γ = 2.

		1) FLM	2) XLP	3) FM
(a) CB probability				
	Concentrate	0.11928%	0.11931%	0.13000%
	Split	0.10000%	0.10010%	0.11530%
(b) Trading cost				
	Concentrate	25.8615	25.8702	25.8961
	Split (first period)	8.8259	8.839	8.8715
	Split (second period)	5.7603	5.7672	5.7843
(c) Cost threshold				
		11281	11260	9752
(d) Liquidity (first period)				
	Concentrate	1.6163	1.6169	1.619
	Split	2.2065	2.2098	2.2186

comparing to the FLM when the discretionary liquidity trader either concentrates or splits orders, respectively. Initially advocated by the Brady Commission following the Black Monday of 1987 and subsequently implemented in 1988, the U.S. circuit breaker was triggered only once in October 27, 1997 before the 2020 coronavirus pandemic. However, recently we have witnessed more frequent episodes of extreme price movements and liquidity breaks over very short time periods. This result provides an interpretation of this evidence. In addition, by comparing the increments from FLM to XLP and then from XLP to FM, we see that much of the increments in crash probability come from the DMM's adverse selection cost due to the endogenous liquidity provision of HFT. The result in Table 2-5 shows how dramatically algorithmic trading activity could lead to more volatile market outcomes.

Second, with the introduction of HFT and endogenous liquidity provision, the trading cost for the discretionary trader increases and market liquidity in the early period decreases no matter whether the discretionary trader splits or concentrates trades. This is because the fast matching and information processing speeds of high-frequency ELP impose adverse selection on other market participants. Further, the discretionary liquidity trader has a greater incentive to concentrate trading in the early period. At the threshold cost of not being able to trade, the discretionary liquidity trader is indifferent in splitting and concentrating trades. When the cost of not being able to trade is higher than the threshold value, the discretionary liquidity trader concentrates orders. Hence, a lower threshold value of cost of not being able to trade means that the discretionary trader is more likely to concentrate trades in the early period.

Third, when the discretionary liquidity trader concentrates orders, the probability of triggering circuit breakers, market liquidity and trading costs all increase. This is consistent with the predictions about market quality in Subrahmanyam (1994). At the same time, the introduction of HFT and endogenous liquidity provision mean that the discretionary liquidity trader has a greater incentive to concentrate her trading in the early period. These outcomes describe a mechanism through which algorithmic trading can lead to more volatile market conditions. As an example, when the cost of not being able to trade is 10,000 (which is almost 400 times of discretionary liquidity trader's trading cost with concentrated orders), the discretionary liquidity trader would split their trades for the cases of FLM and XLP but concentrate orders for the case of ELP. This means that the introduction of ELP increases the probability of triggering circuit breaker by 30% (from 0.1% to 0.13%, which is large relative to small base).

Finally, we focus on the full model (FM). Recall that the discretionary liquidity trader always chooses to split orders without a circuit breaker. Hence, the FM can be used to examine the influence of circuit breaker on endogenous liquidity provision. When introducing circuit breaker into financial market, the discretionary liquidity trader might concentrate her order to mitigate the cost of not being able to trade in the later period. In this situation, the aggregate liquidity trading in the early period increases. As the result, the endogenous mechanism of ELP reduces ELP's chance to take liquidity. In other words, with ELP participating in market, the introduction of circuit breaker might improve market by lowing the probability of liquidity breaks.

2.4 The Risk of Simultaneous Liquidity Withdrew: A Multi-ELPs Model

This section extends the basic model of Section 2 to examine the competition effects of multi-ELPs.[36] We assume that the information sets of ELPs are correlated. Hence, without explicit obligations, investors who endogenously supply liquidity might simultaneously scale back when market conditions are unfavourable or hit by systematic extreme shocks. Consequently, the increase in the presence of algorithmic market-makers can exacerbate liquidity breaks and ultimately amplify market volatility.

2.4.1. The Model with Multi-ELPs

To incorporate multi-ELPs, we modify the information structure and introduce market clearance conditions that depart from those used in Section 2.1. To focus on the effect of multi-ELPs, we consider the market with limited DMM participation (that is, $\tau = 0$).

Agents: Different from Section 2.1, there are N (≥ 2) ELPs in the financial market. Hence, except the strategic interactions among ELP, informed trader and DMM, this

[36] With an increasing presence of new algorithmic market-makers, such as investment banks or dedicated trading firms, they compete liquidity provision with not only traditional DMMs but also high-frequency trading institutions. The 2020 March global bear markets in the history have reflected the scale of the coronavirus crisis and how stretched valuations already were at its onset. Analysts and investors have argued that the violence of the rout was also exacerbated by a dramatic drought in market liquidity.

section further captures the coordinative movement among different ELPs. We have seen increasing presence of "new" algorithmic market-makers competing liquidity provision with not only traditional DMMs but also high-frequency trading institutions.[37] In most of the time this competition benefits financial market due to faster and better liquidity provision; however, when markets become turbulent, not only high-frequency traders but also these "new" algorithmic market-makers automatically ratchet back their quoted prices and order size. Hence, the increasing endogenous liquidity provision can exacerbate the severity of systematic liquidity withdrew.

***Asset*:** To explore the simultaneously scale back of liquidity provision, we modify the payoff structure by

$$\tilde{V} = \underbrace{\overbrace{\tilde{\delta}_f^s}^{\substack{\text{common views}\\ \text{for ELPs}}} + \overbrace{\sum_{i=1}^{N} \tilde{\delta}_f^i}^{\substack{\text{distinct views}\\ \text{for ELPs}}}}_{\substack{\text{distinct views between}\\ \text{ELPs and informed trader}}} + \tilde{\delta}_d + \underbrace{\tilde{\delta}_0}_{\substack{\text{common views between}\\ \text{ELPs and informed trader}}}. \tag{2-12}$$

Here $\tilde{\delta}_d$ and $\tilde{\delta}_0$ remain the same as in Section 2; while $\tilde{\delta}_f^s \sim N(0, \rho\sigma_f^2)$ and $\tilde{\delta}_f^i \sim N(0, (1-\rho)\sigma_f^2)$ are mutually independent and normally distributed. We maintain multiple sources of randomness in the risky payoff, i.e., the informed trader and ELP have both distinct and common views about the payoff. Furthermore, we introduce the distinct and common views for different ELPs as well. Specifically, all ELPs' signals are related to $\tilde{\delta}_f^s$; but only ELP i's signal is related to $\tilde{\delta}_f^i$. This captures different ELPs, high-frequency traders and investment banks, who might have different information sources.

***Information*:** We then turn to the information sets. At date 1 before trading, the informed trader knows the realization of $\tilde{s} = \tilde{\delta}_0 + \tilde{\delta}_d$ while ELP i observes private information described by

$$\tilde{\theta}_i = \tilde{\delta}_0 + \tilde{\delta}_f^s + \tilde{\delta}_f^i + \tilde{\varepsilon}_{\theta i}.$$

As in Section 2.1, the innovation $\tilde{\varepsilon}_{\theta i} \sim N(0, \sigma_{\varepsilon,\theta}^2)$ measures the ELP's information precision, which can be influenced by the degree of technology investment. The parameter ρ

[37] Vowing never to bail out banks again, policymakers slapped them with stricter capital requirements and "proprietary" trading with their own money was strangled. As the result, the amount of money that banks can allocate to "market-making" — essentially acting as intermediaries between buyers and sellers — has shrunk. Instead, they have copied the techniques of high-frequency traders, replacing capital with speed.

measures the information similarity among different ELPs. As ρ increases, the information from different ELPs becomes more corelated. This information structure is consistent with the sole ELP considered in Section 2.1; the combination of signals $\tilde{\delta}_f^s$ and $\tilde{\delta}_f^i$ behaves the same as the signal $\tilde{\delta}_f$, while the total variance of signal ($Var(\tilde{\theta}_i) = Var(\tilde{\theta}) = \sigma_f^2 + \sigma_0^2 + \sigma_{\varepsilon,\theta}^2$) and covariance between signal and fundamental ($Cov(\tilde{\theta}_i, \tilde{V}) = Var(\tilde{\theta}, \tilde{V}) = \sigma_f^2 + \sigma_0^2$) are the same for both sole ELP and multi-ELPs setups.

***Market Clearance*:** Based on the ELPs switching behaviour outlined in equation (2-3), there are $N+1$ equilibrium regimes: a 'taker regime' where all ELPs scale back of liquidity provision and consume liquidity simultaneously; and N maker regimes where at least one of the ELPs provides liquidity. The different market regimes depend on the number of ELPs who provide liquidity in the market. In the taker regime all traders (informed, ELPs, and liquidity) take liquidity and in equilibrium their orders are absorbed by the DMM, who is the only liquidity provider. Therefore, the market clearing condition in the taker regime becomes

$$\tilde{u}_T + x_D(\tilde{u}_T, \tilde{p}_T) = 0 \quad \text{with} \quad \tilde{u}_T = x_I(\tilde{s}) + \textstyle\sum_{i=1}^N x_T(\tilde{\theta}_i) + \tilde{z}. \tag{2-13}$$

In the maker regimes, the informed and liquidity traders take liquidity; while at least one of ELPs supplies liquidity or multiple ELPs compete for providing liquidity. We differentiate different market regimes by the number $j = \{1, \dots, N\}$ of ELPs who provide liquidity. When several ELPs ($i = 1, \dots, j$) compete for liquidity provision, they share the aggregate orders,

$$\tilde{u}_M^j + \textstyle\sum_{i=1}^j x_M(\tilde{\theta}_i; \tilde{u}_M^j, \tilde{p}_M^j) = 0 \quad \text{with} \quad \tilde{u}_M^j = x_I(\tilde{s}) + \textstyle\sum_{l=j+1}^N x_T(\tilde{\theta}_l) + \tilde{z}. \tag{2-14}$$

From (2-13) and (2-14), all the ELPs provide liquidity only in the maker regimes. In addition, due to information similarity, ELPs might withdrew from providing liquidity simultaneously. With no obligation to supply liquidity, ELPs switch between making and taking liquidity endogenously. We formalize a rational expectations cut-off equilibrium, which requires market clearing under both maker and taker regimes. Based on the market clearance conditions (2-13) and (2-14), we would expect the overall market liquidity to be improved with multi-ELPs in normal market conditions, but market crashes could become extremely severe when all ELPs withdraw liquidity simultaneously.

2.4.2. Equilibrium Price and Demand Functions

This subsection characterizes the nonlinear equilibrium with respect to the equilibrium demand and price functions and the endogenous threshold value of the ELPs. The analysis is similar to Section 2.1 but becomes more complicated with multi-ELPs. Due to the nonlinearity in DMM's and ELPs' posterior beliefs, we do not have a closed form equilibrium solution. To characterize the equilibrium, we first derive an exogenous equilibrium when the threshold value is given exogenously. Similar to Lemma 2.1, the optimization problems lead to following Lemma 2.4 that describes the demand functions of the informed, ELPs, and DMM given the threshold value.

Lemma 2.4 (Demand functions with ELP competition): *Given threshold value θ^*, the demand function of the informed trader is given by*

$$x_I(\tilde{s}) = \gamma\tilde{s} - \gamma f_{p,I}(\tilde{s});$$

the demand functions of ELP i when consuming or suppling liquidity are given, respectively, by

$$x_T(\tilde{\theta}_i) = \gamma \frac{\sigma_f^2 + \sigma_0^2}{\sigma_f^2 + \sigma_0^2 + \sigma_{\varepsilon,\theta}^2}\tilde{\theta}_i - \gamma f_{p,E}(\tilde{\theta}_i), \qquad x_M(\tilde{\theta}_i; \tilde{u}, \tilde{p}) = \gamma f_{V,E}(\tilde{\theta}_i, \tilde{u}, \tilde{p}) - \gamma\tilde{p};$$

and the demand function of the DMM is given by

$$x_D(\tilde{u}, \tilde{p}) = \gamma f_{V,D}(\tilde{u}) - \gamma\tilde{p};$$

where

$$f_{p,E}(\tilde{\theta}_i) = \underbrace{Pr(j = 0|\tilde{\theta}_i)\, E(\tilde{p}_T|\tilde{\theta}_i, |\tilde{\theta}_{l\neq i}| > \theta^*)}_{taker\ regime} + \underbrace{\sum_{k=1}^{N-1} Pr(j = k|\tilde{\theta}_i)\, E(\tilde{p}_M^k|\tilde{\theta}_i, j = k)}_{maker\ regimes};$$

$$f_{V,E}(\tilde{\theta}_i, \tilde{u}, \tilde{p}) = E(\tilde{V}|\tilde{\theta}_i, \tilde{u}, |\tilde{\theta}_{l\neq i}| > \theta^*)\mathbf{1}_{\varpi=0} + \sum_{k=2}^{N} Pr(j = k|\tilde{\theta}_i, \tilde{u}, \varpi \neq 0)\, E(\tilde{V}|\tilde{\theta}_i, \tilde{u}, \tilde{p}, j = k, \varpi \neq 0)\, \mathbf{1}_{\varpi\neq 0};$$

$$f_{p,I}(\tilde{s}) = \underbrace{Pr(j = 0|\tilde{s})\, E(\tilde{p}_T|\tilde{s}, |\tilde{\theta}_l| > \theta^*)}_{taker\ regime} + \underbrace{\sum_{k=1}^{N} Pr(j = k|\tilde{s})\, E(\tilde{p}_M^k|\tilde{s}, j = k)}_{maker\ regimes};$$

$$f_{V,D}(\tilde{u}) = E(\tilde{V}|\tilde{u}, |\tilde{\theta}_l| > \theta^*);$$

are, respectively, the rational expectations functions of the ELP about the equilibrium price and fundamental value, the informed about the equilibrium price, and the DMM *about the fundamental value conditional on their information. Besides, the expression* ϖ *and* $Pr(j=k)$ *are defined by*

$$\varpi = \tilde{p} - \frac{\tilde{u}}{\gamma} - E\left(\tilde{V}\middle|\tilde{\theta}_i, \tilde{u}, \left|\tilde{\theta}_{l\neq i}\right| > \theta^*\right);$$

$$\psi(x, y; \theta^*, \rho) = \int_{-\infty}^{+\infty}\left[F_{\tilde{\delta}_{fi}}(\theta^* - x - y - z) - F_{\tilde{\delta}_{fi}}(-\theta^* - x - y - z)\right] f_{\tilde{\varepsilon}_\theta}(z)dz\,;$$

$$Pr(j=k) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} C_N^k \psi(x, y; \theta^*, \rho)^k (1 - \psi(x, y; \theta^*, \rho))^{N-k} f_{\tilde{\delta}_{fs}}(x)dx\, f_{\tilde{\delta}_0}(y)dy.$$

The most significant difference of Lemma 2.4 from Lemma 2.1 is that the rational expectations function of the ELP about the fundamental value $f_{V,E}(\tilde{\theta}_i, \tilde{u}, \tilde{p})$ depends on the equilibrium price $\tilde{p}$. When market making ELP i is the only market maker, the equilibrium price reflects the aggregate order flow and his own signal. Thus, $\tilde{\theta}_i$ and $\tilde{u}$ contain all the information in $\tilde{p}$. However, when there are more than two ELPs who provide liquidity, the equilibrium price reflects additional information about other ELPs' private information. Hence, ELPs should also use the equilibrium price to infer the fundamental information. This complicates dramatically the reinforcement learning procedure.[38] Besides, in the equilibrium, we are especially interested in the probability $Pr(j=0)$ in the taker regime with a systematic withdrew of supplying liquidity. At this stage, conditional on the signal correlation ρ, this probability depends on the exogenously given threshold value. Therefore, the endogenous threshold value is also influenced by the correlation ρ. We examine how this influences the simultaneously scale back of liquidity provision. Applying Lemma 2.4 to market clearance conditions (2-13) and (2-14), we obtain the following exogenous equilibrium price.

[38] It is difficult to incorporate price into the expectation of market maker (either the DMM or the ELPs who provide liquidity) when applying reinforcement learning. The reason is that, without considering the price as the information to infer fundamental value, the demand function of market maker is monotonic decreasing with price, which guarantees that there always exists a price clearing the market. However, when taking price into consideration, the demand function becomes nonlinear with respect to the equilibrium price, which may lead a nonexistence of market clearing price under some specific scenarios. Because of the randomness in the reinforcement learning, it is almost sure the nonexistence of price, especially in the early stage. Therefore, in Subsection 5.3, we consider $N = 2$ so that we do not incorporate price information for market maker.

Proposition 2.6 (Price functions with ELP competition): Given the threshold value θ^*, the equilibrium price functions under taker and maker regimes ($N+1$ regimes in total) can be represented, respectively, by

$$\tilde{p}_T(\tilde{u}) = \frac{\tilde{u}}{\gamma} + f_{V,D}(\tilde{u}), \qquad \tilde{p}_M^j(\tilde{\theta}_1, \ldots, \tilde{\theta}_j, \tilde{u}) = \frac{\tilde{u}}{j\gamma} + \sum_{i=1}^{j} \frac{f_{V,E}(\tilde{\theta}_i, \tilde{u}, \tilde{p})}{j}, \qquad j = 1, \ldots, N.$$

From Proposition 2.6, we see that market liquidity improves when more ELPs provide liquidity, a consequence of more intense competition among market making ELPs. We next characterize the endogenous equilibrium threshold value. Unlike the DMM, ELPs endogenously provide liquidity, depending on their information and market conditions. In the equilibrium threshold signal value, ELP's expected payoffs under both the taker and maker regimes are the same. Proposition 2.7 provides the ELPs' threshold signal that fulfils this requirement.

Proposition 2.7 (Cut-off strategy with ELP competition): *The endogenous threshold value θ^* is determined by*

$$\theta^* = \left\{ \theta \left| \left(\frac{\left(\rho + \sqrt{1-\rho^2}\right)\sigma_f^2 + \sigma_0^2}{\sigma_f^2 + \sigma_0^2 + \sigma_{\varepsilon,\theta}^2} \theta - f_{p,E}(\theta) \right)^2 = Pr(\mathbf{1}_{\varpi=0}|\theta)\, E\left[\left(\frac{\tilde{u}}{\gamma}\right)^2 \middle| \tilde{\theta}_i = \theta, j = 1 \right] + Pr(\mathbf{1}_{\varpi\neq 0}|\theta)\, \Pi \right. \right\}.$$

where the expression of Π *is given by*

$$\Pi = E\left[\sum_{k=2}^{N} Pr(j = k|\tilde{\theta}_i, \tilde{u}, \varpi \neq 0)\, E\left[\left(\frac{j-1}{j} f_{V,E}(\tilde{\theta}_i, \tilde{u}, \tilde{p}) - \frac{\tilde{u}}{j\gamma} - \sum_{l\neq i} \frac{f_{V,E}(\tilde{\theta}_l, \tilde{u}, \tilde{p})}{j} \right)^2 \middle| \tilde{\theta}_i, j = k, \varpi \neq 0 \right] \middle| \tilde{\theta}_i = \theta, j \geq 2 \right].$$

From Proposition 2.7, the endogenous threshold value depends on the expected profits under the two regimes. Further, when providing liquidity an ELP faces two possibilities. I.e., either he is the only market maker or there are other market making ELPs. Proposition 2.6 and Proposition 2.7 show that the equilibrium is determined by the conditional expectation functions of the informed trader, ELP and DMM. In the following discussion we focus on the filtering problem of traders' expectation functions (in Lemma 2.4), instead of dealing with the equilibrium demand and price functions directly. Again, we apply the method outlined in Section 2 to solve the nonlinear equilibrium.

2.4.3. Systematic Liquidity Withdrawal

In this subsection, we consider a case with two ELPs (i.e., $N = 2$) to illustrate the influence of increasing information similarity between ELPs. Relative to the sole ELP case, the two ELPs scenario can be treated as the introduction of new algorithmic market-makers, such as investment banks or dedicated trading firms. These entities compete for liquidity provision with not only traditional DMMs but also high-frequency traders. Hence, our focus is on different types of ELPs instead of simply s more crowded high-frequency industry. We first consider the linear case with exogenous liquidity provision in Lemma 2.5 to provide an intuition about the equilibrium; then use the reinforcement learning to characterize the overall effect.

Lemma 2.5 (Exogenous liquidity provision with ELP competition): *Assume that the exogenous probabilities μ_1 and μ_2 that ELP1 and ELP2 take the liquidity are same (i.e., $\mu_1 = \mu_2 = \mu$), there are three equilibrium regimes: (i) DMM provides liquidity (taker regime); (ii) one ELP provides liquidity (either ELP1 or ELP2); and (iii) both ELPs compete for providing liquidity. The equilibrium price functions under the taker and maker regimes can be represented, respectively, by*

$$\tilde{p}_T(\tilde{u}) = \left(\frac{1}{\gamma} + \phi_D\right)\tilde{u}, \tilde{p}_M^1(\tilde{\theta}_i, \tilde{u}) = \left(\frac{1}{\gamma} + \phi_E\right)\tilde{u} + \phi_{E,\theta}\tilde{\theta}_i,$$

$$\tilde{p}_M^2(\tilde{\theta}_1, \tilde{\theta}_2, \tilde{u}) = \left(\frac{1}{2\gamma} + \phi_{M,E}\right)\tilde{u} + \phi_{M,E,\theta}(\tilde{\theta}_1 + \tilde{\theta}_2),$$

in which the parameter $\{\phi_D, \phi_E, \phi_{E,\theta}, \phi_{M,E}, \phi_{M,E,\theta}\}$ are depicted in Appendix A1 and A2.

Lemma 2.5 illustrates that the DMM's adverse selection cost are severity when both ELPs withdrew from providing liquidity in the market. To explore the influence of the information correlation among different types of ELPs, we use the reinforcement learning methodology. When there are only two ELPs, market making ELP i is either the only market maker (the equilibrium price becomes $\tilde{p}_M^1(\tilde{\theta}_i, \tilde{u})$ this case) or competes with the other market making ELP (the equilibrium price becomes $\tilde{p}_M^2(\tilde{\theta}_1, \tilde{\theta}_2, \tilde{u})$ in this case). Hence, when $\varpi \neq 0$, both ELPs provide liquidity and learn from each other on the information from equilibrium price. This means that the ELPs' demand functions do not contain the

Table 2-6

Market condition with multi-ELPs

This table reports how the information similarity (ρ) among two ELPs influences the threshold value (θ^*) for the endogenous liquidity provision decision, the probability ($\Pr(j=0)$) that both ELPs take liquidity simultaneously, and the overall liquidity, respectively. They show that an increase in the information similarity (the 1st column) among two ELPs increases ELPs' liquidity provision (the 2nd column) and the overall market liquidity (the last column), but reduces the probability (the 3rd column) of both ELPs to take the liquidity simultaneously.

ρ	θ^*	$Pr(j=0)$	Overall liquidity
0.000	20.632	0.011	1.438
0.100	21.311	0.009	1.413
0.200	22.147	0.008	1.384
0.300	23.581	0.006	1.348
0.400	25.777	0.003	1.309
0.500	30.272	0.001	1.264

equilibrium price (using both private information instead).[39] Table 2-6 and Figure 2-5 report numerical analysis through reinforcement learning methodology.

Under endogenous liquidity provision equilibrium, increasing information similarity creates a trade-off on systematic liquidity withdrew. Increasing information similarity raises the coordination among ELPs as illustrated in exogenous situation. However, information similarity increases the endogenous threshold value for liquidity provision, which ultimately dominates the trade-off. Specifically, in the intermedium equilibrium (when the threshold value of ELP2's liquidity provision decision is exogenously given $\theta^* = 20$), we report ELP1's expected payoff for submitting limit orders (U_M) and market

[39] Otherwise, when $N \geq 3$, if $\varpi \neq 0$ ELP i would not know whether there are one or more ELPs making market, apart from himself. Hence, ELP i could only infer other ELPs' signal from price when forming the expectation about fundamental value. This, as mentioned before, would make the demand function becomes nonlinear with respect to the equilibrium price, which may lead a nonexistence of market clearing price under some specific scenarios.

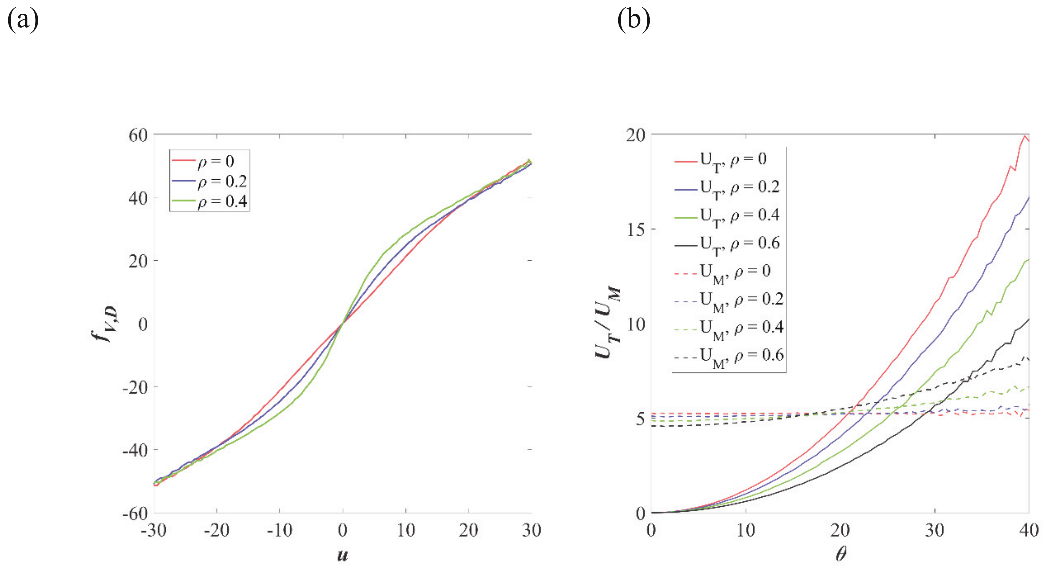


Figure 2-5

The price impact and order choice in multi-ELPs equilibrium. Panel (a) shows the DMM's expectation functions ($f_{V,D}$) about the payoff condition on the aggregate order flow for different level of information similarity (ρ) in equilibrium. It indicates an increasing concavity of the price impact function as the information similarity increases (from $\rho = 0$ to 0.2 and then 0.4). Panel (b) shows, given the threshold value of ELP2 ($\theta^* = 20$), the expected payoff of ELP1 with respect to the realized θ for submitting limit orders (U_M) and market orders (U_T) for different level of information similarity (ρ). It indicates that ELP1 prefers limit orders when his signals are low but market orders when his signals are high. However, as the information similarity increases (from $\rho = 0$ to 0.2, 0.4 and then 0.6), ELP1' utility of limit orders (U_M) increases, while the utility of market orders (U_T) decreases. This means that the probability of making the liquidity increases in information similarity. This becomes more significant as his signals become more extreme. We choose the additional parameter values: $\sigma_0 = 0$, $\sigma_f = 10$, $\sigma_d = 10$, $\sigma_{\varepsilon,\theta} = 8$, $\sigma_Z = 2$, $\gamma = 2$, $\tau = 0$.

orders (U_T) for different level of information similarity (ρ) and given threshold value (θ) with respect to his signals $\theta \in (0, 30)$ in Figure 2-5 (b). We have two observations. First, given the information similarity, ELP1 prefers limit orders when his signals are low (up to about $\theta = 20$) but market orders when his signals are high. Second, ELP1's utility of limit orders (U_M) increases but the utility of market orders (U_T) decreases as the information similarity increases (from $\rho = 0$ to 0.2, 0.4 and then 0.6). If we assume that the threshold value for ELP2 is given (as a constant), the unconditional probability of taking/making liquidity for ELP2 remains constant. For ELP1, with increasing information similarity, his profit for market order decreases. This is because ELP2 can predict better ELP1's market order when ELP2 is market making. When ELP1 provides liquidity, he can better predict ELP2's market order in the equilibrium regime when ELP2 is consuming liquidity. Hence, with more correlated information, the advantage for providing liquidity increases, as is illustrated in Figure 2-5 (b). This leads to an increasing endogenous threshold value (in the 2nd column in Table 2-6). These observations imply that an increase in information similarity raises the probability of making liquidity. This becomes more significant when ELP1's signals are more extreme.

In the full endogenous equilibrium (when the threshold value of the ELP's liquidity provision decision is determined endogenously), we report the DMM's expectation functions (E_V^D) with respect to the aggregate order flow for different level of information similarity (ρ) in Figure 2-5 (a). These show an increase in the concavity of the DMM's expectation functions as the information similarity increases (from $\rho = 0$ to 0.2 and then 0.4). Table 4 reports how information similarity (ρ) among two ELPs influences the threshold value (θ^*) for the endogenous liquidity provision decision, the probability ($\Pr(j = 0)$) that both ELPs take liquidity, and the overall liquidity, respectively. Both Table 2-6 and Figure 2-5 (a) provide several key results with respect to the increasing information similarity between two types of ELPs. First, individually, ELPs prefer to provide liquidity (the threshold value in the 2nd column of Table 2-6 for ELPs to provide liquidity increases). Second, even though information becomes more correlated, there is less risk for systematic liquidity withdrawal (the endogenous probability in the 3rd column of Table 4 for both ELPs to take liquidity decreases). Third, the overall market liquidity (in the last column of Table 2-6) improves (the average price impact under the three equilibrium regimes decreases). Last, the DMM's adverse selection cost is intensified (Figure 2-5 (a)), indicated by the increasing concavity of the price impact function as ρ increases. In general, the results illustrate that, in most of the time, more correlated ELPs is a boom, reducing the trading cost and stabilizing the market. It benefits the "new" algorithmic market-makers, such as investment banks and dedicated trading firms, through reducing their adverse selection cost. This is a real boost to conditions when the market is running smoothly. However, risk does not disappear. Rather it is transferred from one regime to another. In turbulent markets, with a small chance that ELPs scale back liquidity provision simultaneously, liquidity evaporates faster, and markets can break down in a more dramatic fashion.

In summary, competition among ELPs can ease their systematic and simultaneous liquidity withdrawal when their information becomes more correlated. This will improve the overall market liquidity in normal time. However, this can also speed up losses of liquidity during market turmoil.

2.5 Concluding Remarks

We study a nonlinear rational expectation equilibrium model of fragile liquidity with high-frequency endogenous liquidity provision. Both the designated market maker (DMM) and the endogenous liquidity provider (ELP) can provide liquidity in a setting that includes both an informational and a participation friction. We show that the ELP's endogenous liquidity provision, coupled with DMM's limited participation (a consequence of developments in trading technology and automation), can curtail liquidity from the DMM, and expand aggregate liquidity demand during atypical market activity.

We further demonstrate how reinforcement learning techniques can be used to analyse nonlinear equilibria in complex models of price formation. When exploring how the ELP's endogenous liquidity provision decision affects the DMM's adverse selection, we introduce a linear approximation to obtain a closed-form linear equilibrium, in addition to machine learning method that characterizes nonlinear equilibrium model. Contrasting results between the nonlinear equilibrium and linear approximation show the linear approximation fails to capture important nonlinearities and overestimates liquidity available under the taker regime. Most empirical frameworks focus on this linear situation structure, which can be helpful in capturing some general mechanisms but may fail to deliver results precisely. However, when market participants are diversified across numerous risky assets and where market structures are more complex, being able to characterize accurately how innovative and comprehensive trading behaviours affect price and liquidity dynamics is very important for investors and regulators.

Finally, we extend our model to study the influence of endogenous liquidity provision taking either circuit breakers or competition among HFTs into consideration. With circuit breakers, the presence of the ELP encourages the discretionary liquidity trader to concentrate trading in the early period. This increases the probability of both the trigging a circuit breaker and trading costs of the discretionary trader. It also results in lower market liquidity. However, the introduction of circuit breaker can reduce the probability of liquidity breaks, thereby improving market liquidity. With competition among HFT, an increase in information correlation among ELPs reduces the risk of systematic liquidity withdrawals but intensifies the DMM's adverse selection cost, thereby speeding up a loss of liquidity.

Chapter 3

The Fast and the Furious: Exchange Latency and Ever-fast Trading

Speed was once a ticket to fortune in financial markets. The extreme fast speed guarantees proprietary trading firms[40] to accumulate wealth through millions of razor-thin profits from capturing fleeting price differences across markets. Meanwhile, exchanges have spent hundreds of thousands dollars to build fast trading infrastructures to reduce order execution and communication latencies for catering to their quickest customers.[41] With ever-improved connections, exchanges have supplied high-frequency trading (HFT) to faster traders so that they can profit from price discrepancy between the most up-to-date (and more informative) prices and older (and less informative) prices offered by slower

---

[40] Proprietary trading firms, such as Citadel, Virtu, Flow Traders or Jump Trading, bet their own capital and largely serve as market makers, facilitating the buying and selling of shares, currencies, and other assets for investors. These high-frequency institutions speed up execution through technology investment on computer hardware, algorithms, and fast connection to exchange servers. They have made fortunes by being first to pounce on favourable prices since decades ago.

[41] Financial markets have made costly investments in trading infrastructure to reduce order execution and communication latencies. For instance, the trading latency on the NYSE dropped from 350 milliseconds in 2007 to 5 milliseconds in 2009. In October 2015, the Bombay Stock Exchange operated the fastest platform in the world, with a round-trip latency of 6 microseconds. Within a time window of eight years, trading platforms experienced a 50,000-fold speed increase. This process has gone beyond equities and futures markets to options, bonds, and currency exchange markets.

traders. In doing so, slower traders have had no choice but to invest in fast trading technology as well. This leads to a "latency war" in financial markets as both traders and exchanges speed up their trading.

Some studies view this trading speed evolution as beneficial for financial market participants, as extreme fast trading reduces the cost for liquidity supplier.[42] Others argue that the emerging of algorithms and fast trading technology bring investors' attention on both "excessive speed" demand from speculators and supply from trading platforms, contributing to increasingly extreme volatilities and liquidity breaks. Economists and policy makers have expressed concerns about some potentially negative effects of speed arms race, including (i) high-frequency trading might raise adverse selection costs for slow and liquidity traders (Biais et al., 2015, Hoffmann, 2014); (ii) quick response to order flow information would reduce the profitability of producing fundamental information and thereby make asset prices less informative (Dugast and Foucault, 2018); and (iii) the desire for speed could lead technology suppliers to provide traders faster access and lead to a perpetual arms race in speed, increasing market volatility (Budish et al., 2015).

Concerning about HFT would be harmful for financial markets and economy, leading to market instabilities or failures, a broad array of policies to slow down trading in financial markets have been proposed. For instance, some trading platforms (e.g., IEX or Alpha) now use so called "speed bumps" to delay the time elapsed between the moment at which an order arrives trading platforms and the moment in which the order enters the matching engine of the trading platforms (see Table C1 for exchanges that have introduced speed bumps). To reduce market volatility, Budish et al. (2015) propose to switch from continuous electronic order books (the current market organization in many electronic markets) to frequent batch auctions, in which market clearings take place only at discrete points in time (say, every second).

[42] This has been analysed in the literature through three channels: i) intensifying competition among liquidity providers, ii) lowering their inventory holding costs, and iii) expediting the search for trading counterparties (see Biais and Foucault (2014), Foucault and Moinas (2018), Menkveld (2016) for excellent surveys on both theoretical and empirical research about HFT).

Implicitly, these proposals view that market forces alone are not able to coordinate investment in fast trading technology and trading speed is excessive for social welfare. Accordingly, regulatory intervention is needed. This however raises the following fundamental questions. Why should we expect trading speed to be excessive? How much speed technology should investors and trading platform acquire? From a social welfare viewpoint, why market forces fail to coordinate at an optimal speed level? What are the implications for "price discovery" (information aggregation and price informativeness) and social cost? Can regulatory interventions help to mitigate these market failures? An economic analysis of these questions is nascent, and this chapter is devoted to such analysis.

Data for fast trading investment is scarce since both trading technologies are proprietary and the emerging of trading technologies takes years to complete. Thus, it is important to understand these fundamental questions from theoretical perspective. Within a unify framework, this chapter theoretically evaluates how technological innovations drive the speed investment of investors and trading platforms in fast trading technology. Although some recent work delves into the consequences of speed of either speculators (Huang and Yueshen, 2018) or platforms (Pagnotta and Philippon, 2018) and there have been empirical analyses of excessive speeds, to our knowledge, this is the first theoretical work that directly studies the endogenous speed acquisition decision of speculators and trading platforms simultaneously in a general framework.

To keep the model as simple (and as general) as possible, we consider two key features; trading speed demanded by fast speculators and supplied by exchanges (or trading platforms). Indeed, a salient feature of modern exchange architecture is its two-layer structure (gateway and matching engine) that involves two steps. Traders first send orders to exchanges and the exchanges then send orders to matcher after verification. Intrinsically underlying the first and second steps are the notions of the speed of speculators and the speed of exchanges, respectively. Motivated by the strategic trading

Table 3-1

Types of Equilibrium with respect to Endogenous Speed Acquisition

	Endogenous speed acquisition		
	Short term	Medium term	Long term
Speculators	No	Yes	Yes
Trading platforms	No	No	Yes

model in Froot et al. (1992),[43] we use endogenous decision on the probability that traders' order can be executed at earlier period to characterize the demand for the speed by speculators. In order to differentiate the speeds of speculators from the speed of exchanges, we decompose the uncertainty of order execution timing into agency latency and exchange latency (as in equation (3-1) below). The model is notable for allowing endogenous decisions for both intensive margin (the optimal trading speed invested by fast speculators) and extensive margin (the endogenous fraction of fast speculators) . While the exchange latency has the same impact on both fast and slow speculators, the fast speculators who use extraordinarily high-speed and sophisticated computer programs for generating, routing, and executing orders can significantly reduce their agent latency.

We characterize the equilibrium from short, medium, and long term perspectives, respectively, depending on whether speculators and exchanges are able to endogenously adjust their speed investment. As summarized in Table 3-1, from the speed evolution perspective, speculators can make endogenous speed acquisition decision in medium and long term while exchanges can only do this in long term. We then examine how the advances in (speculators or/and exchange) speed technology influence the investment in fast trading technology that leads to potential market failures with respect to excessive

[43] Consider an economy populated by three types of risk neutral agents trading a risky asset: N rational speculators, competitive market makers, and liquidity traders. Before trading, rational speculators, who observe the information about fundamental, face both extensive (i.e., whether become fast speculator) and intensive (i.e., how much speed to acquire) decisions on endogenous speed acquisition. Following Froot et al. (1992), we model the speed as the uncertainty of order execution timing. In our model, speed is characterized as the probability that traders' order executes at earlier period. The speed technology allows one to execute his trade faster; hence, his order has higher chance of executing before the information is capitalized into prices — fast speculators have a "first-mover advantage".

speed investment and speed mismatch between the desired exchange speed for policy makers and the optimal speed supplied by exchanges.

In the medium term, the endogenous acquisition of the speed technology for speculators depends on two cost components: a fixed cost and a variable cost that increases in the probability of order execution in early period. We show that, on the one hand, a reduction in the fixed cost promotes the population of fast speculators but demotes their demand on the speed. On the other hand, a reduction in the sensitivity to the varying cost promotes the speed acquisition of fast speculators, while it can promote (demote) the population of fast speculators when their agent-latency is high (low). This endogenous speed acquisition of speculators leads to an *excessive speed investment* of speculators due to that fast speculators (having the first-mover advantages) do not internalize the negative externality of their fast trading to other market participants. Essentially, the increasing trading from more fast speculators brings price closer to fundamental value, which then shrinks the net profit of slow speculators, forcing them to become faster. In equilibrium, netting the speed acquisition cost, fast and slow speculators have the same net profit, which is however significantly lower comparing to the situation without fast speculators. Furthermore, the advancing in speculators' speed technology makes it cheaper acquire speed, while intensified competition between fast and slow speculators reduce their profit. Therefore, the speed over-investment becomes worse with respect to technology advancing.

In the long term, apart from endogenous speed acquisition of speculators, the speed supply decision of exchanges depends endogenously on exchange speed technology cost that increases in provision of fast trading. Different sources of technology innovations (to reduce the fixed and variable costs for fast speculators and the variable cost for exchanges to provide fast trading speed) have similar impact on market quality and the optimal speed supplied by exchanges but affect speculators' speed investment differently. A novel finding is that an advancing in exchange speed technology can feed back into speculators' speed demand, making the speed of speculators more concentrated, i.e., a faster exchange shrinks market fraction of fast speculators but stimulates their trading speed. Intensive margin (the optimal investment speed level) and extensive margin (the endogenous fraction of fast speculators) are affected to each other. When taking this interaction into consideration, the change in exchange latency influences directly (indirectly) the optimal speed level (the fraction of fast speculators), which in turn affects indirectly (directly) the fraction of fast speculators (the optimal speed level) that feeds back to the optimal speed level (the fraction of fast speculators). We show that reducing exchange latency can increase (decrease) directly (indirectly) the optimal speed level and the fraction of fast speculators. In equilibrium, the direct effect dominates on the optimal speed level; by contrast, the indirect effect is dominating the fraction of fast speculators. Therefore, faster speculators become more concentrated, i.e., faster exchange speed shrinks market fraction of fast speculators but stimulates their optimal trading speed.

A policy maker might want to regulate the excessive speed investment for at least two potential reasons. Firstly, if price informativeness at long horizons is more important for economic decisions (like physical investment), then long-term information acquisition should be encouraged. Secondly, based on the risk-neutral Kyle (1985) framework, speed acquisition costs from speculators and trading platforms measure the deadweight loss for the economy. Policy makers, who care about price discovery and simultaneously want to limit the deadweight loss (resulting from the speed investment), can have a different objective from exchanges that simply maximize operation profits. Specifically, on the one hand, faster exchange speed increases the trading intensity in short run but decreases the trading intensity in long run. Due to the dominance of the trading intensity in long run, lowing exchange latency demotes price discovery. Hence, from price discovery perspective, regulator desires a default exchange speed which is lower than the endogenous optimal exchange speed. On the other hand, when policy makers only care about deadweight loss, there is a trade-off between decreasing aggregate speed acquisition cost for speculators and increasing cost of speed supply by exchange. This makes policy makers desire a moderate exchange speed that is even higher than the endogenous optimal exchange speed. In general, depending on speed technology level and the balance between efficiency and deadweight loss for policy makers, there is speed mismatch, either speed overinvestment or underinvestment, between the desired exchange speed from policy makers and the optimal speed supplied by exchanges. The findings echo recent concerns from market regulators about market failure with respect to excessive speed investment and speed mismatching caused by superior trading capabilities of fast trading.

Related Literature

Development of financial markets is largely driven by technological innovations. Starting with the computerization of tasks on trading floors, through the introduction of completely electronic markets, to Algorithmic Trading (AT), trading becomes much faster. Recently, the rise of HFTs, that is, investors rely on sophisticated computer programs to trade over extremely short time periods, is a major change in financial market (see Biais and Foucault (2014), Foucault and Moinas (2018), Menkveld (2016) for excellent surveys on both theoretical and empirical research about HFTs). Trading speed is important for HFTs; it is the key for the profitability of their strategies, such as directional trading on very short-lived signals and high-frequency market making. This chapter thus contributes to exploring how technological innovations and policy regulations drive the investment of investors and trading platforms in fast trading technology.

The speed characteristic considered in this chapter is largely motivated by the literature which features both "fast" and "slow" investors in sequential trading rounds (Banerjee et al., 2018, Froot et al., 1992, Hirshleifer et al., 1994). Particular, this chapter is related to Kendall (2018), Holden and Subrahmanyam (1996), Dugast and Foucault (2018) that examine the endogenous decisions to be either "fast" or "slow" (short- versus long-horizon; raw versus processed; quick and imprecise versus slow and precise signals). In modern electronic financial markets, there is an order-processing uncertainty that orders placed at the same time may not be executed simultaneously.[44] In this spirit, our

[44] As Huang and Yueshen (2018) explain, "the notion of speed roots in the course of financial securities trading, …, from the trading desk and onward, the speed of execution depends on technology investments on computer hardware, algorithms, and connection to exchange servers (co-location, fibre-optic cables, and microwave towers)". Over the last decade, the rising of algorithmic and HFT technologies amplify order-processing uncertainties. The institutional investors, such as investment banks and hedge funds, observe, process, and trade faster than normal investors, leading higher probability for their orders to be executed faster than others. As well documented in academia research and industry practice, these order-processing uncertainties are essential components of financial market in nowadays.

model is closely related to the theoretical analysis in Froot et al. (1992), who assume that informed traders face an ex-ante uncertainty of order execution timing. Following Froot et al. (1992), we characterize the speed by the probability that traders' order can be executed before the information is capitalized into prices, i.e., fast speculators have a "first-mover advantage". On the probability, instead of exogenously given in Froot et al. (1992), it is determined endogenously in equilibrium in our model. The model essentially takes a dynamic problem and compresses it into a simple two period's model.

A salient feature of the modern exchange architecture is its two-layer structure. It involves two steps: first, traders send orders to exchanges; second, after verification, the exchanges send orders on matcher. Prior theoretical work on fast trading investment focuses on their role in either speed supply, i.e., endogenous speed acquisition of trading platforms, or speed demand, i.e., endogenous determination of speculators' speed, but not in both. For example, Pagnotta and Philippon (2018) analyse theoretically how trading platforms compete in speed, while Riordan and Storkenmaier (2012), Ye et al. (2013) empirically examine the liquidity effect of faster exchange speed.[45] In Pagnotta and Philippon (2018), the speed of trading platform determines the rate at which buyers and sellers are matched. Regarding the demand for speed, several theoretical studies have analyzed high-frequency traders from either liquidity consuming or liquidity providing perspectives (see, for instance, Biais et al. (2015) and Aït-Sahalia and Saglam (2013)).[46] Furthermore, Huang and Yueshen (2018) novelty study investors' speed acquisition, alongside their information acquisition. Speed heterogeneity arises in equilibrium,

---

[45] Empirical studies show that exchange speed upgrades have an ambiguous effect on liquidity. Riordan and Storkenmaier (2012) find that a latency reduction at the German stock exchange had no detectable effect on the effective spread for most sample stocks, only for quartile of smallest stocks do they report a spread reduction. Ye et al. (2013) study NASDAQ exchange latency reductions and find that quoted spread did not change.

[46] Studies that model speed generically as a market contact rate find that speed is beneficial for liquidity (Baron et al., 2019, Hoffmann, 2014). Essentially, gains from trade are realized more quickly. Research on changing speed for specific category of traders finds that an increase in high-frequency market makers' speed benefits liquidity (Aït-Sahalia and Saglam, 2013) whereas an increase in high-frequency speculators' speed hurts liquidity (Biais et al., 2015). The underlying mechanism of these results is driven by adverse-selection cost that high-frequency speculators inflict on high-frequency market makers. Even more specifically, within the group of high-frequency speculators, if one makes some high-frequency speculators faster, this then benefits liquidity as it enables high-frequency market makers to learn more quickly (Roşu, 2019).

fragmenting information aggregation process and resulting nonmonotone impact on the overall price informativeness.

While these research are important in providing a basic understanding from either speculators' perspective or trading platforms' perspective, focusing solely on one of them neglects the interaction in speed investment between speculators and trading platforms and thus understates the fundamental merits of fast trading investment mechanism. Given that the speeds of speculators and trading platforms are not mutually exclusive, this chapter fills this gap. By developing a model to decompose the uncertainty of order execution timing into agency latency and exchange latency (as in equation (3-1) below), our work complements these studies. Menkveld and Zoican (2017) also consider the two-layer structure in a standard limit-order market model but with completely different focus,[47] and the speed is exogenous. By contrast, our focus is on the interaction of the endogenous speed acquisitions between speculators and trading platforms.

The chapter is organized as follows. Section 3.1 develops the model and considers speed acquisition decision from short, medium, and long-term perspectives. Section 3.2 characterizes market equilibrium. Sections 3.3 and 3.4 examine medium- and long-term equilibrium, respectively. Section 3.5 examines speed mismatch and provides some policy implications. Section 3.6 concludes. All the proofs and additional discussions are presented in the Appendix.

3.1 The Model

Market exchanges have made costly investments in trading infrastructure to reduce order execution and communication latencies. This benefits not only fast speculators but also

[47] Menkveld and Zoican (2017) propose a model to analyze the impact of exchange latency on liquidity in the presence of two canonical HFT strategies. They find that speeding up the exchange does not necessarily improve liquidity. On the one hand, more speed enables a high-frequency market maker (HFM) to update his quotes faster on incoming news. This reduces his payoff risk and thus lowers the competitive bid-ask spread. On the other hand, HFM's price quotes are more likely to meet speculative high-frequency bandits and less likely to meet liquidity traders. This raises the spread. The net effect depends on security's news-to-liquidity-trader ratio.

slow speculators in general. By contrast, some institutions speed up execution through technology investment on computer hardware, algorithms, and fast connection to exchange servers. These fast speculators who use “extraordinarily high-speed and sophisticated computer programs for generating, routing, and executing orders (SEC, 2010, p. 45)” can significantly lower agent latency and create the speed advantage; hence, fast speculators that "beat the gun" are able to profit. This section establishes a model with speed hierarchy. The framework is akin to Froot et al. (1992) in terms of order execution uncertainty. Instead of focusing on herding, we are interested in optimal speed demand from speculators and speed supply from exchange.

3.1.1. Financial Market

Assets: Consider a financial market trading for a risky asset with zero net supply. The asset pays a liquidation dividend of v, which is a normally distributed with mean of zero and variance of σ^2. There exists one exchange for speculators trading the asset.[48]

Speculators: There are total N risk neutral rational speculators. To focus the analysis on the uncertainty of order execution timing,[49] we assume they all observe the fundamental value v. Specifically, all speculators submit asset demands simultaneously at $t = 0$ but they are uncertain whether their orders will be executed without delay (at $t = 1$) or with delay (at $t = 2$). Speculators are large enough to affect market price, and they take this into account when formulating their demands.

Speed technology: Speculators can be different in trading speed subject to speed technology investment. Let q be the probability for an order to be executed at early round

[48] To simply the analysis, we consider a monopoly exchange situation, i.e., only one exchange charging a membership fee. Regarding exchange competition, Pagnotta and Philippon (2018) analyze how trading platforms compete for the speed. In their model, the speed of trading on a platform determines the rate at which buyers and sellers are matched. To focus on the interaction between demand and supply of the speed and different objection between exchange and speculators, we exclude the exchange competition and consider only one exchange.

[49] Our work thus contributes to the market microstructure literature on multi-dimensional sources of randomness with both risk-neutral agents in static (Avery and Zemsky, 1998, Gervais and Hall, 1997, Romer, 1992) or dynamic models (Back et al., 2018, Li, 2013) and risk averse investors (Banerjee and Green, 2015, Gao et al., 2013). Different from this literature, we consider the uncertainty of order execution timing.

$(t = 1)$, $1 - q$ be the probability for the order to be executed at later round $(t = 2)$, μ be the fraction of fast speculators with $q = q_F$, and $1 - \mu$ be the fraction of slow speculators with $q = q_S$ $(< q_F)$. Then there are μN fast speculators and $(1 - \mu)N$ slow speculators. With $q_F > q_S$, this means that fast speculators have a "first-mover advantage" in that their orders have a higher chance to be executed before the information is capitalized into market prices.

Agent and exchange latency: A salient feature of the modern exchange architecture is its two-layer structure: gateway and matching engine.[50] It involves two steps. First, traders send orders to exchange. Second, after verification, exchange send orders on matcher. This means that the trading speed of speculator i is determined by

$$q_i = l_i \times l_e, \qquad (i = F, S), \tag{3-1}$$

in which l_F, l_S and l_e represent the agent latency for fast and slow speculator and exchange latency, respectively. Agent (exchange) latency reflects the probability that speculators' order is not delayed at first (second) layer. The speed technology of fast speculators only affects agent latency, while the latency investment from trading platforms has symmetric influence on both fast and slow speculators.

Timeline: There are four dates $t \in \{0,1,2,3\}$. The timeline and order of events are described in Figure 1. At $t = 0$, speculators make a speed acquisition decision to endogenously determine both the fraction μ of fast speculators that measures the *extensive margin*, and the agent latency level l_F of fast speculators that measures the *intensive margin*. While the exchange also makes a speed acquisition decision to endogenously determine the exchange latency l_e. Besides, the speculators and exchange decide the membership fee f

[50] Once a trader's message arrives at the exchange, it enters a gateway (the first layer). This gateway processes the order. For example, it checks whether the order contains all required information and if the submitter is allowed to send the type of order he is sending. If all is verified then the order is routed on to the matching engine, or matcher for short (the second layer). Thus, there are three different sources of latency that an order experiences after leaving trader's server on its way to the exchange matcher: network-to-gateway, gateway-processing, and gateway-to-matching-engine. Agent latency is reflected in the first source while exchange latency is the sum of the second and third sources.

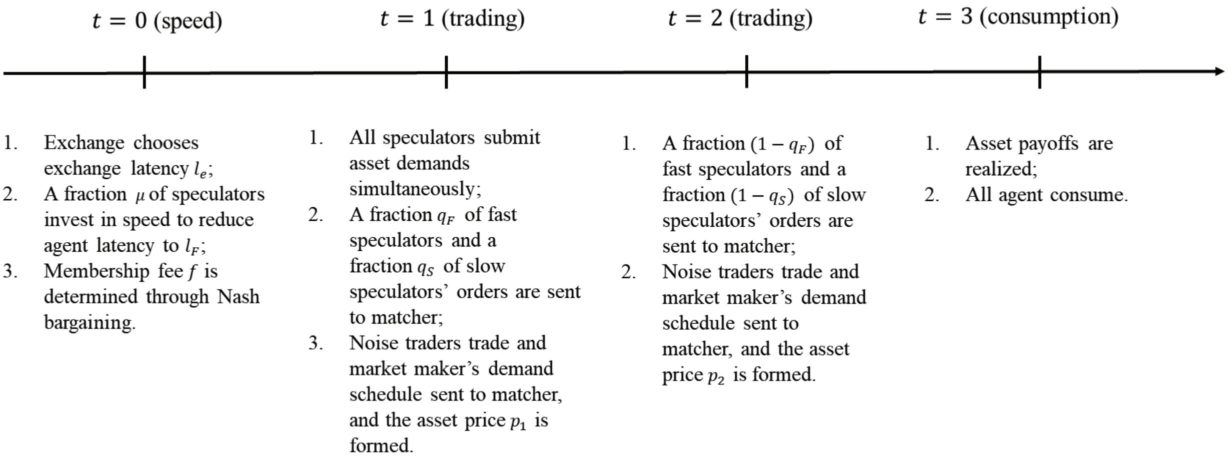


Figure 3-1: Timeline and order of decision making and trading events

through Nash bargaining as introduced below. Time $t \in \{1,2\}$ are trading rounds.[51] Speculator $i(=F,S)$ submits order $x_i = \{\mathbb{x}_i(v; q_F, q_S, \mu)\}|_{i=F,S}$ based on the payoff information v, fraction μ of fast speculators and the trading speeds q_F and q_S of fast and slow speculators. Finally, at $t = 3$, the risky asset pays off and all speculators consume their terminal wealth.

Trading: In each trading round $t = 1, 2$, there is a noise demand z_t, which is *i.i.d.* normal with zero mean and variance σ_z^2, to capture exogenous trading (due to sentiment or hedging). Therefore, the aggregate demand at date $t = 1, 2$ is given by, respectively,

$$w_1 = q_F \mu N x_F + q_S (1-\mu) N x_S + z_1; \tag{3-2}$$

$$w_2 = (1-q_F)\mu N x_F + (1-q_S)(1-\mu) N x_S + z_2. \tag{3-3}$$

There are competitive market makers who set the price based on the aggregate order flow information,

$$p_t = \mathbb{E}[v | \{w_\tau\}_{\tau \le t}]. \tag{3-4}$$

This means that market makers cannot distinguish orders from informed speculators and uninformed liquidity traders; they can only observe the total order flow. Because they are

---

[51] In equilibrium, the fraction q_F of fast traders and q_S of slow traders have their orders executed at date 1 and the other fraction $1 - q_F$ of fast traders and $1 - q_S$ of slow traders have their order executed at date 2. After trading at date 1 or date 2, all speculators finish their trading and liquidation value is revealed. This assumption follows the standard models of informed speculation in which speculators have long horizons and can hold the asset forever. In addition, speculators also have short horizons in that they need to unwind their position before information is released.

risk neutral and competitive, they expect a zero profit. Thus, the market clearing price is the market makers' expectation of v, conditional on what they learn about v from the aggregate order flows.

3.1.2. Speed Demand and Supply

Speed acquisition costs for speculators and exchange: All speculators are slow by default (with agent latency of l_S). One can instead become fast to reduce agent latency from l_S to l_F by paying a speed acquisition cost. We can think of speculators as financial institutions and their speed acquisition activities typically involve both a fixed cost C_F (such as the overhead cost of establishing a research department) and a varied cost $C_V(l_F)$ (such as hiring programmer to produce computer trading program). We focus on entropy form of the varied speed acquisition cost based on probabilistic choice models,[52]

$$C_V(l_F) = \alpha \times \left[l_F \times \ln\frac{l_F}{l_S} + (1 - l_F) \times \ln\frac{1 - l_F}{1 - l_S}\right], \tag{3-5}$$

where the infrastructure multiplier $\alpha(> 0)$ measures how difficult to improve the speed for speculators. The higher probability l_F he wants to increase from the alternative l_S, the more effort is needed.[53] Intuitively, the fixed cost C_F affects market fraction of fast speculators, while the variable cost multiplier α affects the speed acquisition for the fast speculators, however they can affect each other indirectly.

Trading platforms have made costly investments in trading infrastructure to reduce order execution and communication latencies (i.e., investments in trading

[52] There are many probabilistic choice models in economics literature, with the multinomial logit model as the prime example. These models are usually derived in an additive random-utility approach, with McFadden (1974) as the pioneering contribution. The randomness of choice is then interpreted either in terms of heterogeneity of tastes in a population of decision makers, or in terms of choice attributes hidden to the analyst, see Anderson et al. (1992). We instead consider a single decision maker who needs to make some effort in order to steer the choice probabilities in the desired direction. The higher probability he wants to put on any alternative, the more effort (cost) is needed. With no effort, the choice probabilities are given by an exogenous "default" choice distribution.

[53] In the classical microeconomic model, every economic agent can effortlessly obtain any desired choice alternative with certainty. In most real-life situations, the decision maker cannot guarantee any desired outcome with certainty but can optimally make effort to reduce the uncertainty. For example, a careless driver may drive off the road, an absentminded shopper might buy a wrong item, the CEO who does not carefully prepare meetings, or does not motivate and monitor his staff well, may find that the firm has taken suboptimal actions, etc. In the present model, agent can also obtain any desired choice alternative with certainty; however, the marginal disutility of such effort at zero mistake probability is assumed to be infinite.

infrastructure[54]). Similarly to speculators' speed acquisition cost in (3-5), the cost of reducing exchange latency $C_e(l_e)$ can be described as

$$C_e(l_e) = \alpha_e \times \left[l_e \times \ln\frac{l_e}{l_e^0} + (1 - l_e) \times \ln\frac{1 - l_e}{1 - l_e^0}\right], \tag{3-6}$$

which depends on a default exchange speed l_e^0 and an infrastructure multiplier $\alpha_e(> 0)$ measuring how difficult to improve the speed for the exchange. When the infrastructure multiplier $\alpha_e < N\alpha$, trading platforms has scale effect that encourages exchange to reduce the latency of trading platforms when the scale effect becomes more significant (i.e. when α_e becomes smaller).

Nash bargaining and membership fee: To trade and make profit from their private information, speculators must search for a trading platform to trade on it. Both information (from speculators) and trading venue (from trading platform) are scarce resource and thus should earn an economic rent. Inspired by Gârleanu and Pedersen (2018), we use Nash bargaining as a modelling device to characterize profit sharing between speculators and platform, whose outcome yields a membership fee f for trading on the platform.[55] We let ψ describe the bargaining power of the trading platforms.[56] When $\psi = 1$ the trading platform has full bargaining power and extracts all surplus resulting from informed trading. When $\psi < 1$, speculators as platform buyers have non-negligible bargaining power and can retain some surplus resulting from informed trading.

[54] For instance, the trading latency on the NYSE dropped from 350 milliseconds in 2007 to 5 milliseconds in 2009. In October 2015, the Bombay Stock Exchange operated the fastest platform in the world, with a round-trip latency of 6 microseconds. Within the space of eight years, trading platforms experienced a 50,000-fold speed increase. This indicates that exchange latency is endogenously determined by the costly investment.

[55] Our approach is similar to that of Gârleanu and Pedersen (2018), who employ Nash bargaining to determine the fee charged by mutual funds, and is also consistent with Anand and Galetovic (2000), who argue that even when information sellers have local monopoly power, information prices are often determined through bilateral bargaining between the buyer and the seller.

[56] One might argue that $\psi = 1$ is more applicable in a setting with a single trading platform. Note that in our setting, the single-exchange assumption is merely a simplification, and being the only trading platform does not imply that it can extract all the rents. In fact, if the trading platform were able to extract all rents, no speculators will trade anymore, eliminating any profit opportunity for fast exchanges.

3.2 Equilibrium Characterization

This section characterizes the equilibrium from short, medium, and long-term perspectives, respectively. The differences among these equilibria are whether speculators and trading platforms can endogenously adjust their speed investment. As summarized in Table 3-1, speculators can make endogenous speed acquisition decision in medium and long term while trading platforms can only do this in long term. Apart from common exogenous parameters $\mathcal{E} = \{N, \sigma, \sigma_z, l_S\}$, the short, medium, and long term economies can be defined by several tuples of exogenous parameters $\mathcal{E}_s = \{\mathcal{E}, l_F, \mu, l_e\}$, $\mathcal{E}_m = \{\mathcal{E}, C_F, \alpha, l_e\}$ and $\mathcal{E}_l = \{\mathcal{E}, C_F, \alpha, l_e^0, \alpha_e, \psi\}$.

3.2.1. Short Term Equilibrium: Financial Market Equilibrium

Without endogenous speed investment adjustment, short term equilibrium can be treated as financial market equilibrium. As in Kyle (1985), we search for a perfect Bayesian equilibrium. Because fast (slow) speculators are ex-ante identical, we look for symmetric equilibrium of financial market in which all fast (slow) speculators choose the same demand x_F^* (x_S^*).

Definition 3.1: *A short term equilibrium (financial market equilibrium) is characterized by price functions* $\mathbb{p}_t(\{w_\tau\}_{\tau \le t}; \mathcal{E}_s)$ *and demand schedules* $\{\mathbb{x}_k(v; \mathcal{E}_s)\}|_{k=F,S}$ *of fast and slow speculators such that in trading round* $t = \{1, 2\}$,

1) *demand schedules* $\{\mathbb{x}_k(v; \mathcal{E}_s)\}|_{k=F,S}$ *maximize fast (or slow) speculator i's expected profit conditional on her information and trading speed; and*
2) *price functions* $\mathbb{p}_t(\{w_\tau\}_{\tau \le t}; \mathcal{E}_s)$ *clear the market almost surely.*

Based on Definition 3.1, the demand x_F^* of fast speculator i solves

$$x_F^* \in \underset{x_F^i}{argmax}\, \mathbb{E}\left\{x_F^i[l_F l_e(v - p_1) + (1 - l_F l_e)(v - p_2)] \,\middle|\, v, \mathcal{E}_s, \{x_F^j\}\big|_{j \neq i} = x_F^*, \{x_S^j\} = x_S^*\right\};$$

where p_t is given by market maker's pricing as in Equation (3-4). Similarly, the demand x_S^* of slow speculator i solves

$$x_S^* \in \underset{x_S^i}{argmax}\, \mathbb{E}\left\{x_S^i[l_s l_e(v-p_1)+(1-l_s l_e)(v-p_2)]\,\middle|\, v, \mathcal{E}_s, \{x_S^j\}\big|_{j\neq i} = x_S^*, \{x_F^i\} = x_F^*\right\}.$$

Therefore, speculators and market makers are strategical in their trading and pricing, characterizing "*forecast the forecasts of others*" (Foster and Viswanathan, 1996). In other words, a trader's trading strategy depends on the trading strategies of other traders and market maker's pricing rules, which in turn affect the trader's trading strategies. This creates an endogenous feedback of "*one agent's strategy affects other agents' strategies that affect himself own strategy*" (Foster and Viswanathan, 1996). The procedure of obtaining the equilibrium follows the standard Kyle (1985) model. Based on market maker's pricing functions, we first solve speculators' optimal investment problem. We then characterize the fixed-point problem of the optimally strategic trading among traders. Finally, we determine market maker's pricing functions and obtain the following equilibrium. Details of the model derivation are in the appendix.

Proposition 3.1 (Financial market equilibrium): *There exists a linear Bayesian equilibrium, in which each fast (slow) speculator submits a linear demand schedule,*

$$\{\mathbb{x}_k(v;\mathcal{E}_s)\}|_{k=\{F,S\}} = \beta_k(\mathcal{E}_s)v, \qquad (k=F,S), \tag{3-7}$$

while the equilibrium prices satisfy

$$p_1 = \lambda_1(\mathcal{E}_s)w_1, \tag{3-8}$$

$$p_2 = \rho_1(\mathcal{E}_s)w_1 + \lambda_2(\mathcal{E}_s)w_2, \tag{3-9}$$

where the coefficients $(\beta_F, \beta_S, \lambda_1, \rho_1, \lambda_2)$ *are the functions of a tuple of seven exogenous parameters* $\mathcal{E}_s = \{N, \sigma, \sigma_z, l_S, l_F, \mu, l_e\}$ *given in Appendix A1.*

3.2.2. Medium Term Equilibrium: Demand for Speed

In the medium term, speculators can endogenously adjust fast trading investment. Similar to Goldstein and Yang (2017), we study both extensive margin μ^* (i.e., whether to become fast speculator) and intensive margin l_F^* (i.e., how much speed to acquire). Since speculators are ex-ante identical, we look for symmetric equilibrium of speed acquisition in which all fast speculators choose the same optimal agent latency l_F^*.

Definition 3.2: *A medium term equilibrium is characterized by a fraction function $\mathbb{u}(\mathcal{E}_m)$ of fast speculators and an optimal speed demand schedule $\mathbb{l}_F(\mathcal{E}_m)$ of fast speculators such that at $t = 0$,*

1) *speed demand schedule $\mathbb{l}_F(\mathcal{E}_m)$ maximizes fast speculator's expected profit conditional on the speed acquisition cost $C_F + C_V(\mathbb{l}_F(\mathcal{E}_m))$; and*
2) *no slow speculator rushes to be fast given the fraction function $\mathbb{u}(\mathcal{E}_m)$ of fast speculators, that is: i) $\phi(\mathbb{l}_F(\mathcal{E}_m), \mathbb{u}(\mathcal{E}_m)) = C_F + C_V(\mathbb{l}_F(\mathcal{E}_m))$ for $0 < \mathbb{u}(\mathcal{E}_m) < 1$; ii) $\phi(\mathbb{l}_F(\mathcal{E}_m), 0) \leq C_F + C_V(\mathbb{l}_F(\mathcal{E}_m))$ for $\mathbb{u}(\mathcal{E}_m) = 0$ and iii) $\phi(\mathbb{l}_F(\mathcal{E}_m), 1) \geq C_F + C_V(\mathbb{l}_F(\mathcal{E}_m))$ for $\mathbb{u}(\mathcal{E}_m) = 1$, where $\phi(l_F, \mu)$ is the value of speed acquisition define by*

$$\phi(l_F, \mu) = \mathbb{E}\{\mathbb{E}\{\mathbb{x}_F(v; \mathcal{E}_s)[l_F l_e(v - p_1) + (1 - l_F l_e)(v - p_2)] \\ - \mathbb{x}_s(v; \mathcal{E}_s)[l_s l_e(v - p_1) + (1 - l_s l_e)(v - p_2)] | v, \mathcal{E}_s\} | \mathcal{E}_m, \mu, l_F\}.$$

To characterize the equilibrium, we first derive two constrained equilibrium variants in which either the fraction of fast speculators is exogenous given or fast speculators have exogenous speed. The two constrained economies can be described by the best response functions (3-10) and (3-11) below, which then jointly determine the equilibrium for speculators' speed demand in Proposition 3.2 with respect to the optimal fraction of fast speculators μ^* and the optimal speed l_F^*. When the equilibrium choice of other fast speculators l_F^* and fraction of fast speculators μ are given, at time $t = 0$, fast speculator i investes in speed technology l_F^i to maximizes his expected payoff,

$$l_F^* \in \underset{l_F^i}{argmax}\, \mathbb{E}\left\{\mathbb{E}\{\mathbb{x}_F(v; \mathcal{E}_s)\left[l_F^i l_e(v - p_1) + \left(1 - l_F^i l_e\right)(v - p_2)\right] - C_F - C_V\left(l_F^i\right) | v, \mathcal{E}_s\} \middle| \mathcal{E}_m, \mu, \{l_F^j\}\big|_{j\neq i} \\ = l_F^*\right\}.$$

The first order condition then implicitly determines the optimal l_F^* for given value of μ,

$$l_F^* = l_F(\mu; \mathcal{E}_m). \tag{3-10}$$

We then examine the optimal endogenous extensive margin given the intensive margin l_F. A speculator chooses to become fast only when his extra expected payoff from fast trading is sufficient to cover the cost of acquiring the speed, compared to be a slow speculator. In equilibrium, by taking into account the costs, a speculator is indifferent between being fast and slow, which generates an interior equilibrium solution. Otherwise, a corner equilibrium may arise and all speculators stay slow (rush to be fast) when the speed cost is too high (low). Corollary 3.1 describes the optimal response function of μ^* for given value of l_F.

Corollary 3.1: *The optimal response function of μ^* for given value of l_F (and exchange latency l_e) is expressed as*

$$\mu^* = \begin{cases} 1, & C_F \leq \phi(l_F, 1) - C_V(l_F) \\ \mu(l_F; \mathcal{E}_m), & \phi(l_F, 1) - C_V(l_F) < C_F < \phi(l_F, 0) - C_V(l_F) \\ 0, & C_F \geq \phi(l_F, 0) - C_V(l_F) \end{cases}, \tag{3-11}$$

where $\mu(l_F; \mathcal{E}_m)$ is implicitly determined by $\mu^* \in \{\mu | \phi(l_F, \mu) = C_F + C_V(l_F)\}$.

Based on the above analysis, Proposition 3.2 describes the overall equilibrium.

Proposition 3.2: *Given the tuple of exogenous parameters $\mathcal{E}_m = \{N, \sigma, \sigma_z, l_S, C_F, \alpha, l_e\}$, the two best response functions* (3-10) *and* (3-11) *jointly determine the equilibrium intensive and extensive margins,*

$$l_F^* = \mathbb{l}_F(\mathcal{E}_m). \tag{3-12}$$

$$\mu^* = \mathbb{u}(\mathcal{E}_m). \tag{3-13}$$

Furthermore, relative to some thresholds $\hat{l}_e(C_F)$ and $\check{l}_e(C_F)(< \hat{l}_e)$ (defined in Appendix A3) there exists a unique equilibrium such that

> ***1) (corner)*** *all speculators stay slow ($\mu^* = 0$) when $l_e \geq min\{\hat{l}_e(C_F), 1\}$;*
>
> ***2) (interior)*** *a fraction μ^* of speculators acquire the speed with the optimal speed level of l_F^* when $max\{\check{l}_e(C_F), 0\} \leq l_e < min\{\hat{l}_e(C_F), 1\}$, while the equilibrium μ^* and l_F^* are determined by equations* (3-12) *and* (3-13)*;*
>
> ***3) (corner)*** *all speculators acquire the speed ($\mu^* = 1$) when $l_e < max\{\check{l}_e(C_F), 0\}$, while the optimal speed level is determined by* (3-12)*.*

Based on Proposition 3.2, Figure 3-2 (a) describes an overview for speculators' endogenous speed acquisition decision in equilibrium. With the chosen parameter values, it illustrates the regimes of different types of equilibrium in the parameter space

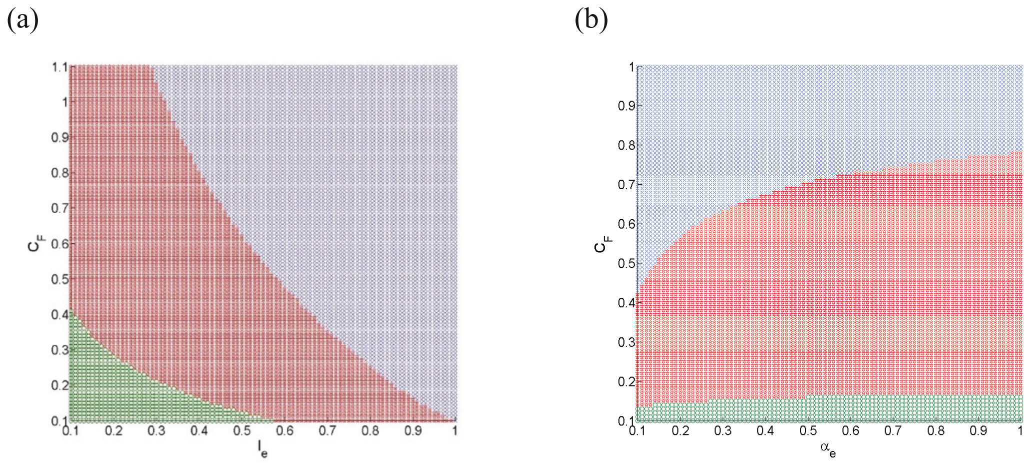


Figure 3-2

Equilibrium types for exogenous and endogenous exchange speed acquisition. Panels (a) and (b) plot the equilibrium types in parameter space of (l_e, C_F) and (α_e, C_F) respectively, where parameter l_e controls the exchange latency, α_e denotes infrastructure multiplier for the exchange, and C_F denotes the fixed cost of speculators' speed acquisition. The green, red, and blue regions represent the equilibrium fraction of fast speculators with $\mu = 1$, $0 < \mu < 1$, and $\mu = 0$ respectively. The variable cost of speed acquisition takes an entropy form in (3-5) and (3-6). Apart from the common parameter values $N = 100, \sigma = 20, \sigma_z = 10, l_S = 0.4, \alpha = 0.5$ in panel (a), the extra parameter value for panel (b) $l_e^0 = 0.35$.

of $\{C_F, l_e\}$.[57] In general, as C_F and l_e become larger, an equilibrium with a lower μ^* is more likely to prevail. Intuitively, an increase in the fixed cost reduces the absolute benefit to be fast, while a reduction in exchange latency reduces relatively the first-mover advantage, reflecting a crowding-out effect of faster exchange speed on fast trading. This contributes to the determination of endogenous speed supply.

[57] Specifically, depends on the speed acquisition cost and exchange latency, there are three types of endogenous equilibria. First, when C_F, the fixed cost in speed acquisition, is sufficiently small, all speculators become fast, that is $\mu^* = 1$; a case that mimics Verrecchia (1982) in information acquisition literature. Here, speculators choose to become fast when all others are fast. Second, when C_F takes an intermediate value, an interim proportion of the speculators chooses to become fast: $\mu^* \in (0, 1)$. The equilibrium is pinned down by a condition that makes traders indifferent between producing the equilibrium amount of speed and remaining slow, characterized by (3-11) and the first-order condition (3-10) that guarantees an optimal level of the speed. Finally, when C_F is sufficiently large, no speculator chooses to become fast ($\mu^* = 0$).

3.2.3. Long Term Equilibrium: Supply of Speed

In the long term equilibrium, both speculators and trading platform can adjust their fast trading investment optimally, which is characterized in Definition 3.3.

Definition 3.3: *A long-term equilibrium is characterized by optimal exchange latency $l_e^* = \mathbb{l}_e(\mathcal{E}_l)$ of the trading platform and membership fee function $\mathbb{f}(l_e, \mathcal{E}_l)$ such that at date $t = 0$:*

1) *the membership fee $\mathbb{f}(l_e, \mathcal{E}_l)$ is determined via Nash bargaining between the speculators and trading platform; and*
2) *$l_e^* = \mathbb{l}_e(\mathcal{E}_l)$ maximizes the exchange's expected profit.*

To characterize the long term equilibrium, the membership fee is determined as an outcome of Nash bargaining between the speculators and trading platforms[58] to maximize the product of the utility gains from the agreement,

$$\max_f (R_i - f)^{1-\psi} f^{\psi} \implies f^* = \psi R_i. \tag{3-14}$$

Here, ψ describes the bargaining power of the trading platforms. The exchange charges speculator a membership fee f so that the speculator can trade on the trading platform. In equilibrium, based on the Nash bargaining, the net profit R_i generated from speculator i is divided between the trading platform and the speculator as follows,

$$R_i = f^* + (R_i - f^*) = \underbrace{\psi R_i}_{exchange's\ rent} + \underbrace{(1-\psi) R_i}_{speculators'rent}. \tag{3-15}$$

The shares of the gains are determined by the bargaining power $(1 - \psi)$ of the speculator in negotiating prices with trading platform. For $\psi < 1$, speculator has non-negligible bargaining power and retains some surplus resulting from informed trading. Corollary 3.2 characterizes the membership fee function $\mathbb{f}(l_e, \mathcal{E}_l)$.

---

[58] Recall that the bargaining equilibrium depends on agent's utilities in the event of agreement or no agreement (the latter is called the "outside option"). For trading platform, if an agreement is reached, its additional profit is given by the membership fee f, because the trading platform's speed acquisition cost C_e is sunk and there is no marginal cost to taking on more speculators. For speculator, the utility in an agreement of a fee of f is $(R_i - f)$. If no agreement is reached, the speculator's outside option is no-trading, yielding a zero profit. Hence, the speculator's gain from the agreement is $(R_i - f)$. At the bargaining stage, the speculator's acquisition cost is also sunk.

Corollary 3.2: *Given exchange latency l_e, the optimal membership fee for the trading platform can be described by*

$$f^* = \begin{cases} \psi \bar{\pi}_S(0, l_e) & , \quad l_e \geq min\{\hat{l}_e(C_F), 1\} \\ \psi[\bar{\pi}_F(\mathbb{u}(\mathcal{E}_m), \mathbb{l}_F(\mathcal{E}_m), l_e; \mathcal{E}_m) - C_F - C_V(\mathbb{l}_F(\mathcal{E}_m))], & max\{\check{l}_e(C_F), 0\} \leq l_e < min\{\hat{l}_e(C_F), 1\} \\ \psi[\bar{\pi}_F(1, \mathbb{l}_F(\mathcal{E}_m), l_e; \mathcal{E}_m) - C_F - C_V(\mathbb{l}_F(\mathcal{E}_m))] & , \quad l_e < max\{\check{l}_e(C_F), 0\}, \end{cases}$$

in which parameters $\hat{l}_e(C_F)$ and $\check{l}_e(C_F)$ are defined in Proposition 3.2, while $\bar{\pi}_F$ and $\bar{\pi}_S$ are defined by

$$\bar{\pi}_k = \mathbb{E}\{\mathbb{E}\{\mathbb{x}_k(v; \mathcal{E}_s)[l_k l_e(v - p_1) + (1 - l_k l_e)(v - p_2)]|v, \mathcal{E}_s\}|\mathcal{E}_m\}, \ \ k = F, S.$$

We finally examine the exchange's decision on the speed supply and the overall equilibrium. The trading platform chooses exchange latency to maximize operation profit,

$$l_e^* \in \underset{le}{argmax}\left\{N\mathbb{f}(l_e, \mathcal{E}_l) - \alpha_e \times \left[l_e \times \ln\frac{l_e}{l_e^0} + (1 - l_e) \times \ln\frac{1 - l_e}{1 - l_e^0}\right]\right\}. \tag{3-16}$$

This problem can be solved in two steps. First, we use the financial market equilibrium condition in Proposition 3.1 and speculators' endogenous speed acquisition conditions in Proposition 3.2 to jointly solve for intensive and extensive margins. Second, based on the results in the first step, we calculate the net profits for speculators and use Nash bargaining and Corollary 3.2 to solve (3-16) for the optimal latency for the exchange. This generates the optimal exchange latency

$$l_e^* = \mathbb{l}_e(\mathcal{E}_l). \tag{3-17}$$

Depending on the exogenous parameter values, we obtain either interior or corner equilibrium solution, which is summarized in Proposition 3.3 and illustrated numerically in Figure 3-2 (b).

Proposition 3.3: *Given a tuple of exogenous parameters $\mathcal{E}_l = \{N, \sigma, \sigma_z, l_S, C_F, \alpha, l_e^0, \alpha_e, \psi\}$, depending on the fixed speed acquisition cost C_F relative to some thresholds $\hat{C}_F(\alpha_e)$ and $\check{C}_F(\alpha_e)(< \hat{C}_F)$ (defined in Appendix A5),*

1) (corner) *all speculators stay slow ($\mu^* = 0$) when $C_F \geq \hat{C}_F(\alpha_e)$;*

2) (interior) *a fraction μ^* of speculators acquire the speed with the optimal speed level of l_F^* when $\check{C}_F(\alpha_e) \leq C_F < \hat{C}_F(\alpha_e)$, while the equilibrium μ^* and l_F^* are jointly determined by the optimal exchange latency $\mathbb{l}_e(\mathcal{E}_l)$ and equations* (3-12) *and* (3-13)*;*

3) (corner) *all speculators acquire the speed* ($\mu^* = 1$) *when* $C_F < \breve{C}_F(\alpha_e)$*, while the optimal speed level is determined by equation* (3-12).

Figure 3-2 (b) provides an overview for the endogenous speed acquisition decisions for speculators and the exchange in equilibrium. With the chosen parameter values, it illustrates the regimes of different types of equilibrium in parameter space of $\{C_F, \alpha_e\}$. In general, as C_F becomes larger and α_e becomes smaller, an equilibrium with a lower μ^* is more likely to prevail. Higher C_F reflects more cost to become fast; while lower α_e captures the crowding out effect of the exchange to fast trading. Intuitively, with relatively cheaper speed acquisition cost α_e, the exchange has more incentive to reduce exchange latency, which reduces the relative first-mover advantage of the fast, crowding out fast speculators.

3.3 Medium Term Equilibrium: Excessive Speed Demand from Speculators

Electronic trading groups that made fortunes by being first to pounce on favorable prices are finding it tougher to stay ahead in markets where transactions unfold in billionths of a second. Their costs have risen while volatility and volumes, essential elements for profitable trading, have ebbed. As a result, only the largest half dozen or so trading companies remain in the race for speed. In a sign of relentless cost pressures, they are finding ways to economize so that the fixed cost to be fast speculators decreases.[59] This can increase the relative cost of the speed for fast speculators but reduce overall sensitivity to the cost due to technology innovations. Therefore, investment in trading technology affects intensive and extensive margins as well as their interaction. With the medium term equilibrium, this section studies the implications of an advancing fixed and varied

[59] Top groups including DRW, IMC, Jump Trading and XR Trading have banded together to construct a speedy wireless-and-cable route between Chicago and Tokyo called Go West, according to a person briefed on the project. These stylized facts just reflect that the fixed cost to be fast for speculators decreases.

speculators speed technology (decreasing in parameters C_F and α, respectively) on speculators' investment in fast trading technology. We show that this can lead to excessive speed investment from speculators, a market failure to spurs speed arms race in HFTs.

3.3.1. Cross Effects: Complementarity and/or Substitution

Advancing in fixed and varied speed technology can affect market fraction of fast speculators and their optimal speed directly and indirectly (named as cross effect below), characterized in the following proposition.

Proposition 3.4: *In the interior equilibrium, an advancing in C_F promotes fast population but demotes fast speculators' speed; while an advancing in α promotes the speed of fast speculators and promotes (demotes) fast population when α is high (low). Mathematically, advancing speed technology has a direct effect: $\partial\mu/\partial C_F < 0$ and $\partial l_F/\partial\alpha < 0$; and a cross effect: $\partial l_F/\partial C_F > 0$ and $\partial\mu/\partial\alpha > 0$ (< 0) for small (large) α.*

The result on the direct effect is intuitive; (i) the lower is the cost C_F to be the fast, the higher is the fraction of fast speculators; (ii) the lower is the cost α to reduce agent latency, the lower is agent latency. These direct effects are illustrated in the blue solid line in Figure 3-3 (a) and the red dotted line in Figure 3-3 (b), respectively. The red dotted line in Figure 3-3 (b) and blue solid line in Figure 3-3 (b) illustrate the indirect cross effects of α on μ and C_F on l_F, respectively, showing that lowing C_F increases agent latency, while lowing α can be either complementarity or substitution to μ. Intuitively, lowing the fixed cost increases market fraction of fast speculators and hence the competition among the fast, which dominates the marginal benefit of lowing agent latency. On the other hand, the effect of lowing the variable cost on the fraction of the fast is ambiguous, depending on the trade-off between the value and cost of the speed.

To better understand the indirect cross effects, particularly the nonmonotone effect, we consider two constrained model variants (best response functions (3-10) and (3-11))

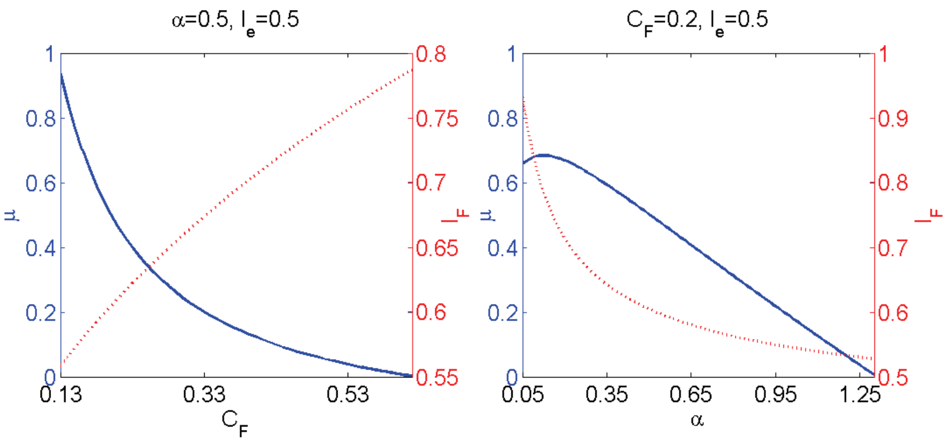


Figure 3-3

The influence of speculators' speed acquisition cost on speed acquisition. This figure plots the effects of advancing speed technology with respect to C_F (in the left panel) and α (in the right panel) on speed acquisition with respect to market fraction μ of fast speculators (on the left axis) and the optimal agent latency l_F (on the right axis), respectively, where the exchange variable cost function parameter $\alpha_e = 0.7$ and other parameter values are the same as Figure 3-2.

to examine the cross effect between the intensive and extensive margins and how they are further influenced by various competition effects.

The indirect effect of α to the fast population μ illustrated be the blue solid line in Figure 3-3 (b) depends on the cross effect of l_F on μ. The left and right panels in Figure 3-4 (a) illustrate the determination of constrained model variant $\mu^* = \mu(l_F; \mathcal{E}_m)$ for given l_F. With lowering fast speculators' agent latency (i.e., higher l_F; specifically, moving from scenario 1 to 2 in the left panel or from scenario 3 to 4 in the right panel), both the value of the speed (Φ) and the speed acquisition cost (C) increase, no matter under compliment or substitute scenario.[60] However, due to the competition effect, as fast speculators' agent latency decreases, the marginal cost of being faster is increasing, becoming more significant (from C_3 and C_4 than from C_1 and C_2), while the marginal revenue of being fast is decreasing (indicated by a smaller reduction from ϕ_3 to ϕ_4 than

[60] On the speed acquisition cost, fast speculator needs to make some effort in order to steer the choice probability (l_F) in the desired direction. The higher probability he wants to put on the alternative, the more effort is needed, the higher is the cost. On the value of speed, ever-fast trading brings fast speculator more speed advantage. Noticing that the fraction of fast speculators is relatively low under all the scenarios. Even though unconscious about more intensified competition, this can affect fast speculators' revenue positively due to relatively low fraction of fast speculators.

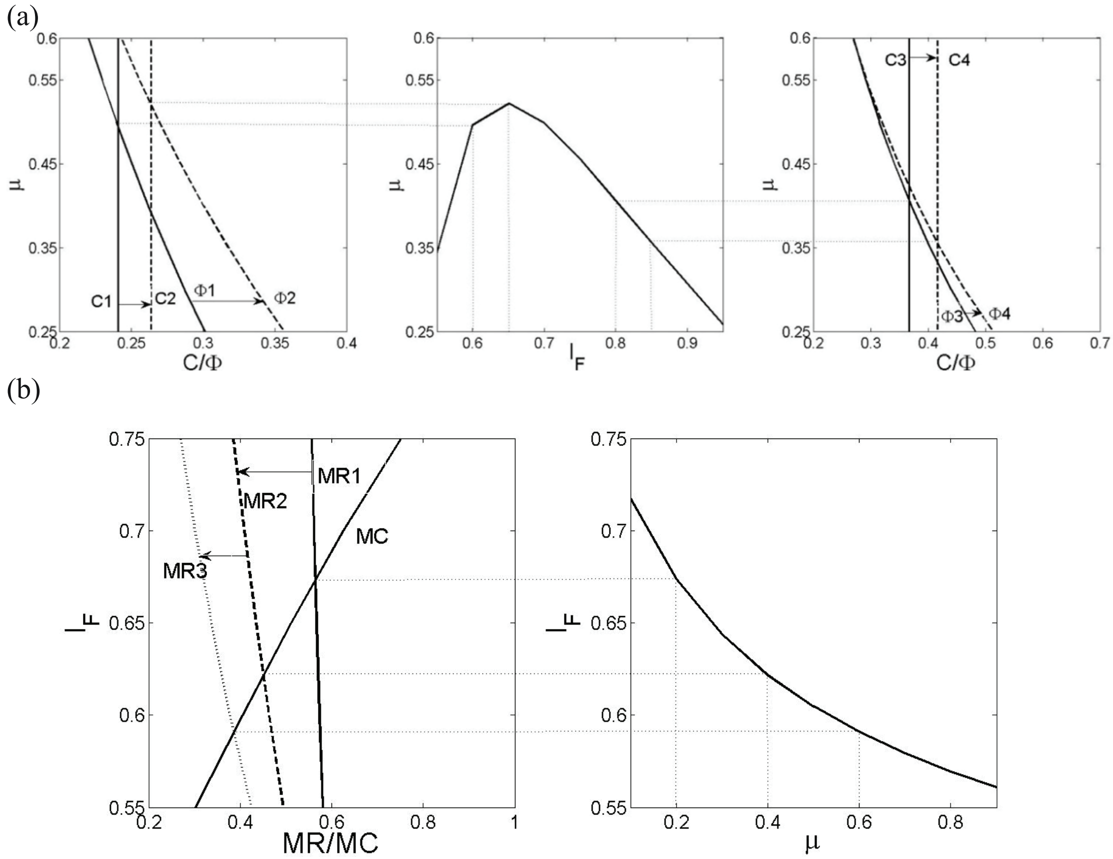


Figure 3-4

The cross effect and its decomposition. This figure illustrates how an increasing in either the fraction of fast speculators (the bottom panels) or the equilibrium speed advantage (the top panels) affects each other. Panel (a) shows that ever-fast trading can either promotes the incentive to become fast when agent latency is high or demotes the incentive to become fast when agent latency is low. Panel (b) shows that higher fraction of fast speculators demotes the optimal investment in the speed to reduce the agent latency. In panel (a), the value of the speed and speed acquisition cost jointly determine the equilibrium extensive margin. In panel (b), the marginal revenue and marginal cost jointly determine the equilibrium intensive margin. The parameter values are: $N = 100, \sigma = 20, \sigma_z = 10, l_S = 0.4, l_e = 0.5, \alpha = 0.5, C_F = 0.2$.

from ϕ_1 to ϕ_2). Therefore, when fast speculators' agent latency is high, the increasing in the value of speed dominates the trade-off and increases the fraction of fast speculators, leading to a *complementarity effect*. However, when the fast speculators' agent latency is low, the increase in the cost dominates the trade-off and reduces the fraction of fast speculators, leading to a *substitution effect*. Both the cross effects are illustrated in the middle panel of Figure 3-4 (a) for the constrained model variant $\mu^* = \mu(l_F, l_e)$ defined

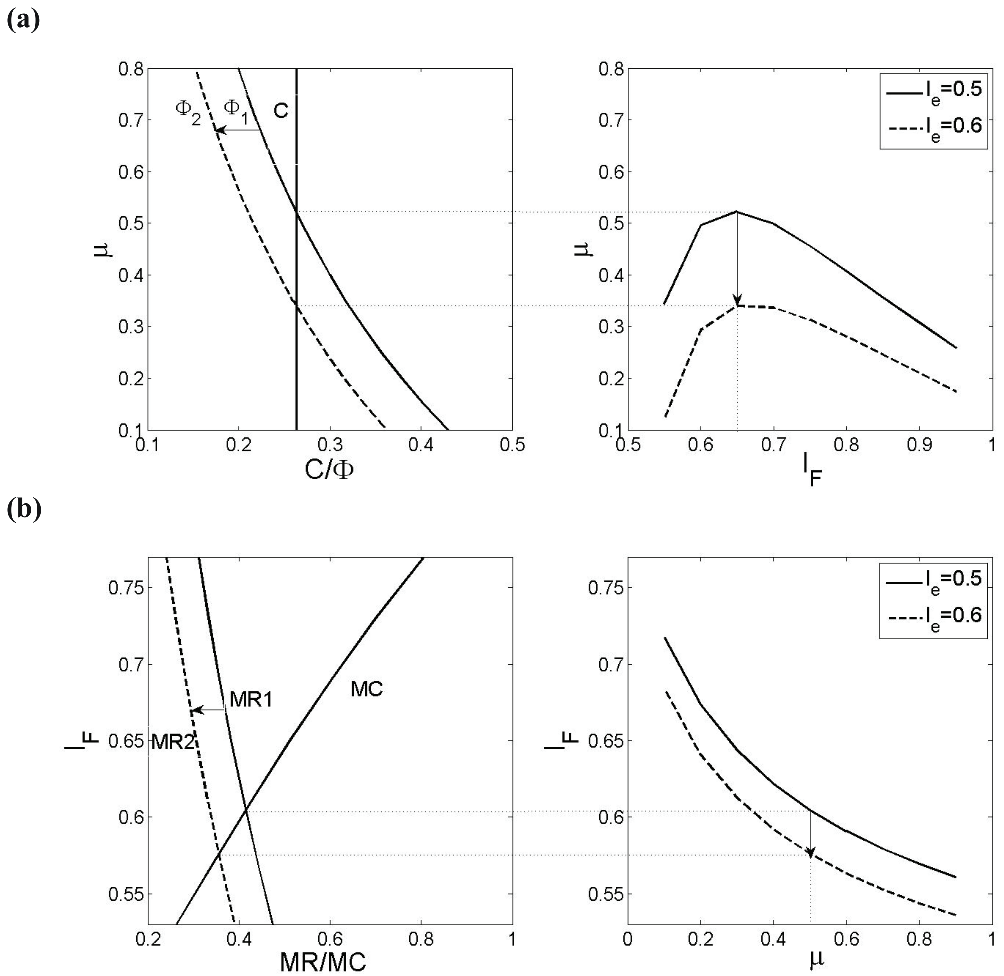


Figure 3-5

The influence of lowering exchange latency on the cross effect. This figure illustrates how an increasing in exchange speed affects the fraction of fast speculators (the top panels) and the optimal speed of the fast (the bottom panel) for given speed of fast trading and the fraction of the fast, respectively. In panel (a), the value of speed and speed acquisition cost jointly determine the equilibrium extensive margin; though the exchange latency has no influence on the speed acquisition cost, the value of speed decreases with more fast speculators. In panel (b), the marginal revenue and marginal cost jointly determine the equilibrium intensive margin; with increasing exchange speed, the marginal cost is not influenced while the marginal benefit shrinks (the left shifting MR curves). Parameters are the same as in Figure 3-4.

by (3-11). Combining the direct effect of α on l_F in Figure 3-3 (b), this cross effect leads to the nonmonotonic indirect effect of α on μ in Figure 3-3 (b).

The indirect effect of C_F on the fast trading latency l_F illustrated by the red dotted line in Figure 3-3 (a) depends on the cross effect of μ to l_F. The left panel in Figure 3-4 (b) depicts the determination of the constrained model variant $l_F^* = l_F(\mu; \mathcal{E}_m)$ for given μ.

With a higher fraction of fast speculators (moving from scenario 1 to 2 or from 2 to 3), based on equation (3-5), the marginal cost (hence the MC curve) remains the same for all three different scenarios. In contrast, the marginal revenue (MR) shrinks due to the intensified competition among fast speculators, indicated by the left shifting MR curves. Due to the "first-mover advantage", the speed technology allows fast speculator to execute his trade faster so that his order has higher chance to be executed before the information is capitalized into the prices. When the fraction of fast speculators increases, the competition in earlier period also increases. In this circumstance, an extra speed advantage brings less benefit (MR) comparing to a less intensifies competition situation. In general, the left shifting MR curve and the constant MC curve jointly lead to the decreasing l_F^* with respect to μ, illustrated in the right panel of Figure 3-4 (b) for the constrained model variant $l_F^* = l_F(\mu; \mathcal{E}_m)$ defined by (3-10). Combining the direct effect of C_F on μ in Figure 3-3 (a), this cross effect leads to the indirect effect of C_F on l_F in Figure 3-3 (a). We summarize these cross effects in Corollary 3.3.

Corollary 3.3*: In the interior equilibrium, more fast speculators demotes the optimal investment in the speed to reduce the agent latency; while lower latency of fast speculators can promote (demote) the fast population when fast speculators' agent latency is high or (low). Mathematically, $\partial l_F(\mu; l_e)/\partial\mu < 0$ and $\partial\mu(l_F; l_e)/\partial l_F > 0$ (< 0) for small (large) l_F.*

Corollary 3.3 implies less concentrated fast trading with the decreasing in the fixed cost. This means that, as the fixed cost of fast trading reduces, the fraction of fast speculators increases, while the optimal fast trading speed decreases. The effect of the exchange latency on fast trading can also be examined and results are summarized as follows.

Corollary 3.4*: In the interior equilibrium, lowering exchange latency demotes fraction of fast speculators and fast speculators' speed under both constrained model variants (3-10) and (3-11), respectively. Mathematically, $\partial l_F(\mu; \mathcal{E}_m)/\partial l_e < 0$ and $\partial\mu(l_F; \mathcal{E}_m)/\partial l_e < 0$.*

Corollary 3.4 is illustrated by the right panels (a) and (b) of Figure 3-5, showing that lowering exchange latency demotes both the optimal investment in the speed and the fraction of fast speculators. The left panels (a) and (b) of Figure 3-5 demonstrate a decomposition of the impact of exchange latency through equations (3-10) and (3-11). They show that the reduction in the exchange latency has no influence on the speed acquisition cost and marginal cost (from equation (3-5)) but decreases the value of speed and marginal revenue. Specifically, lowering exchange latency spurs the speed of both fast and slow speculators, hence more speculators trade at earlier period. Fast speculators, who trade relatively earlier, outperform the slow speculators due to the "first-mover advantage". With lower exchange latency, speculators, in general, become faster and the competition in the early period becomes more intensified. This reduces not only fast speculators' speed advantage, i.e., the value of speed, but also the marginal benefit from the speed advantage, i.e., marginal revenue. In general, the left shifting in the value of the speed and the fixed speed acquisition cost jointly lead to a decrease of μ^* with respect to l_e. Similarly, the left shifting MR curve and constant MC curve jointly lead to the decreasing l_F^* with respect to l_e. Therefore, lowering exchange latency reduces the incentives for speculators to become fast and to trade faster.

3.3.2. Excessive Speed Demand

High frequency trading, relying on sophisticated computer programs and gaining tremendous revenue, is very attractive for Wall Street's companies..[61] With the increasing fast trading and decreasing profitability of HFT, this section analyses the impact on high-frequency trading arms race and explores market failure with respect to excessive speed investment of speculators. This is illustrated in the left panel of Figure 3-6. The solid line shows the net profit $R_F = \bar{\pi}_F - C_F - C_V(\, or \, R_S = \bar{\pi}_S)$ for speculators (same for both fast and

[61] Ever since the SEC introduced regulations for alternative trading systems, fast trading technology started to take flight. Compare to the situation around 2000s when HFT represented less than 10% trading shares, they accounted for nearly 50% of all equity trades in the US by 2012. According to Popper (2012), "HFT is also rapidly growing in Europe and Asia, accounting for approximately 45% of stock trading volume in the European Union, 40% in Japan, and 12% in the rest of Asia as of late 2012". Although massive investment in HFT has already been seen over the last decades, anecdotal evidence suggests that the profitability of high-frequency traders has decreased in recent years (Bias et al. 2015).

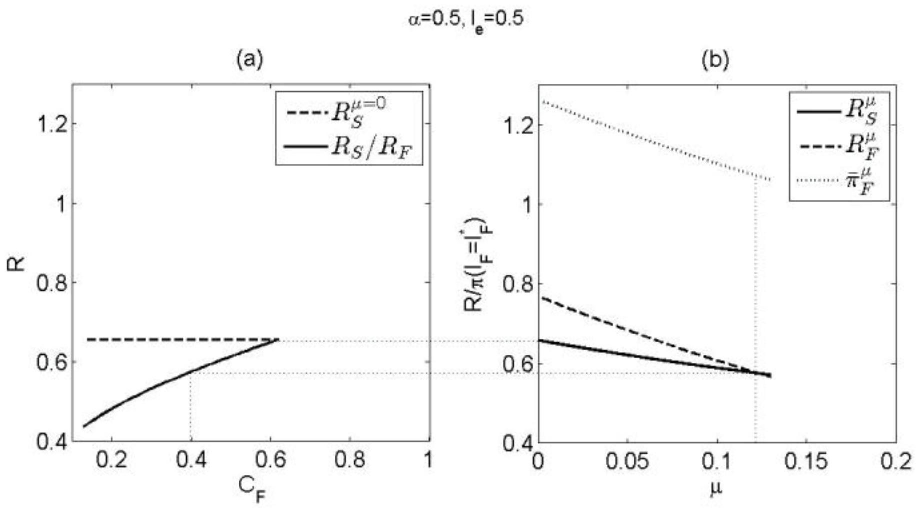


Figure 3-6
The influence of technology advancement on the gross, net profit, and speed acquisition cost. The left panel plots the net profit under both equilibrium and no speed investment scenarios, while the right panel shows the evolution of speed over-investment with respect to C_F. The parameter values are the same as in Figure 3-4.

slow speculators in interior equilibrium) under the endogenous speed acquisition equilibrium; while the blue dotted line shows the net profit $R_S^{\mu=0}$ when all speculators stay slow. It delivers two main results.

Firstly, the endogenous speed acquisition of speculators always induces excessive speed investment. This is illustrated in the right panel of Figure 3-6 showing the evolution of extensive margin equilibrium for given optimal intensive margin l_F^*. Speculators, who rush to trade fast through investment in speed technology, profit from their first-mover advantage. This is indicated by the higher net profit being fast $R_F^{\mu<\mu^*}$ comparing to staying slow $R_S^{\mu<\mu^*}$ for low fraction of fast speculators before reaching the equilibrium μ^*. Therefore, fast speculators do not internalize the negative externality that their investment speed imposes on other market participants. The increasing trading from more fast speculators moves equilibrium prices closer to the fundamental value, which shrinks the slow speculators net profit (indicated by the decreasing R_S^{μ}, i.e., $\partial R_S^{\mu}/\partial\mu < 0$). Any arbitrage profit opportunity between fast and slow speculators makes more speculators become fast, which further reduces slow speculators revenue. In equilibrium, netting the speed acquisition cost, fast and slow speculators have the same net profit. That is, the fraction of fast speculators increases until $R_F^{\mu=\mu^*} = R_S^{\mu=\mu^*}$ (which is at the level l_F^*). This net profit in equilibrium is significantly lower comparing to no speed investment situation $R_F^{\mu=\mu^*} = R_S^{\mu=\mu^*} < R_S^{\mu=0}$, illustrated in the left panel of Figure 3-6. Therefore, because speculators do not internalize the loss of slow speculators, the speed is over invested.

Secondly, the advancing in speculators speed technology makes it cheaper acquire speed. As showed in Proposition 3.4 and Figure 3-3, technology advancing either stimulates more traders becoming fast or promotes fast speculators to acquire more speed. Hence, the intensified competition between fast and slow speculators reduce equilibrium profit. This implies that speed demand becomes even more excessive with respect to technology advancing.

3.4 Long Term Equilibrium: Supply of Speed by Exchange

Recent decades, exchanges have spent thousands of hundreds of dollars to build fast trading infrastructures, trying to cater to their quickest customers and supply high-frequency trading shops with ever-improved connections.[62] This in turn affects fast trading behavior of speculators. In this section, we analysis the long-term equilibrium with focuses on the supply of speed by exchange. Specifically, we are interested in how the advancing in exchange' speed supply technology feeds back into speculators' speed demand and financial market by comparative statics with respect parameters governing the trading platform' speed supply cost α_e. We further examine how faster exchange

[62] Financial exchanges have made costly investments in trading infrastructure to reduce order execution and communication latencies. For instance, the trading latency on the NYSE dropped from 350 milliseconds in 2007 to 5 milliseconds in 2009. In October 2015, the Bombay Stock Exchange operated the fastest platform in the world, with a round-trip latency of 6 microseconds. Within the space of eight years, trading platforms experienced a 50,000-fold speed increase. This process has gone beyond equities and futures to options, bonds, and currency exchange markets.

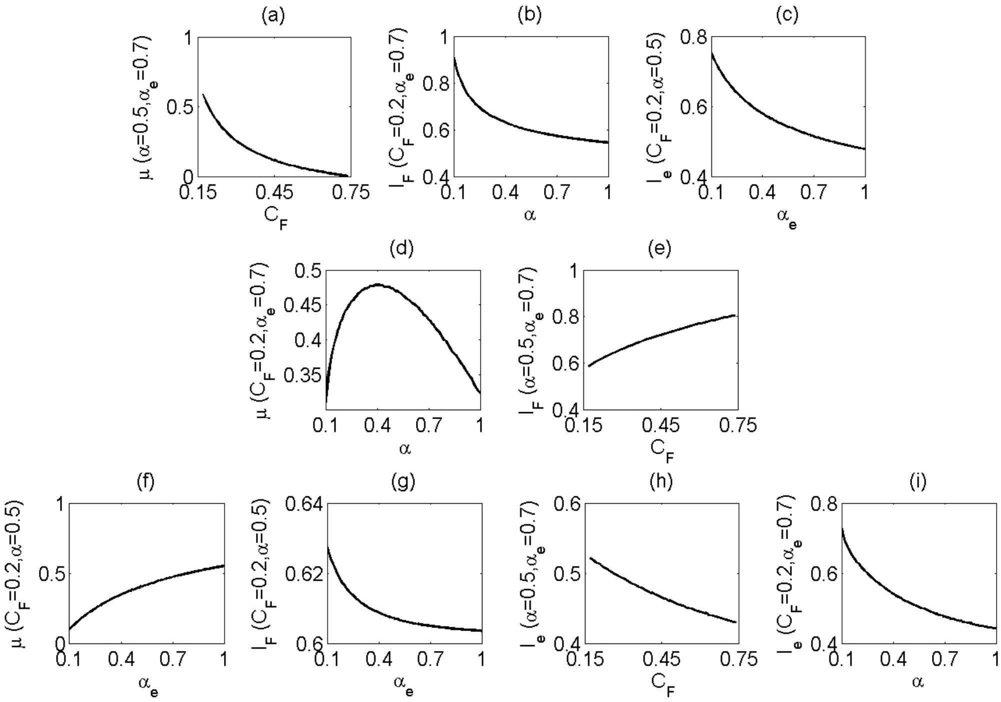


Figure 3-7
Effects of the fixed cost, c_F, cost multipliers, α and α_e of speed acquisition for the fast speculators and the exchange on equilibrium fraction of fast speculators, μ, and the speed latencies, l_F and l_e, of the fast speculators and the exchange. Panels (a), (b) and (c) plot the direct effects of c_F, α and α_e on μ, l_F and l_e, respectively; panels (d) and (e) plot the spill-over effects of α and c_F on μ and l_F, respectively; panels (f) and (g) plot the spill-over effects of α_e on μ and l_F, respectively; panels (h) and (i) plot the spill-over effects of c_F and α on l_e, respectively; where the exchange variable cost function parameter $\alpha_e = 0.7$ and other parameter values are the same as in Figure 3-4.

influence market quality, which plays important role in degerming socially optimal exchange speed as discussed in Section 3.5.

3.4.1. Equilibrium Properties

From long-term equilibrium perspective, both speed demand from the speculators and speed supply by the exchange are endogenously determined by cost parameters C_F, α, and α_e as illustrated in Table 3-1. The influence of technology advancements is examined through the comparative analysis with respect to parameters controlling the cost of being fast. The results illustrated in Figure 3-7 can be divided into three categories, i.e., directly effect, cross effect and spill-over effect. Firstly, the direct effects are intuitive: the lower is C_F, α and α_e, the lower is the cost to be fast, to reduce agent and exchange latencies, and the higher is the fraction of the fast and the lower are the agent and exchange latencies, respectively. These direct effects are illustrate by Figure 3-7 (a), (b) and (c). Secondly, the cross effects of speculator's speed technology on speculator's investment remain robust as in Section 4. Following Proposition 3.5 summarizes the cross effects.

Proposition 3.5: *The comparative statics stated in* Proposition 3.4 *hold whether or not the exchange can endogenously supply the speed.*

The cross effects in Proposition 3.5 are illustrated in Figure 3-7 (d) and (e), which share the same patterns as in Figure 3-3. They show that speculator's speed technology has similar influence from both medium and long-term equilibrium perspectives. On the cross effect of l_F on μ, it should be emphasized that Proposition 3.4 and Proposition 3.5 describe a nonmonotone pattern via the trade-off between the value of speed and the cost of speed acquisition. This trade-off might also be affected by the endogenized speed supply by the exchange through the following spillover mechanism.

Lastly, Figure 3-7 (f), (g), (h) and (i) demonstrate how the advancing in either speculators or exchanges speed technology can spillover to speed demand and supply. We focus on how the advancing in exchange's speed acquisition technology (lowing α_e) feeds back into speculator's speed demand (via μ and l_F) in panels (f) and (g). Intuitively, an advancing in exchange's speed acquisition technology stimulates trading platform's speed. Based on Corollary 3.4, faster exchange speed then directly demotes the fraction and speed of fast speculators. However, the interaction between cross effect further determine the final impacts of the spill-over effect. More explicitly, Corollary 3.5 shows how the advancing in exchange speed technology can spill overs to speculator's investment in fast trading technology, supporting the above intuition.

Corollary 3.5: *From the medium term equilibrium perspective, the effect of exchange latency l_e on the optimal speed and the fraction of fast speculators are given by*

$$\frac{di^*}{dl_e} = \overbrace{M\frac{\partial i(j;\mathcal{E}_m)}{\partial l_e}}^{Direct\ effect} + \overbrace{M\frac{\partial i(j;\mathcal{E}_m)}{\partial j}\frac{\partial j(i;\mathcal{E}_m)}{\partial l_e}}^{Indirect\ effect} \quad (3\text{-}18)$$

for i, j=l_F, μ, where M is a multiplier defined by

$$M = \left(1 - \frac{\partial l_F(\mu;l_e)}{\partial \mu}\frac{\partial \mu(l_F;l_e)}{\partial l_F}\right)^{-1}. \quad (3\text{-}19)$$

Figure 3-8 (b), (c) and (d) illustrate the decomposition of the spill over effect of faster exchange speed on the fraction of fast speculators under compliment scenario based on Corollary 3.5.[63] Figure 3-8 (c) and (d) illustrate the direct and indirect effects, respectively, which can be divided into an initial effect and a multiplier effect in (3-19). There are two consequences: (i) a reduction in the exchange latency crowds out fast speculators initially (indicated by the green arrow); (ii) due to the inter-connection between the two best response functions via the cross effect, the initial effect on μ spills over to l_F that feeds back to μ (indicated by the black arrows). When these cross effects operate in different directions[64], the initial impact is shrunk through the feedback effect (indicated by the change from the longer downward green arrow to the downward red arrow). Consequently, Figure 3-8 (b) shows the direct effect (indicated by the red arrow) and the indirect effect (indicated by the blue arrow) of lowering exchange latency (as in equation (3-18)). Therefore, the indirect effect enhances the direct effect, leading to an unambiguous amplified crowding out of fast speculators, as shown in Figure 3-7 (f).

In general, the direct effect of lowing exchange latency on the optimal speed level and the fraction of fast speculators is negative as indicated in Corollary 3.4; while the indirect effect, coming from the feedbacks through the interaction, can be either positive or negative for the fraction of fast trader but always positive for the optimal speed level. For the fraction of fast speculators, the negatively direct effect dominates the trade-off;

[63] From Corollary 3.3, the term $\partial\mu(l_F;l_e)/\partial l_F$ can be either positive (compliment) or negative (substitute). We leave the analysis for other three cases (the fraction of fast speculator under substitute scenario, agency latency under both substitute and substitute scenarios) in Appendix B3.

[64] An exogenous change in the fraction of fast speculators can spill over to the fast speculators' agent latency when $\partial l_F(\mu;l_e)/\partial\mu \neq 0$ and the direction (either positive or negative) depends on the sign of $\partial l_F(\mu;l_e)/\partial\mu$ (which is negative). The change of the fast speculators' agent latency also feeds back to the fraction of fast speculators (either positively or negatively), depending on the sign of $\partial\mu(l_F;l_e)/\partial l_F$ (which is positive here).

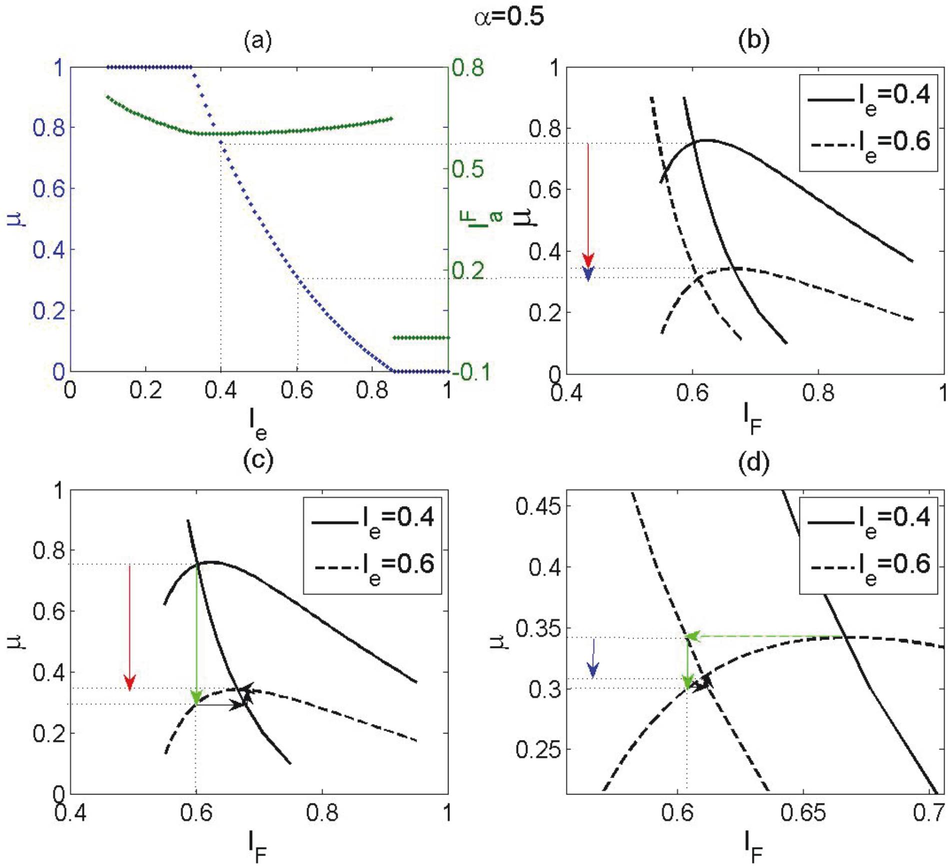


Figure 3-8

The effect of lowering exchange latency on speed investment. This figure illustrates how an increasing in exchange speed endogenously affects speed acquisition decision. Panel (a) depicts that lowering exchange latency shrinks the fraction of fast speculators but stimulates up the optimal speed level. Panel (b) depicts the decomposition of the spill-over effects of lowering exchange latency on the fraction of fast speculators under compliment scenario. Panels (c) and (d) show the decompositions of the direct and indirect effect, respectively. Parameters are the same as in Figure 3-4.

while in contrast, the positively indirect effect dominates for the optimal speed level. Therefore, the speed of speculators becomes more concentrated, i.e., faster exchange speed shrinks the market fraction of fast speculators but stimulates up their optimal trading speed. The results show that speed that was once a ticket to fortunes in financial market is the bare minimum, only the largest trading companies remain in the race for speed. [65] Figure 3-7 (h) and (i) show how the advancing in speculators' speed technology

[65] In a sign of relentless cost pressures, top groups including DRW, IMC, Jump Trading and XR Trading have banded together to construct a speedy wireless-and-cable route between Chicago and Tokyo called Go West (according to a person briefed on the project).

spills over to speed supply by the exchange. We summarize the overall spill-over effects as follows.

Proposition 3.6: *In the interior equilibrium, an advancing in C_F and α promote the exchange's speed; while an advancing in α_e promotes the fast speculator's speed but demotes the fraction of fast traders. Mathematically,* $\partial l_e/\partial C_F < 0$, $l_e/\partial\alpha < 0$ *and* $\partial\mu/\partial\alpha_e > 0$, $\partial l_F/\partial\alpha < 0$.

3.4.2. Implications of Lowering Exchange Latency for Market Quality

We now examine the implications to market quality with respect to price discovery and market liquidity. We use the reciprocal of conditional variances $\epsilon_1 = \left(VAR(V|w_1)\right)^{-1}$ and $\epsilon_2 = \left(VAR(V|w_1, w_2)\right)^{-1}$to measure the price discovery about the fundamental payoff in the early and late trading rounds. Speed advantages are typically measured in microseconds and below and therefore the price discovery in the early trading rounds can be treated as temporary price discovery; while the one in the late trading rounds echoes to permanent price discovery. We focus on the permanent price discovery (in the late trading period).

Following the literature, parameters λ_t $(t = 1{,}2)$ measure market depth or bid-ask spread: a smaller λ_t means that liquidity trading has less price impact in a deeper market. Different from price discovery, the price impacts in both early and late trading rounds matter for market liquidity due to liquidity traders participating in both trading rounds. Considering liquidity for both periods, we use the aggregate expected loss L from liquidity trading in both trading rounds to measure the overall liquidity. Figure 3-9 depicts the overall impact of lowering exchange latency on market quality. Faster exchange speed stimulates less speculators to trade fast, impounding less information. This process benefits liquidity traders in general, shown in Figure 3-9 (a), but hurts the permanent price discovery, illustrated in Figure 3-9 (b).

To better understand the positive (negative) impact of lowing exchange latency on market liquidity (price discovery), we examine how trading intensity affects market quality. Denote $\xi_1 = q_F\mu N\beta_F + q_S(1-\mu)N\beta_S$ and $\xi_2 = (1-q_F)\mu N\beta_F + (1-q_S)(1-\mu)N\beta_S$ the trading intensities in the early and late periods, respectively. Then the (permanent) price discovery is determined by

$$\epsilon_2 = \frac{(\xi_1^2 + \xi_2^2)\sigma^2 + \sigma_z^2}{\sigma^2 \sigma_z^2}$$

while market liquidity is measured by

$$L = -(\lambda_1 + \lambda_2)\sigma_z^2 \ = -\left(\frac{\xi_1 \sigma^2 \sigma_z^2}{\xi_1^2 \sigma^2 + \sigma_z^2} + \frac{\xi_2 \sigma^2 \sigma_z^2}{(\xi_1^2 + \xi_2^2)\sigma^2 + \sigma_z^2}\right).$$

Faster exchange speed increases the trading intensity in short run but decreases the trading intensity in long run, illustrated in Figure 3-9 (c) and (d), respectively. Due to the dominance of the trading intensity in long run, lowing exchange latency demotes the price discovery. On market liquidity, lowing exchange latency reduces short run price impact,[66] meaning that the liquidity traders are better off in early round, but worsen in late round, as in Figure 3-9 (c) and (d) respectively. However, the price impact is much higher in short run than in long run (meaning a large loss for liquidity traders in early trading round). Hence, with lower exchange latency, the decreasing short run price impact contributes more to the improvement of the overall liquidity. Regarding the expected aggregate loss from liquidity trading, the zero-sum trading (due to the risk-neutral Kyle (1985) framework) means that the expected trading profit from all speculators should be the expected trading loss of liquidity traders. Though the trading profits of fast and slow speculators increase in faster exchange speed, the aggregate trading profit decreases due to the crowding-out of fast speculators. This further supports our result about liquidity. Therefore, faster exchange speed has different impacts on market liquidity and price discovery.

[66] More aggressive trading can influence market maker and thus market liquidity through two perspectives. On the one hand, it brings more information to the market maker, which contributes positively to liquidity; on the other hand, it intensifies market maker's adverse selection risk, which contributes negatively to liquidity. In short run, the information effect dominates, leading to better liquidity. In long run, liquidity is jointly determined by short and long run trading intensities. The dominance of the trade-off between the information and adverse selection effects depends on the relatively fraction of fast speculators trading in earlier period, which is further determined by the varied cost.

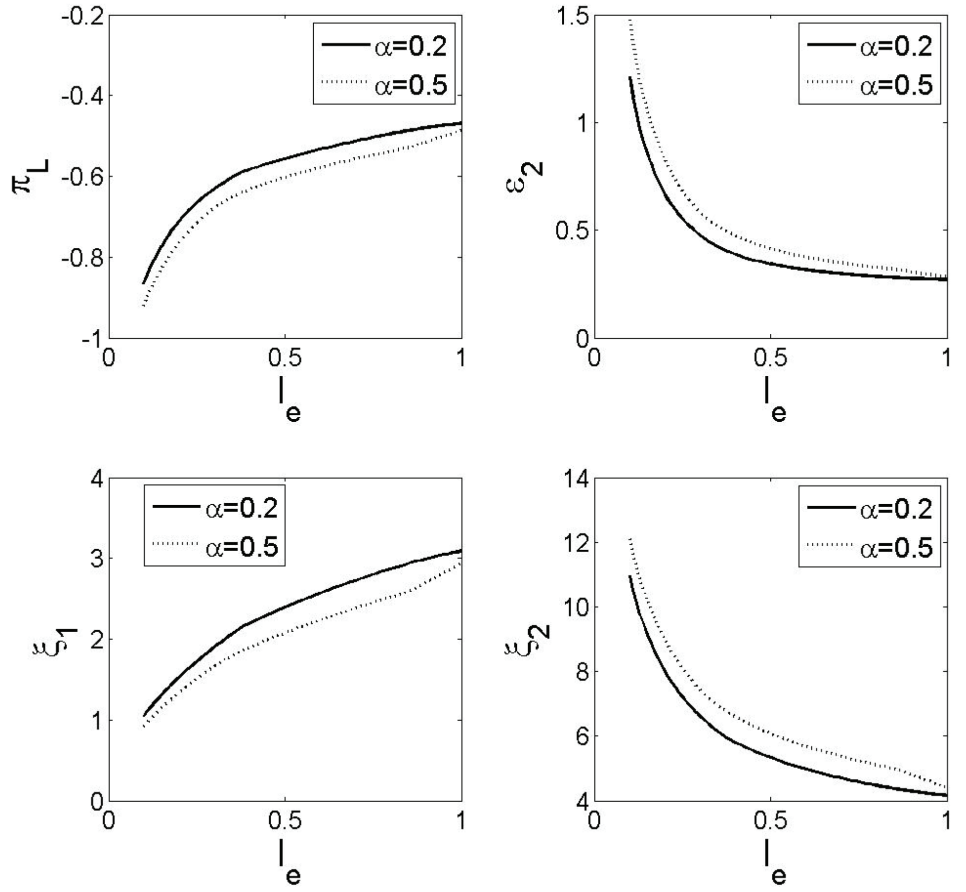


Figure 3-9

Market quality and trading intensity. This figure plots market quality with respect to market liquidity π_L (panel (a)) and price discovery ϵ_2 (panel (b)), together with the trading intensities in short run ξ_1 (panel (c)) and long run ξ_2 (panel (d)), as functions of the exchange latency parameter for two different varied cost $\alpha = 0.2$ (the solid line) and $\alpha = 0.5$ (the dotted line), parameter values are the same as in Figure 3-4**.**

3.5 Exchange Speed Mismatch

The excessive speed investment to maximum profits from both speculators' and exchange's perspective has raised some concerns from policy makers and market regulators about the superior capabilities of fast trading and their impact on market quality and social welfare. When policy makers aim to improve price discovery and to reduce deadweight loss (resulting from speed investment) simultaneously, their desired exchange speed can be different from the optimal speed supplied by exchanges, generating a speed mismatch.

3.5.1. Policy Maker's Social Welfare

Price discovery plays a very important role in economic decisions and investment. When policy makers aim to improve price informativeness in long-run, information acquisition should be encouraged. Also, the aggregate cost on speed investment from speculators and trading platforms is simply a deadweight loss for economy.[67] Therefore, market regulator should also minimize this social waste of speed investment. Considering both perspectives, facing HFT arms race, policy makers may want to regulate exchange speed, l_e, for both price discovery, $\epsilon_2(l_e)$, and social waste, $\mathbb{m}(l_e)N \times \left[C_F + C_V\left(l_F^*(l_e)\right)\right] + C_e(l_e)$.[68] Therefore a social welfare of market regulator can be measured by

$$\mathbb{W}(l_e) \equiv \omega \frac{\epsilon_2(l_e) - \epsilon_2(l_e = 1)}{\epsilon_2(l_e^0) - \epsilon_2(l_e = 1)} + (1-\omega)\frac{-\mathbb{m}(l_e)N \times \left[C_F + C_V\left(l_F^*(l_e)\right)\right] - C_e(l_e) - \underline{\varpi}}{\overline{\varpi} - \underline{\varpi}}, \tag{3-20}$$

in which ω and $1-\omega$ represent the weights for price efficiency, $\epsilon_2(l_e)$, and welfare gain, $-\mathbb{m}(l_e)N \times \left[C_F + C_V\left(l_F^*(l_e)\right)\right] - C_e(l_e)$, that is defined by the saving on market deadweight loss on speed investment, respectively, and parameters $\overline{\varpi}$ and $\underline{\varpi}$ are the maximum and minimum value of the welfare gain. To make market efficiency and social gain comparable, we normalize both components (so that the maximum value of $\mathbb{W}(l_e)$ equal to one).

---

[67] In aggregate, the expected trading revenue from fast and slow speculators, plus exchange profit, should be equal to the expected trading loss of liquidity traders. This zero-sum loss does not contribute to the social welfare of economy.

[68] Liquidity, measured by liquidity traders' expected loss, is not considered here. The gain from the investment in speeds by fast speculators is at the expense of liquidity traders. Traders with faster access to information make profits at the expense of liquidity providers who are slow to update their quotes when new information arrives. Market makers who are fast in obtaining time priority capture a larger fraction of the rents associated with time priority at the expense of slower market makers. The redistribution of the gains from liquidity traders to fast speculators and market makers is however not sufficient to conclude that fast speculators and market makers should be slowed down; because there is no reason that liquidity traders' welfare would be more important than that of informed speculators and market makers.

Thus, a marginal increase in the level of exchange speed has the following effect on the welfare,

$$\frac{\partial \mathbb{W}(l_e)}{\partial l_e} = \underbrace{\frac{\omega}{\epsilon_2(l_e^0) - \epsilon_2(l_e = 1)} \frac{\partial \epsilon_2(l_e)}{\partial l_e}}_{(-)} \tag{3-21}$$
$$- \frac{1-\omega}{\psi - \varpi}\left[\underbrace{\mathbb{u}(l_e) N \frac{\partial C_V\big(l_F^*(l_e)\big)}{\partial l_F} \frac{\partial l_F^*(l_e)}{\partial l_e}}_{(+)} + \underbrace{\frac{\partial C_e(l_e)}{\partial l_e}}_{(+)}\right.$$
$$\left. + \underbrace{\frac{\partial \mathbb{u}(l_e)}{\partial l_e} N\big[C_F + C_V\big(l_F^*(l_e)\big)\big]}_{(-)}\right].$$

The first term is negative, because an increase in exchange speed l_e reduces long term price efficiency, measuring market quality defacement incurred by feedbacks from advancing in exchange' speed supply technology into speculators' speed demand and financial market (see Subsection 3.4.2, the market efficiency decreases in faster exchange speed).

Besides, faster exchange speed also increases varied speed costs for both exchange and speculators; directly, exchange needs to incur higher cost to speed up, indirectly, higher exchange speed stimulates speculators' speed investment. Donating these social costs (in the second line of (21)) by $C_{ar}(l_e)$, we have

$$\frac{\partial \mathbb{W}(l_e)}{\partial l_e} = \left[\underbrace{-\frac{(1-\omega)N\big[C_F + C_V\big(l_F^*(l_e)\big)\big]}{\psi - \varpi} \frac{\partial \mathbb{u}(l_e)}{\partial l_e}}_{Scial\ Value\ of\ Fast\ Exchange(+)}\right] \tag{3-22}$$
$$- \left[\underbrace{-\frac{\omega}{\epsilon_2(l_e^0) - \epsilon_2(l_e = 1)} \frac{\partial \epsilon_2(l_e)}{\partial l_e} - C_{ar}(l_e)}_{Social\ Cost\ of\ Fast\ Exchange(+)}\right].$$

Thus, a marginal increase in l_e has two opposite effects on the social welfare. On the one hand, faster exchange crowds out fraction of speculators investing in speed technology, reducing speculators' speed acquisition costs which are treated as deadweight loss. This benefit captures "*social value of fast exchange*", the second term in equation (3-22). On the other hand, faster exchange and fast speculators pay more varied speed acquisition cost. Together with the reduction of market efficiency with faster exchange, these social costs capture "*social cost of fast exchange*", the first term in equation (3-22).

The socially optimal level of exchange speed, l_e^{so} (i.e., the value of l_e maximizing $\mathbb{W}(l_e)$), depends on the trade-off between the social benefit and social cost of fast exchange. When the social cost dominates the trade-off, policy makers should slow down trading in financial markets. In fact, financial markets and economy have proposed a broad array of policies to address this concern.[69] When the social benefit dominates, policy makers can encourage exchange to provide fast trading in financial markets. Therefore the desired exchange speed for policy makers can be different from the optimal speed supplied by exchanges depends on technology level and policy weight, leading to speed mismatch (i.e., $l_e^{so} \neq l_e^*$ which is discussed in Corollary 3.7).

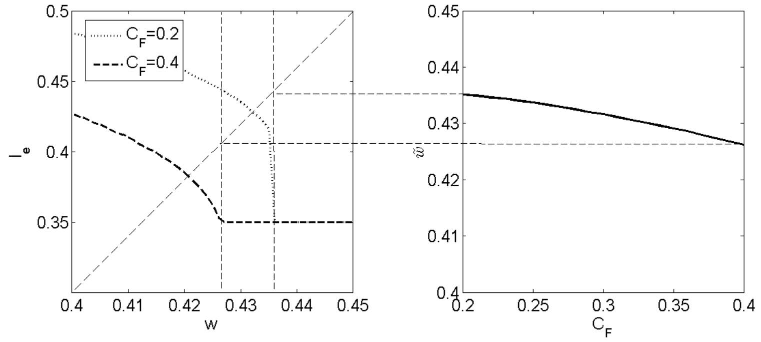
3.5.2. Speed Mismatch to the Default Exchange Speed

The optimal level of exchange speed l_e^{so} is not necessarily at the default exchange speed. In fact, at the default exchange speed $l_e = l_e^0$, the social value of the fast exchange can exceed its social cost and hence policy makers may desire a faster exchange speed, which means that one should not necessarily ban the speed investment of exchanges. Following Corollary 3.6 provides a necessary and sufficient condition for $l_e^{so} > l_e^0$.

Corollary 3.6: *The socially optimal level of exchange speed is strictly larger than the default exchange speed, i.e., $l_e^{so} > l_e^0$, if and only if $\omega < \hat{\omega}(\mathcal{E}_l)$, where $\hat{\omega}(\mathcal{E}_l)$ is a threshold. In addition, this threshold decreases in the fixed cost C_F of speculators' speed acquisition.*

[69] For instance, some trading platforms (e.g., IEX or Alpha) now use so called "speed bumps" to delay the time elapsed between the moment at which an order arrives trading platforms and the moment in which the order enters the matching engine of the trading platforms (see Table C1 for exchanges that have introduced speed bumps).

(a)



(b)

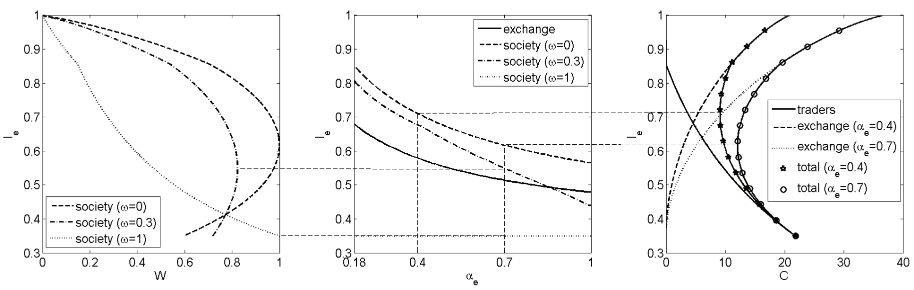


Figure 3-10

Speed mismatch between the desired speed of policy maker and the optimal speed supplied by the exchange. Panel (a) depicts the dependence of the desired exchange speed of policy makers on policy weighting and the fixed cost, c_F; while panel (b) shows how the speed mismatch between the desire exchange speed of policy maker and the optimal speed supplied by the exchange varies with respect to fast trading technology advancement measured by the cost multiplier of the exchange, α_e, under different policy weighting $\omega = 0, 0.3, 1$.

Investment in the fast exchange is socially optimal only if the gain is large enough relative to the social cost of speed technology. The social value of the fast exchange is driven by the efficiency of exchange speed investment, i.e., relatively cheaper cost for exchange speed investment comparing to individual speculator speed investment. From (3-22), at $l_e = l_e^0$, this efficiency gain is given by the social value of fast exchange,

$$\frac{(1-\omega)N\left[C_F + C_V\left(l_F^*(l_e^0)\right)\right]}{\psi - \varpi}\frac{\partial \mathbb{U}(l_e^0)}{\partial l_e},$$

which decreases in ω. Hence, when regulators relatively value the deadweight loss more, $\omega < \widehat{\omega}(\mathcal{E}_l)$, the social value of fast exchange is higher. It is optimal for regulator to allow faster exchange speed investment other than completely ban fast exchange. This is illustrated in the left panel of Figure 3-10 (a). Specifically, the investment in fast exchange is socially optimal if and only if $\omega < \widehat{\omega}(\mathcal{E}_l)$; besides, the socially optimal exchange speed decreases in ω for $\omega < \widehat{\omega}(\mathcal{E}_l)$.

With moderate exchange speed, relatively high fraction of speculators acquire speed and incur speed acquisition cost. Under this situation, higher exchange speed benefits all speculators and significantly decreases speculators' speed investment through crowding out fast trading investment, reducing deadweight loss and social waste. Also, with relatively high fixed speculator speed investment cost and regulators value the deadweight loss more, exchange speed investment could be more efficient. Hence, the threshold value $\widehat{\omega}(\mathcal{E}_l)$ decreases with fixed speculators speed acquisition cost, which is illustrated in the right panel of Figure 3-10 (a). In general, Corollary 6 states that when policy makers relatively care more about deadweight loss, the socially optimal level of exchange speed is always higher than the default speed.

3.5.3. Speed Mismatch to the Optimal Exchange Speed

Previous subsection shows that regulator should not necessarily ban exchanges from investing in speed. Moreover, under some situations, the socially optimal level of exchange speed l_e^{so} might be even higher than the optimal endogenous exchange speed. Essentially, policy makers, who care about price discovery and simultaneously want to limit the deadweight loss (resulting from the speed investment), might have different objective from exchanges that simply maximize operation profits, causing speed mismatch (i.e., $l_e^{so} \neq l_e^*$).

Corollary 3.7: *The socially optimal exchange speed decreases in ω, i.e., $\partial l_e^{so}(\omega)/\partial\omega \leq 0$. In particular, (i) when $\omega = 1$, policy makers focus only on market efficiency and they would desire the lowest exchange speed (the default exchange latency), i.e., $l_e^{so} = l_e^0$; (ii) when $\omega = 0$, policy makers care only the deadweight loss and their desire exchange speed* *is always higher than the equilibrium endogenous speed supplied by the exchange, i.e.,* $l_e^{SO} > l_e^*$*, leading to a speed mismatch.*

With $\omega = 1$, the result is intuitive. Perhaps more interestingly, with $\omega = 0$, there is a trade-off between the decreasing aggregate speed acquisition cost for speculators (i.e., $\mu^*(l_e)N\left[C_F + C_V\left(l_F^*(l_e)\right)\right]$) and increasing cost of speed supply by exchange (i.e., $C_e(l_e)$) as illustrated in the right panel of Figure 3-10 (b).[70] This makes policy makers desire a moderate exchange speed that is higher than the default exchange speed. Moreover, we can show that this socially optimal exchange speed is always higher than the endogenous optimal exchange speed. Indeed, the optimization problem of the exchange in (3-16) can be formally rearranged as

$$l_e^* \in \underset{l_e}{argmax}\, \psi(\lambda_1 + \lambda_2)\sigma_z^2 + (1 - \psi)\mu^*(l_e)N\left[C_F + C_V\left(l_F^*(l_e)\right)\right] \underbrace{-\left\{\mu^*(l_e)N\left[C_F + C_V\left(l_F^*(l_e)\right)\right] + C_e(l_e)\right\}}_{Goal\ of\ policy\ maker}. \tag{3-23}$$

Comparing the objective of the exchange with policy maker's deadweight loss objective,

$$l_e^{SO}(\omega = 0) \in \underset{l_e}{argmax} -\left\{\mu^*(l_e)N\left[C_F + C_V\left(l_F^*(l_e)\right)\right] + C_e(l_e)\right\},$$

the additional term $\psi(\lambda_1 + \lambda_2)\sigma_z^2 + (1 - \psi)\mu^*(l_e)N\left[C_F + C_V\left(l_F^*(l_e)\right)\right]$ measures liquidity traders' loss, as well as fast trader's speed investment cost. As discussed in subsection 3.4.2, slower exchange speed not only impedes market liquidity and hence increases the aggregate loss from liquidity traders but also eases speculator's speed investment cost. Therefore, both $(\lambda_1 + \lambda_2)\sigma_z^2$ and $\mu^*(l_e)N\left[C_F + C_V\left(l_F^*(l_e)\right)\right]$ decrease with exchange speed, implying that policy makers would desire a higher exchange speed in order to reduce the social waste on speed investment. Mathematically, the above discussion implies

---

[70] The aggregate speed cost for fast speculators decreases with lowering exchange latency (though the individual speed cost is increasing with lower exchange latency). Intuitively, lowering exchange latency shrinks the fraction of fast speculators who become faster. For individual speculator, the speed investment is increasing; but the aggregate cost decreases due to the dominance of crowding out of fast speculators, while the exchange speed acquisition cost increases with exchange speed. With the trade-off between the increasing cost for the exchange and the decreasing cost of fast speculators, a lower exchange latency may increase the deadweight loss. Therefore, there exists an optimal speed level that minimizes this deadweight loss.

$$\left.\frac{\partial\{\psi(\lambda_1+\lambda_2)\sigma_z^2+(1-\psi)\mu^*(l_e)N[C_F+C_V(l_F^*(l_e))]\}}{\partial l_e}\right|_{l_e=l_e^{SO}}$$
$$-\underbrace{\frac{\partial\{\mu^*(l_e^{SO})N[C_F+C_V(l_F^*(l_e^{SO}))]+C_e(l_e^{SO})\}}{\partial l_e}}_{policy\ maker\ FOC}$$
$$=\left.\frac{\partial\{\psi(\lambda_1+\lambda_2)\sigma_z^2+(1-\psi)\mu^*(l_e)N[C_F+C_V(l_F^*(l_e))]\}}{\partial l_e}\right|_{l_e=l_e^{SO}}<0.$$

Intuitively, fast exchange increases the cost for the exchange to supply the speed but, due to the significant crowding out effect, reduces the cost for fast speculators that dominates the aggregate cost on speed investment from speculators and exchange. Therefore, policy makers desire a faster exchange speed than the optimal exchange speed, as illustrated in the middle panel of Figure 3-10 (b).

For $0<\omega<1$, Corollary 3.7 shows that the socially optimal exchange speed decreases in the policy weight ω. This result is also illustrated in the left panel of Figure 3-10 (a). This general case can be treated as a weight average of the two extreme cases, $\omega=0$ and $\omega=1$, which is illustrated in the left panel of Figure 3-10 (b). Intuitively, with higher ω, policy makers care more about market efficiency and would desire relatively lower exchange speed. For instance, with $\omega=1$, policy makers desire the default exchange speed, meaning that the exchange over-invests in the speed. By contrast, with $\omega=0$, policy makers desire a moderate speed that is always higher than the endogenous optimal exchange speed, meaning that the exchange under-invests in the speed. However, with $0<\omega<1$, the socially optimal exchange speed decreases from $l_e^{SO}(\omega=0)>l_e^*$ to the default speed l_e^0. Hence, the desire speed from policy makers' perspective can be either higher, lower, or equal to the optimal endogenous exchange speed. These results are illustrated in the middle panel of Figure 3-10 (b).

3.5.4. The Effect of Speed Technology on Speed Mismatch

The speed mismatch can be affected by the advancing speed technology, characterized by the reduction in α_e (while the influences of speculators' speed technology advancement has been analyzed in Appendix C). A deduction in α_e increases the optimal speed for exchanges, l_e^* , due to speed is cheaper for trading platforms. A deduction in α_e

might also increase the socially optimal speed. Intuitively, with advancing in exchange speed acquisition technology, the investment from exchanges becomes more efficient. This is illustrated in the right panel of Figure 3-10 (b) with $\omega = 0$. The deadweight loss is jointly determined by the costs of speculators and exchange. With a lower $\alpha_e = 0.4$ (comparing to $\alpha_e = 0.7$), the speculators' cost is not influenced (indicated by the constant solid line), but the exchange cost is reduced (indicated by the downward shifting dotted line). In aggregate, the deadweight loss (the dashed line) decreases with $\alpha_e = 0.4$. Hence, policy makers desire a higher socially optimal speed l_e^{SO} due to relatively more efficient exchange speed investment, illustrated in the middle panel of Figure 3-10 (b).

Interestingly, with moderate policy weight (i.e., $0 < \omega < 1$), say $\omega = 0.3$, the speed mismatch could first decrease and then increase with technology advancing as showed in the middle panel of Figure 3-10 (b),

$$\frac{l_e^{SO}(\omega = 0.3, \alpha_e) - l_e^*(\alpha_e)}{\partial \alpha_e} < 0.$$

Specifically, with relatively low exchange speed technology (higher value of α_e, say, $\alpha_e = 0.7$), the dashed dotted line ($\omega = 0.3$) is below the optimal endogenous exchange speed l_e^*,, indicating speed overinvestment. By contrast, with relatively advancing exchange speed technology (lower value of α_e say, $\alpha_e = 0.4$), the dashed dotted line ($\omega = 0.3$) is above the optimal endogenous exchange speed l_e^* , indicating speed underinvestment. The mechanism behind is the rapid increasing socially optimal speed with respect to advancing in speed investment technology.[71]

Finally, the speed mismatch could be mitigated by policy makers trying to balance efficiency and social cost at certain speed technology cost level. This is demonstrated by the crossing point between the optimal speed supplied by the exchange and the desired exchange speed from policy makers with $\omega = 0.3$ at $\alpha_e \approx 0.9$ in the middle panel of Figure 3-10 (b). In summary, policy makers, who care about price discovery at long horizons and simultaneously want to limit the deadweight loss (resulting from speed investment), can have different objective from exchange that simply maximizes operation profits. This

[71] The reason why the dashed dotted line ($\omega = 0.3$) increases faster with respect to dashed line ($\omega = 0$) is that the difference between maximum and minimum value of social waste ($\overline{\varpi} - \underline{\varpi}$) is also changed with respect to the advance in exchange speed acquisition cost.

can generate mismatch in the desired speed between the exchange and policy makers, which can be amplified by technology innovation. Therefore, market forces would not lead to the optimal speed level of exchange.

3.6 Concluding Remarks

The speed of trading has considerably increased in recent years, due to technology advancements such as progress in information technologies and automation of the trading process. This evolution raises many questions about the effects of trading speed and has led to several proposals to slowing down markets, calling for regulatory interventions. In this chapter, we incorporate both speed demand from speculators and speed supply by exchanges endogenously to examine the effects, market failures, and policy implications. On the one hand, fast speculators do not internalize the negative externality of their first mover advantage on other market participants, leading to market failure of excessive speed investment of fast speculators. On the other hand, policy makers, who care about price discovery at long horizons and simultaneously want to limit the deadweight loss, can have different objective from the exchanges that simply maximize operation profits. This can generate mismatching between the desired exchange speed for policy makers and the optimal speed supplied by exchanges. Hence, market forces would lead to market failures with respect to excessive speed investment and speed mismatch. The findings echo recent concerns from market regulators about the superior trading capabilities of fast trading.

Chapter 4

Information Diffusion and Speed Competition

There is extensive evidence that decentralized and heterogeneous information diffusion influences investors' trading behaviour.[72] At the same time, trading technology improvements have changed how information is produced and spread and therefore price discovery is affected. On the one hand, trading technology improvements enable some investors to trade in advance on less precise information before news[73] moves stock price closer to its fundamental value. For instance, some investors can speed up execution through technology investment on computer hardware, algorithms, and fast connection to exchange servers (e.g., algorithmic and high-frequency trading technology). On the other hand, some investors can extrapolate information from earlier trades and anticipate future price movements. How does this evolution in speed technologies influence information production and strategic trading among different investors and hence price discovery?

[72] Varian (1989) survey institutional investors in the NYSE and find that the majority attributes their most recent trades to discussions with peers or word of mouth communication. Similar evidence is also found for households (Ivković and Weisbenner, 2007), fund managers' portfolio choices (Hong et al., 2004), and on-line communication between retail foreign exchange traders(Heimer and Simon, 2012).
[73] Either through a diffusion process or via inference from trades.

This question is important because more informative prices enhance the efficiency of capital allocation (see Bond et al. (2012) for an excellent survey).

Intuitively, as more investors rush to trade fast on less precise information, stock prices are less likely to reflect their fundamentals. The literature of time-consuming information processing indeed predicts that, with more fast traders, price discovery is reduced. This is due to either a decline in raw information cost (Dugast and Foucault, 2018) or a rise in time cost of public news and inference from trades (Kendall, 2018). With homogenous information, this literature is limited in exploring the information aggregation function of financial markets. This is particularly relevant when considering that investors can process only a subset of publicly available information (Hong and Stein, 1999) because either they have limited processing capabilities (Jensen and Heckling, 1995, Simon, 1955) or searching over all possible forecasting models (using publicly available information) is costly (Hong et al., 2007). With heterogeneous information, this chapter shows that fast trading can aggregate heterogeneous and imprecise information from early equilibrium prices and orders and render slow traders more informative. Thus, a rush of fast trading due to low information dissemination can *improve* price discovery.

In our model, an information diffusion process is described by two characteristics: *information disclosure* level, which is measured by the precision of the initial information signals, and *information dissemination* rate at which heterogeneous information signals about fundamental payoff disperse across investors. Specifically, we decompose fundamental value information into n independent sub-innovations with the same precision. In the early period, each investor randomly receives one of the sub-innovations, which are then dispersed to other investors in the late period. Information diffusion brings more precise signals, but it takes time. To account for this, we assume that each trader observes $(\delta - 1)$ new innovations in the late period. Therefore, an increase in $\theta = 1/n$ characterizes higher information disclosure, while an increase in δ reflects faster information dissemination.

Before trading, there are two technologies available for investors: information and speed technologies. Information technology provides investors with full information in addition to the partial information received from the diffusion process, while speed technology allows investors to trade in the early period (instead of waiting for the late period). There are three types of rational traders: partially informed fast traders who trade in the early period, partially informed slow traders who trade in the late period, and informed traders who are fully informed about the fundamental information and trade in the late period. In equilibrium, investors optimally choose to become either fast or informed before trading. This provides a framework to analyze how changes in information diffusion affect equilibrium speed acquisition and information production, which then affect price discovery in the early and late periods.

This chapter delivers three main results. First, the "aggregation effect" of fast trading, a driving feature of the model, can significantly affect strategic trading in the late period. In general, the first-mover advantage makes traders rush to trade earlier on less precise information. However, the equilibrium prices aggregate heterogeneous signals, releasing more information to slow traders. This generates strategic trading between early and late trades. More precisely, with the first mover advantage, there is no intra-temporal competition between fast and informed (or slow) traders. However, more aggressive fast trading makes the early equilibrium price more informative, reducing information asymmetry for slower traders and adverse selection for (long-lived) market makers. This asymmetry reduction effect generates *strategic complementarity* of late to early trading so that increasing fast trading intensity strengthens the trading intensity of informed (or slow) traders.

Second, the value of speed (i.e., to trade earlier) decreases in information disclosure but can be ambiguous in information dissemination. The rise of technology investment on computer hardware, algorithms, and fast connection to exchange servers has witnessed the development of high frequency trading (HFT). At the same time, information is disseminated faster through social media such as tweets, newswires, and company reports, etc.[74] Reducing the cost of speed technologies contributes to fast trading intuitively; our results however suggest information dissemination as another channel for increasing fast trading. This is because an increase in information dissemination, on the one hand, increases slow traders' revenue due to more available information, an "*information effect*". On the other hand, it intensifies intra-temporal competition since the same information is observed by more groups of traders, a *"competition effect"*. Considering that information dissemination does not affect fast traders, when slow traders' revenue declines due to competition effect dominates trade-off, value of speed and hence fast trading is spurred.

Lastly, and perhaps surprisingly, price discovery in both periods is improved in disclosure but reduced in dissemination; although both information disclosure and dissemination crowd out fast and informed traders. An increase in information disclosure intensifies the intra-temporal competition among fast trader and hence decreases trader's willingness to become fast. However, this intensive competition, coupled with more informative signals from disclosure, results in more informative fast trading. Channelled by trading complementarity of late to early trading, this also leads to more aggressive trading in the late period, improving price discovery in both periods. Instead, an increase in information dissemination directly reduces the value of information, which further reduces the value of speed due to the complementarity of speed acquisition to information production. Thus, fast trading is less likely be informative through the crowding out of fast traders. Due to the learning of slow traders, these effects can ultimately reduce price discovery in both periods. This chapter shows that the distinction in different information diffusion characteristics, namely disclosure versus dissemination, is important in

[74] In the past, investors learned the news through reading newspapers, watching television, or by word-of-mouth. The advent of social media has fundamentally changed how information is produced and disseminated in financial markets. Take Twitter as an example; people receive short electronic messages known as "tweets" through the Twitter accounts they follow, and they can re-transmit the message by retweeting. This does not affect initial information that traders observe, but incredibly influences transmission speed. Robert Shiller (2015) also mentions that "We were witnessing another explosion of technological innovations that facilitate interpersonal communication, consisting of e-mail and chat rooms, and after 2000 these led to social media… These new and effective media for interactive (if not face-to-face) communication may have the effect of expanding yet again the interpersonal contagion of ideas. They may have allowed enthusiasm for the market to spread much more widely than it would otherwise have. Certainly, we are still learning how to regulate the use of these new media in the public interest".

determining price discovery and can have different policy implications for financial markets.

Related literature

With rapid growth of the high frequency trading arms race, the ubiquity of computerized algorithmic and data abundance, speed hierarchy and information diffusion play very important roles in financial markets. The speed hierarchy not only motivates fast trading with low precision signals but also renders slower trades more informative. This speed characteristic is shared by the literature featuring sequential trading rounds. As in Froot et al. (1992), Hirshleifer et al. (1994), and Banerjee et al. (2018), our model features "early" and "late" investors. Our model differs from this literature in that the numbers of early and late investors are endogenous in equilibrium. Dugast and Foucault (2018), Holden and Subrahmanyam (1996) and Kendall (2018)also examine the endogenous decision to be either early or late (short- versus long-horizon; raw versus processed; quick and imprecise versus slow and precise signals). Distinct from this literature, we model speed acquisition and information production separately.

Recently, by separating speed from information, Huang and Yueshen (2018) make a novel contribution to the literature by identifying endogenous complementarity or substitution between information and speed technologies. We also consider that investors optimally choose to become either fast or informed separately. However, distinct from Huang and Yueshen (2018), our focus is on how information diffusion, characterized by information disclosure and dissemination, contributes to price discovery with a speed hierarchy.

Building on the literature of endogenous information production, the current modelling framework extends to endogenous trading speed acquisition in financial markets. In the literature, the static endogenous information equilibrium model of Grossman and Stiglitz (1980) has been extended from different perspectives by considering: a) dynamic trading (Avdis, 2016, Mendelson and Tunca, 2004); b) traders conditioning their information acquisition decision on public signals (Foster and Viswanathan, 1993); and c) traders pre-committing to receiving signals at particular dates

(Back and Pedersen, 1998, Holden and Subrahmanyam, 2002). A distinguishing feature of our model from this literature is that (partially informed) speculators can endogenously acquire trading speed to capture trading speed competition.

This chapter also contributes to a broad literature that studies the influence of HFT (see Menkveld (2016) for an excellent survey on both theoretical and empirical research regarding HFT). The growing literature focuses on three aspects of these technological changes; (i) what information high-frequency traders use to trade (Hu et al., 2016, Van Kervel and Menkveld, 2019); (ii) how high-frequency trading incorporates information (Brogaard et al., 2014, O'hara et al., 2014); and (iii) what impact of high-frequency trading on market quality (Budish et al., 2015, Weller, 2016, Yang and Zhu, 2020). By introducing information diffusion (with respect to (i)), this chapter addresses questions (ii) and (iii). In particular, we highlight that fast trading provides information for slower trades; however, increases in information disclosure and dissemination can have opposite effects on asset price informativeness, providing important policy implications for market regulators regarding HFT. Our results help to better understand how the evolution of information environment explored in the literature can change financial markets under this "new environment" of abundant information and fast trading.

This chapter further contributes to the literature on information disclosure and transparency (surveyed in Bond et al. (2012), Goldstein and Yang (2017), Verrecchia (2001)). The literature has theoretically explored the implications of information transparency from different aspects, including the cost of capital (Hughes et al., 2007); informative prices about fundamentals (Banerjee et al., 2018) ; endogenous liquidity trading (Han et al., 2016); inefficient coordination on public information (Morris and Shin, 2002); and learning of managers or regulators (Goldstein and Yang, 2019). However, none of these studies have examined the effect of information disclosure and dissemination on market quality when investors can endogenously acquire trading speed. This chapter also relates to the recent literature studying the effect of information diffusion on trading and asset price (see Section 4.1 for more detail).

This chapter is organized as follows. Section 4.1 develops the model of information diffusion and speed competition. Section 4.2 characterizes the equilibrium trading and price in terms of trading intensities and value of speed (information). In the endogenous equilibrium, Section 4.3 examines how information diffusion affects information production, speed acquisition, and price discovery. Section 4.4 talks about the implication for price and trade patterns. Section 4.5 concludes.

4.1 Model

How information is generated by market participants, disseminated, and finally incorporated into prices has been the subject of extensive literature in financial economics. With the increase in fast trading competition, this remains one of the key questions for understanding how financial markets operate. In this section, we first introduce two characteristics of an information diffusion process: *information disclosure level*, measured by the initial information precision, and the *information dissemination rate* at which heterogeneous information is dispersed among investors. We then characterize a perfect Bayesian equilibrium to examine how the changes in information environment affect information production, speed acquisition, and price discovery.

4.1.1. Information Diffusion Process

Consider a financial market with N investors trading a risky asset with normally distributed payoffs $V \sim N(0, \sigma^2)$. To characterize information diffusion and trading speed competition among investors, we introduce a heterogeneous information diffusion process over two periods, indexed by $t = 0, 1, 2$. The fundamental value V can be decomposed into n independent sub-innovations with the same variance:

$$V = \sum_{i=1}^{n} v_i, \tag{4-1}$$

where $v_i \sim N(0, \sigma^2/n)$ $(i = 1, 2, \ldots, n)$ are independently and identically distributed (*iid*) random variables. All investors are grouped evenly into n groups. Prior to trading at $t = 1$, the investors in group i receive v_i. This reflects that investors can process only a subset of public information (Hong and Stein, 1999)because either they have limited processing capabilities (Jensen and Heckling, 1995, Simon, 1955)or searching over all possible

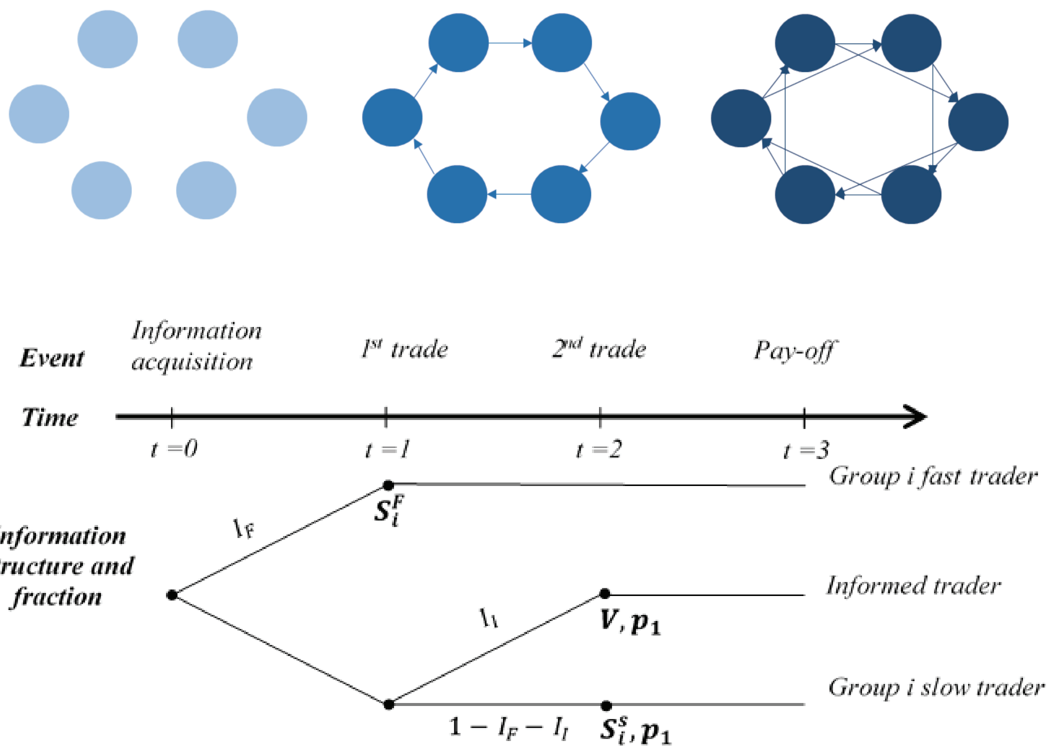


Figure 4-1

Information diffusion process and time line. The upper panels describe the information diffusion process. All investors are grouped evenly into *n (=6)* groups (the upper left panel). At time $t = 1$, each trader in group i observes v_i. After time $t = 1$, information starts to diffuse across different groups. At time $t = 2$, each trader in group i observes δ innovations $\{v_i, v_{i+1}, \dots, v_{i+\delta}\}$, and each innovation v_i is observed by a fraction of δ/n traders (the upper middle panel with $\delta = 2$ and the upper right panel with $\delta = 3$). The bottom panel illustrates the trading flows of fast traders in the early period and informed and slow traders in the late period.

forecasting models using publicly available information itself is costly (Hong et al., 2007). Each trader in group i observes the same innovation and each innovation is observed by a fraction of $1/n$ traders. This information structure can be treated as an islands-connection network in the social-network literature (Han and Yang, 2013, Jackson, 2005). Within each group, all traders are fully connected in the sense that they share the same signals due to friendships, club memberships, or social media, but different groups of traders are not connected with each other.

We use parameter $\theta = 1/n$ to represent information disclosure level; higher θ means traders have more precise heterogeneous public information initially and hence better information disclosure. After time $t = 1$, the information starts to diffuse across different groups. To measure the speed of information diffusion, we introduce parameter δ to measure heterogeneous information dissemination. At time $t = 2$, each trader in group i observes $\delta - 1$ new innovations so that his information set becomes $\{v_i, v_{i+1}, \dots, v_{i+\delta-1}\}$, and each innovation is observed by a fraction of δ/n traders. Therefore, with larger δ, information spreads more widely among traders, reflecting faster information dissemination.

This symmetric rotation feature of the information diffusion is helpful for the following reasons. Firstly, the full information is equally split among the n groups of traders. This provides a consistent information allocation scheme for comparing results when information diffuses (measured by disclosure and dissemination) at different speed. With faster information dissemination, the investors who either trade in the first (fast) or second (slow) period, as a group, have the full information about the fundamental value. Secondly, more importantly, it ensures that on average partially informed traders are equally informed even though the information is heterogeneous and spreads across the population. This symmetry simplifies the "*forecast the forecasts of others*" problem and helps to obtain a closed-form equilibrium. This information diffusion process is illustrated in the upper panel of Figure 4-1.

This information diffusion structure is closely related but also different from the current literature. Firstly, by considering sequential information arrival, Kendall (2018) and Dugast and Foucault (2018) examine the time cost of homogeneous information so that the precision of traders' information increases with time, meaning that information filters out the noise in data but it takes time. The information diffusion shares a similar feature wherein an investor in group i receiving only $\{v_i\}$ at time $t = 1$ observes δ innovations $\{v_i, v_{i+1}, \dots, v_{i+\delta-1}\}$ at time $t = 2$. However, we focus on the time dimension of heterogeneous, instead of homogenous, information.

Secondly, in a dynamic model of opinions, prices, and volume, Hong et al. (2011) consider word-of-mouth communication to describe information diffusion as random evolution of informed traders [75] Similarly, Manela (2014) models homogeneous information diffusion as a fraction of informed investors that evolves deterministically over time[76]. Different from Manela (2014), we focus on heterogeneous information diffusion among partially informed investors; a sub-innovation v_i revealed to a fraction $1/n$ of the population at time $t = 1$ diffuses deterministically so that the same signal is revealed to a fraction δ/n of the population at time $t = 2$. In our framework, the incidence probability $\Gamma(I_1, \delta)$ can be represented as:

$$\Gamma(I_1, \delta) = \frac{\delta - 1}{n - 1}.$$

Lastly and more broadly, this model is related to the literature on information percolation and network theory[77]. Instead of exploring the consequences of complicated technique issues, this chapter embeds a simpler information diffusion technology in a centralized financial market to better understand the joint effect of increasing information disclosure requirements from market regulators, faster information dissemination from the new information environment, and the development of fast trading technology. This sheds some light on the debate about the effect of increasing information transparency and fast trading competition, in particular in the context of HFT.

[75] Hong et al. (2011) allow for intricate information diffusion dynamics at the cost of a simplified asset pricing framework in which agents agree to disagree and behave myopically. In their framework, the fraction of informed traders at each time is a random variable, whose distribution can be calculated by the transition probability from the communication process.

[76] Specifically, a signal is revealed initially to a fraction I_1 of the population. Information then diffuses deterministically from I_1 in this period to I_2 in the next period by $I_2 - I_1 = \Gamma(I_1, \delta)(1 - I_1)$. Here $\Gamma(I_1, \delta)$ is the incidence probability that an uninformed agent in the previous period becomes informed in the next period, depending on the fraction informed that allows for word-of-mouth transmission of information. The incidence probability is increasing in the transmission rate ($\partial\Gamma/\partial\delta > 0$) so that, all else equal, uninformed traders are more likely to receive more information when the transmission rate is higher.

[77] Duffie and Manso (2007) and subsequently Duffie et al. (2009) study information percolation through agents' encounters in large but decentralized markets. In an exogenous information environment, Ozsoylev and Walden (2011) show that social communication improves price informativeness. Walden (2019) constructs a model in which information diffuses through a general network of agents. Agents who are more closely connected have more similar period-by-period trades, and their profitability is closely related to the Katz centrality measure.

4.1.2. Asset Market

Based on the information diffusion structure, the total of N rational traders are grouped into three types[78]: (i) a fraction I_F ($> \frac{n}{N}$) of *fast traders* who pay a speed cost of C_F (> 0) to trade earlier at time $t = 1$; (ii) a fraction I_I ($> \frac{n}{N}$) of *informed traders* who pay an information cost of C_I (> 0) to receive extra (full) information and trade at time $t = 2$; and (iii) a fraction $1 - I_F - I_I$ ($> \frac{n}{N}$) of *slow traders* who receive information from the diffusion process and trade at time $t = 2$. There are *n* groups of fast and slow traders, respectively. Accordingly, fast traders in group i receive information $s_i^F = \{v_i\}$at time $t = 1$, while slow traders in group i receive information $s_i^S = \{v_i, v_{i+1}, \dots, v_{i+\delta-1}\}$ at time $t = 2$. Therefore, fast traders are not only faster-acting but also partially informed about less precise information[79].

In addition, as usual, over the two periods, there is a liquidity trader who trades randomly $z_t{\sim}N(0, \sigma_z^2)$ shares and a competitive market-maker who absorbs the net trading flow and sets market prices p_t to have zero expected profit. At time $t = 0$, all traders independently invest in either information (to be informed) or speed technologies (to be fast) but not both; and the remaining speculators are slow traders. As in Kyle (1985), all traders are risk neutral and the fundamental value is revealed at the end of the second

[78] As in Fishman and Hagerty (1992), we allow the number of speculators to be a continuous variable (to avoid integer issues). The restriction on the fraction is due to the strategic setting (see the Appendix for more detail). To disentangle the effect of trading speed from information diffusion, we assume that fast traders only trade in the first period and no speculators trade in both periods. We relax this assumption in Section 4.4 to allow fast traders to trade in both periods and show that the main results remain unchanged.

[79] Empirically, Jovanovic and Menkveld (2016)find that high-frequency market makers can use index futures information to update their quotes; and Hu et al. (2016) show that high-frequency traders are quick in processing public information. Our assumption is consistent with these empirical studies and the argument that "*the investment decision to become HFT is often modeled as a 'bundle' – once invested, the investor is both informed and fast*" (Huang and Yueshen, 2018). Most observations associate HFT with extremely fast computer algorithms coded by traders who trade for their own account, meaning that fast traders are able to act rapidly. What information HFT utilizes is still hotly debated. Empirical evidence shows that high frequency traders are informed, but without being adversely selected completely on their quotes (Brogaard et al., 2014, Menkveld, 2016). We assume that fast traders receive partial information through information diffusion. By studying the trading around local price trends, Fishe et al. (2019) find that those who best predict either the start or the end of such trends are typically not the fastest traders. These results tend to support our assumption that fast traders are partially informed about the fundamental information.

period. The timeline of the information structure, trading actions, and events is illustrated in the lower panel of Figure 4-1.

With the information structure and three types of traders described above, we now introduce their trading decisions. For the fast traders in group i, they receive (partial) information s_i^F and submit market orders $x_F(s_i^F)$ at time 1. There are $I_F N/n$ fast traders in each group and hence the aggregate trading volume at time $t = 1$ is given by:

$$w_1 = \frac{I_F N}{n} \sum_{\mathrm{i}=1}^{n} x_F(s_i^F) + z_1, \tag{4-2}$$

where z_1 is the order submitted by the liquidity trader. Conditional on price p_1 at time $t = 1$, informed traders submit market orders $x_I(V, p_1)$ while slow traders in group i submit market orders $x_S(s_i^s, p_1)$ at time $t = 2$. Hence, the aggregate trading volume at time $t = 2$ becomes:

$$w_2 = I_I N x_I(V, p_1) + \frac{(1 - I_F - I_I)N}{n} \sum_{\mathrm{i}=1}^{n} x_S(s_i^s, p_1) + z_2, \tag{4-3}$$

where z_2 is the order submitted by the liquidity trader at time $t = 2$.

4.1.3. Equilibrium

As in Kyle (1985), we now introduce the perfect Bayesian equilibrium. We first consider exogenous equilibrium for given fractions of fast and informed traders and then endogenous equilibrium in which information production and trading speed acquisition are endogenized in Section 4.3.

Definition 4.1: *Given the fractions of the fast and informed traders I_F and I_I, respectively, a perfect Bayesian equilibrium is defined by traders' strategy profile and market maker's pricing rules* $\{x_F(s_i^F), x_I(V, p_1), x_S(s_i^s, p_1), p_1(w_1), p_2(p_1, w_2)\}$ *that satisfy the following conditions:*

1) profit maximization at period 1 and 2:

$$x_F^*(s_i^F) = \underset{x_F}{argmax}\, E[x_F(V - p_1)|s_i^F];$$

$$x_I^*(V, p_1) = \underset{x_I}{argmax}\, E[x_I(V - p_2)|V, p_1]; \qquad x_S^*(s_i^s, p_1) = \underset{x_S}{argmax}\, E[x_S(V - p_2)|s_i^s, p_1].$$

2) *market efficiency: the price functions* $p_1(w_1)$ *and* $p_2(p_1, w_2)$ *are determined by the rules that the market maker clears the security market in each period for an expected zero profit;*

$$E[w_2(p_2 - V)|p_1, w_2] = 0; \qquad E[w_1(p_1 - V) + w_2(p_2 - V)|w_1] = 0.$$

3) *all the traders have rational expectations in that each trader's belief about the others' strategies is correct in the equilibrium.*

Definition 4.1 implies that traders and the market maker are strategic in their trading and pricing, characterizing "*forecast the forecasts of others*" (Foster and Viswanathan, 1996). In other words, a trader's trading strategy depends on the trading strategies of other traders and the market maker's pricing rules, which in turn also affect a trader's trading strategies. This creates an endogenous feedback of "*one agent's strategy affects other agents' strategies that affect himself own strategy*" (Foster and Viswanathan, 1996). Following the literature, we consider a linear equilibrium in which the price linearly depends on the aggregate trading volume:

$$p_1 = \lambda_1 w_1, \qquad p_2 = p_1 + \lambda_2 w_2, \tag{4-4}$$

where coefficients λ_1 and λ_2 are endogenously determined in equilibrium. Based on the linear price assumption, as in Kyle (1985), the trading strategies of fast and informed traders are also linear:

$$x_F(s_i^F) = \beta_F s_i^F; \tag{4-5}$$

$$x_I(V, p_1) = \beta_I(V - p_1). \tag{4-6}$$

Equation (4-5) captures how the fast traders trade in advance on their speed advantage and thus face a better quote, while equation (4-6) indicates how aggressively the informed traders trade on their information advantage about the pricing error in the expected fundamental value (based on their information about V) from the equilibrium price in period-one. The coefficients β_F and β_I measure how sensible the fast and informed traders trade to their information. The trading strategy for the slow traders becomes more complicated:

$$x_S(s_i^s, p_1) = \beta_S(s_i^s - p_1) + \gamma_S p_1. \tag{4-7}$$

While the first term has a similar meaning as in the informed traders' trading strategy, the second term captures the learning of the slow traders from the equilibrium price in the first period[80].

4.2 Equilibrium Trading Strategies and Prices

To better understand endogenous speed acquisition and information production and their impact on price discovery, this section focuses on the exogenous equilibrium in Section 4.1. The procedure of obtaining the equilibrium follows the standard Kyle (1985) model. Based on the market maker's linear pricing functions, we first solve the speculators' optimal investment problem. We then characterize the fixed-point problem of strategic and optimal trading among the traders. Finally, we determine the market maker's pricing functions and obtain the following linear equilibrium.

Proposition 4.1: *For given $I_F > 0$ and $I_I > 0$, there exists a linear Bayesian equilibrium specified by equations (4-2)-(4-7), in which the coefficients $(\lambda_1, \lambda_2, \beta_F, \beta_I, \beta_S, \gamma_S)$ are the functions of parameters $(N, \theta, \delta, I_I, I_F, \sigma, \sigma_z)$ given by (E-1) to (E-4) in Appendix E1.*

Based on the equilibrium trading strategies and prices, we examine first the trading intensities of fast, informed, and slow traders and then the value of information and speed. We show how price discovery depends on the trading intensities, and how information disclosure and dissemination and changes in the fraction of fast and informed traders affect trading intensities and therefore price discovery.

[80] In the equilibrium to be characterized, strategy (4-7) can also be written alternatively as $x_S(s_i^s, p_1) = \alpha(E(V) - p_1)$ for some constant $\alpha > 0$. This implies that the order of the slow traders is proportional to their information advantage relative to the market maker. With the linear price, $E(V)$ is also linear in slow traders' private information and equilibrium price. We nonetheless work with (7) since it is the most general form and does not impose any structure.

4.2.1. Trading Intensities and Price Discovery

Definition 4.2: *The trading intensities Φ_F, Φ_I and Φ_S, measuring how aggressively fast, informed, and slow traders, respectively, trade on their information about the fundamental, are defined by:*

$$\Phi_F = \frac{I_F N}{n}\beta_F; \qquad \Phi_I = I_I N \beta_I; \qquad \Phi_S = \delta\frac{(1 - I_I - I_F)N}{n}\beta_S. \tag{4-8}$$

We use the reciprocal of conditional variances $\xi_1 = \left(VAR(V|w_1)\right)^{-1}$ and $\xi_2 = \left(VAR(V|w_1, w_2)\right)^{-1}$ to measure the price discovery about the fundamental payoff in the first and second periods, respectively. They indicate how much residual uncertainty the market maker faces after conditioning on the prices. From Definition 2 and Proposition 1, we have the following result.

Corollary 4.1: *In the exogenous equilibrium, price discovery about the fundamental payoff in the first and second periods is given, respectively, by:*

$$\xi_1 = \frac{\Phi_F^2\sigma^2 + \sigma_z^2}{\sigma^2\sigma_z^2}; \qquad \xi_2 = \frac{(\Phi_F^2 + (\Phi_I + \Phi_S)^2)\sigma^2 + \sigma_z^2}{\sigma^2\sigma_z^2}. \tag{4-9}$$

Corollary 4.1 shows that price discovery is positively affected by the trading intensities of traders. Intuitively, when traders trade more aggressively, they reveal more information to the market maker who then impounds more information into the equilibrium prices. Clearly $\xi_1 < \xi_2$, meaning that price discovery in the late period is always better than in the early period. Also, price discovery is affected by fast trading only in the early period, but jointly affected by fast trading and slow trading in the late period. For convenience, we use $\Phi_L = \Phi_I + \Phi_S$ to represent the aggregate trading intensity in the late period. We refer to ξ_1 and ξ_2, respectively, as *short run* and *long run* price informativeness in the early and late periods[81]. Therefore, trading intensities and hence price discovery depend on strategic trading among fast, informed, and slow traders;

[81] Note that the notion of short and long run is relative to the moment at which information starts to diffuse. Clearly, the long run price informativeness is always higher than the short run price informativeness because the market-maker has more information in the late period.

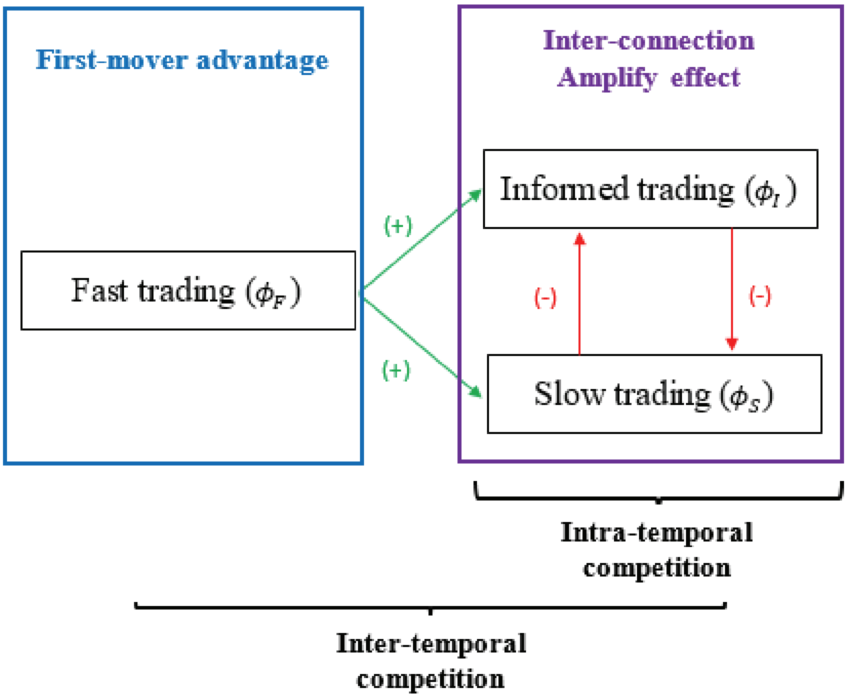


Figure 4-2
Strategic trading among fast, slow and informed traders. Due to the first-mover advantage, fast trading is not influenced by informed and slow trading. However, informed and slow trading intensities not only depend on each other, due to the intra-temporal competition, but also depend on fast trading intensity, due to the inter-temporal competition. In general, informed trading is a substitute to slow trading but a complement to fast trading; while slow trading is a substitute to informed trading but a complement to fast trading.

changes in the fractions of fast and informed trading; and changes in information disclosure and dissemination to be explored in the following.

Strategic Trading Intensity Interactions: Fast trading aggregates heterogeneous and less precise signals and renders slower traders more informative. This *aggregation effect* of fast trading is a driving force for the strategic trading among traders. More explicitly, the strategic interactions among the three types of traders can be characterized by the interaction of their trading intensities.

As for slow trading's impact on informed trading intensity, we show (in Appendix E2) that it can be decomposed into direct and indirect effects of fast and slow trading intensities as follows:

$$\Phi_I = \underbrace{\frac{I_I N}{2(I_I N+1)} \frac{4\Phi_F^2\sigma^2 + 4\sigma_z^2 + \left(\Phi_S\sigma + \sqrt{(\Phi_S^2 + 4I_I N\Phi_F^2)\sigma^2 + 4I_I N\sigma_z^2}\right)^2}{\Phi_S\sigma^2 + \sqrt{(\Phi_S^2 + 4I_I N\Phi_F^2)\sigma^4 + 4I_I N\sigma^2\sigma_z^2}}}_{\text{direct effect from observing signal}} - \underbrace{\frac{I_I N}{I_I N+1}\Phi_S}_{\text{indirect effect from inferring slow traders}} = \Phi_{I,Direct}(\Phi_F, \Phi_S) - \Phi_{I,Indirect}(\Phi_F, \Phi_S). \tag{4-10}$$

This indicates that slow trading strengthens informed trading directly (i.e., $\partial\Phi_{I,Direct}(\Phi_F, \Phi_S)/\partial\Phi_S > 0$). This is because more aggressive slow trading impounds more information and reduces the market maker's adverse selection and price impact. This makes the informed traders trade more aggressively, generating an "*asymmetry reduction effect*". Secondly, slow trading also strengthens the indirect effect (i.e., $\partial\Phi_{I,Indirect}(\Phi_F, \Phi_S)/\partial\Phi_S > 0$), which negatively affects informed trading intensity Φ_I. This captures a "*competition effect*" between the informed and slow traders. Therefore, given the fundamental value, more aggressive slow trading reduces the pricing error and thus the revenue of informed traders, making the informed trade less aggressively. We show (in Appendix E2) that the competition effect dominates the asymmetry reduction effect. This implies that informed trading is a substitute to slow trading, meaning that informed traders trade less aggressively when slow traders trade more aggressively.

As for fast trading's impact on informed trading intensity, due to the first-mover advantage, fast traders trade earlier than informed and slow traders. Hence, fast trading intensity has no influence on informed trading indirectly (no competition), that is $\partial\Phi_{I,Inirect}(\Phi_F, \Phi_S)/\partial\Phi_F = 0$. However, the early period equilibrium price provides information for the market maker and weakens the information asymmetry. Consequently, fast trading leads to a similar asymmetry reduction effect directly, that is $\partial\Phi_{I,Direct}(\Phi_F, \Phi_S)/\partial\Phi_F > 0$. Therefore, informed trading is a complement to fast trading, meaning that informed traders trade more aggressively when fast traders trade more aggressively.

The effect of fast and informed trading on the slow trading intensity can be analyzed similarly. We can show that slow traders trade less aggressively with more informed traders, but more aggressively with informed traders. In general, the first-mover advantage makes traders rush to acquire speed with less precise information. This releases more information to slow traders and affects their trading behaviour. In summary, we have the following results on strategic trading behaviours.

Proposition 4.2: *In the exogenous equilibrium,*

(i) informed trading intensity is a substitute to slow trading intensity, but a complement to fast trading intensity; that is $\frac{\partial \Phi_I}{\partial \Phi_S} < 0$ *and* $\frac{\partial \Phi_I}{\partial \Phi_F} > 0$*;*

(ii) slow trading intensity is a substitute to informed trading intensity, but a complement to fast trading intensity; that is $\frac{\partial \Phi_S}{\partial \Phi_I} < 0$ *and* $\frac{\partial \Phi_S}{\partial \Phi_F} > 0$.

Proposition 4.2 is illustrated in Figure 4-2. It shows that, due to the first-mover advantage, fast trading is not influenced by informed and slow trading. However, informed and slow trading intensities not only depend on each other, due to the intra-temporal competition,[82] but also depend on the fast trading intensity, due to the inter-temporal competition.

Fractions of Fast and Informed Traders: We next examine the effect of changes in the fraction of informed and fast traders[83] on trading intensities and hence price discovery.

Corollary 4.2: *In the exogenous equilibrium,*

(i) an increase in the fraction of fast traders makes the fast and informed traders trade more aggressively but the slow traders trade less aggressively (i.e., $\frac{\partial \Phi_F}{\partial I_F} > 0, \frac{\partial \Phi_I}{\partial I_F} > 0$ *and* $\frac{\partial \Phi_S}{\partial I_F} < 0$*); it also promotes late period trading intensity (i.e.,* $\frac{\partial \Phi_L}{\partial I_F} > 0$*) when* $I_I > I_I^a$ *(defined in (A-18)); but demotes late period trading intensity otherwise.*

(ii) an increase in the fraction of informed traders promotes the trading intensities of informed traders (i.e., $\frac{\partial \Phi_I}{\partial I_I} > 0$*) and trading in the late period (i.e.,* $\frac{\partial \Phi_L}{\partial I_I} > 0$*),*

[82] Figure 4-2 also shows an amplification effect due to the interconnection between informed and slow trading.

[83] When both the fractions of informed and fast traders are endogenously determined, let Q be either disclosure level or dissemination rate. The effect of Q on the late period trading intensity can then be represented as $\frac{d\Phi_L}{dQ} = \frac{\partial \Phi_L}{\partial I_F}\frac{\partial I_F}{\partial Q} + \frac{\partial \Phi_L}{\partial I_I}\frac{\partial I_I}{\partial Q} + \frac{\partial \Phi_L}{\partial Q}$.

demotes the trading intensity of slow traders (i.e., $\frac{\partial \Phi_S}{\partial I_I} < 0$*), but has no impact on trading intensity of fast traders (i.e.,* $\frac{\partial \Phi_F}{\partial I_I} = 0$*).*

Corollary 4.2 (i) shows that, with more fast traders, informed traders trade more aggressively while slow traders trade less aggressively. Then, the late period trading intensity is ambiguous regarding the level of fast traders, depending on the trade-off. Intuitively, when the fraction of informed traders is relatively high (satisfying $I_I > I_I^a$), an increase in informed trading intensity becomes more significant, dominating the trade-off and improving the trading intensity in the long run. According to Corollary 4.2 (ii), with more informed traders, a similar trade-off applies; however, the late period trading intensity deterministically increases for informed traders. Although informed trading intensity Φ_I increases in either more informed or fast traders, the mechanisms behind are different. More fast traders stimulate more informed trading due to the indirect complementary effect; while more informed traders directly increase their trading intensity, becoming more significant. Therefore, the dominance of significant increases in informed trading on the trade-off always stimulates trading in the late period. With today's rapid growth in HFT, Corollary 4.2 suggests that the influence of fast trading on price discovery depends on informed trading in the market. When informed institutions dominate the market, fast trading might help to improve price discovery in both the short and long run.

Information Disclosure and Dissemination: We finally examine the effect of the change of information disclosure and dissemination on trading intensities. Corollary 4.3 shows how information disclosure and dissemination affect trading differently.

Corollary 4.3: *In equilibrium,*

(i) an increase in information disclosure demotes informed trading (i.e., $\frac{\partial \Phi_I}{\partial \theta} < 0$*) when* $I_F < I_F^b$ *and* $I_I < I_I^b$ *(defined in (A-22a) and (A-22b)); otherwise, it promotes informed trading. Nevertheless, it always promotes fast trading (i.e.,* $\frac{\partial \Phi_F}{\partial \theta} > 0$*), slow trading (i.e.,* $\frac{\partial \Phi_S}{\partial \theta} > 0$*) and late period trading (i.e.,* $\frac{\partial \Phi_L}{\partial \theta} > 0$*).*

(ii) an increase in information dissemination demotes informed trading (i.e., $\frac{\partial \Phi_I}{\partial \delta} < 0$*) but promotes slow trading (i.e.,* $\frac{\partial \Phi_S}{\partial \delta} > 0$*) and trading in the late period (i.e.,* $\frac{\partial \Phi_L}{\partial \delta} > 0$*); however, it has no impacts on fast trading (i.e.,* $\frac{\partial \Phi_F}{\partial \delta} = 0$*).*

For informed disclosure, according to Lemma E1 in Appendix E3, the influence on informed trading intensity depends on a direct effect of fast and slow trading and a trading intensity multiplier:

$$\frac{d\Phi_I}{d\theta} = \left(1 - \frac{\partial \Phi_I}{\partial \Phi_S}\frac{\partial \Phi_S}{\partial \Phi_I}\right)^{-1} \left(\underbrace{\left(\frac{\partial \Phi_I}{\partial \Phi_F} + \frac{\partial \Phi_I}{\partial \Phi_S}\frac{\partial \Phi_S}{\partial \Phi_F}\right)\frac{\partial \Phi_F}{\partial \theta}}_{\text{direct effect through fast trader}(+)} + \underbrace{\frac{\partial \Phi_I}{\partial \Phi_S}\frac{\partial \Phi_S}{\partial \theta}}_{\text{direct effect through slow informed}(-)} \right).$$

The direct effect can be decomposed into a positive component through fast trading and a negative component through slow trading. When the fractions of informed and fast traders are low (satisfying $I_F < I_F^b$ and $I_I < I_I^b$), the fraction of slow traders is relatively high. Then the negative component through slow trading becomes more significant, dominating the trade-off and leading to less aggressive informed trading. However, with increasing disclosure, slow traders deterministically trade more aggressively, dominating the aggregate trading intensity in the late period.

For information dissemination, similarly, the influence on informed trading intensity also depends on the direct effect of fast and slow trading and the trading intensity multiplier. It has no impact on fast trading but contributes negatively to the direct effect through slow trading. Therefore, faster dissemination always demotes informed trading. However, with increasing dissemination, slow traders trade more aggressively, dominating the trade-off and leading to more aggressive trading in the late period.

Corollary 4.4: *The price discovery in both periods always improves with information disclosure. With increasing information dissemination, the price discovery is not affected in the early period but improves in the late period.*

Corollary 4.4 implies that, in the exogenous equilibrium, both information disclosure and dissemination can promote trading in the late period and hence improve long run price discovery (through different mechanisms). Recent regulatory efforts (the Sarbanes-Oxley and Dodd-Frank Acts) have brought the debate on information disclosure and dissemination in financial markets to the forefront of market regulation. Our results so far seem to support these regulations. However, we show (in Section 3.3) that this is no longer true in the endogenous equilibrium. Specifically, due to endogenous speed acquisition and information production, price discovery is improved with information disclosure but reduced with dissemination in both the short and long run.

4.2.2. Value of Information and Speed

To develop endogenous equilibrium, we now examine the value of information and speed. At time $t = 1$, a fast trader investing in speed technology has expected profit (per capita) $\pi_F(s_i^F) = x_F(s_i^F)(s_i^F - E[p_1|s_i^F])$ conditional on his signal s_i^F. Similarly, at time $t = 2$, the expected trading profits of informed investing in information technology and slow traders are $\pi_I(V, p_1) = x_I(V, p_1)(V - E[p_2|V, p_1])$ and $\pi_S(s_i^s, p_1) = x_S(s_i^s, p_1)(s_i^s - E[p_2|s_i^s, p_1])$, conditional on their information sets $\{V, p_1\}$ and $\{s_i^s, p_1\}$, respectively. Denote the ex-ante expected profits of speculators by $\bar{\pi}_i \overset{\text{def}}{=} E[\pi_i]$ for $i = F, I, S$. By computing the difference between the unconditional expected revenues of being fast and informed compared to being slow, we obtain the value of speed and information, respectively.

Lemma 4.1: *The value of the speed Γ_F is given by*

$$\Gamma_F(I_F, I_I, \theta, \delta) = \bar{\pi}_F - \bar{\pi}_S = \frac{\sqrt{n}}{\sqrt{I_F N}(I_F N + n)} \sigma \sigma_z - \frac{\delta n}{I_F N + n} \Psi(I_F, I_I, \theta, \delta) \quad , \tag{4-11}$$

while the value of information Γ_I is given by

$$\Gamma_I(I_F, I_I, \theta, \delta) = \bar{\pi}_I - \bar{\pi}_S = (n - \delta)\Psi(I_F, I_I, \theta, \delta) \quad , \tag{4-12}$$

where $\Psi(I_F, I_I, \theta, \delta)$ is defined in (E-24).

<u>Fractions of Fast and Informed Traders</u>: Lemma 4.1 shows that the value of information (speed) decreases with the fraction of investors investing in information (speed) technology. This is consistent with the standard Grossman-Stiglitz effect wherein investing in the same technology (either information or speed) is always a strategic substitute; having more informed (fast) traders reduces a trader's incentive to become informed (fast). Furthermore, we show that speed acquisition is a complement to

information production, while information production can be either a substitute or complement to speed acquisition.

From (4-11), we have $\partial \Gamma_F(I_F, I_I, \theta, \delta)/\partial I_I > 0$, which means that acquiring speed is a complement to producing information. Intuitively, though informed traders have no influence on fast traders' revenue, more informed traders reduce slow traders' revenue due to intensified intra-temporal competition in the late period. Consequently, the value of speed increases.

Perhaps surprisingly, information production can also be a complement to speed acquisition. This is, for some parameter values, the value of information is higher when the market maker learns more from aggregate order flow due to a higher fraction of fast traders. This means that investors trading on the processed information can in fact benefit from a more informative price during the early period. Specifically, from (4-12) we have:

$$\frac{\partial \Gamma_I(I_F, I_I, \theta, \delta)}{\partial I_F} = \underbrace{\frac{(n-\delta)I_F N + n^2}{I_F N + n}\Psi \frac{\partial}{\partial I_F}\left[\frac{(n-\delta)(I_F N + n)}{(n-\delta)I_F N + n^2}\right]}_{competition\ reduction\ effect\ (+)} + \underbrace{\frac{(n-\delta)(I_F N + n)}{(n-\delta)I_F N + n^2}\frac{\partial}{\partial I_F}\left[\frac{(n-\delta)I_F N + n^2}{I_F N + n}\Psi\right]}_{asymmetry\ reduction\ effect\ (-)}. \tag{4-13}$$

Equation (4-13) indicates that more fast traders generate two opposite effects. First, more fast traders make the equilibrium price in the early period more informative. This makes slow traders rely relatively more on the price instead of their own signals and reduces slow traders' trading sensitivity, generating a positive "*competition reduction effect*". Second, a more informative early equilibrium price renders slow traders more informative and hence reduces the information rent for the informed, generating a negative "*asymmetry reduction effect*" for the informed. The overall impact depends on the trade-off between the competition reduction effect and asymmetry reduction effect. Intuitively, more fast traders release more information through early period price and reduce the revenues for both informed and slow traders. This intensifies the competition among informed traders, but also eases the competition between slow and informed traders. When the fraction of informed traders is relatively high $I_I > I_I^c$, informed traders are affected more significantly by their own competition compared to competition with slow traders, which reduces the information value. The above analyzis leads to the following result.

Proposition 4.3: *In equilibrium,*

(i) acquiring speed is a complement to producing information, meaning that the value of speed increases for the informed traders (i.e., $\frac{\partial r_F}{\partial I_I} < 0$*);*

(ii) producing information can be either a substitute or a complement to acquiring speed; more explicitly, the information value decreases (i.e., $\frac{\partial r_I}{\partial I_F} < 0$*) for* $I_I > I_I^c$ *(defined in (D-27)) but increases otherwise for fast traders.*

<u>*Information Disclosure and Dissemination*</u>: We now examine how information disclosure and dissemination influence the value of information and speed. With faster dissemination, it is intuitive that the value of information decreases since slow traders become more informative. However, the influence on the value of speed is more complicated. It affects slow traders' unconditional expected profit, though not fast traders' unconditional expected profit. Specifically, from (4-11):

$$\frac{\partial \bar{\pi}_S(I_F, I_I, \theta, \delta)}{\partial \delta} = \underbrace{\Psi(I_F, I_I, \theta, \delta)\frac{\partial}{\partial \delta}\left[\frac{\delta n}{I_F N + n}\right]}_{information\ effect} + \underbrace{\frac{\delta n}{I_F N + n}\frac{\partial}{\partial \delta}[\Psi(I_F, I_I, \theta, \delta)]}_{competition\ effect}, \tag{4-14}$$

where $\Psi(I_F, I_I, \theta, \delta)$ is defined in (E-24). Based on (4-14), faster dissemination in general brings more information to slow traders through the diffusion process but generates two opposite effects. It increases slow traders' revenue due to greater information availability, generating an "*information effect*". It also implies that the same information is observed by more groups of traders; this intensifies intra-temporal competition, leading to a "*competition effect*". The total effect depends on their trade-off. Intuitively, with fewer informed traders, the competition among the slow traders becomes more significant compared to competition with informed traders. Therefore, the competition effect dominates the information effect, reducing the slow traders' unconditional expected profit and increasing the value of speed.

Similarly, higher information disclosure renders slow traders more informative and hence reduces the information value. Due to the strategic complementary of speed acquisition to information production, the value of speed is also reduced. We therefore have the following result.

Corollary 4.5: *In equilibrium,*

(i) faster information dissemination always decreases information value (i.e., $\frac{\partial r_I}{\partial \delta} < 0$*), and increases speed value (i.e.,* $\frac{\partial r_F}{\partial \delta} > 0$*) when* $I_F < I_F^d$ *and* $I_I < I_I^d$ *(defined in (D-30a)-(D-30b)) but decreases the speed value otherwise.*

(ii) higher information disclosure level always decreases the value of both information (i.e., $\frac{\partial r_I}{\partial \theta} < 0$*) and speed (i.e.,* $\frac{\partial r_F}{\partial \theta} < 0$*).*

The rise of technology investment on computer hardware, algorithms, and fast connection to exchange servers has witnessed the development of HFT and speeded up information acquisition and dissemination. At the same time, we see faster information dissemination through online data such as tweets, newswires, company reports, etc. Corollary 4.5 implies that, in a market with less informed traders, faster information dissemination can be a driving force for increasing speed competition. Otherwise, information disclosure and dissemination crowd out both fast and informed traders.

4.3 Endogenous Equilibrium of Information and Speed

This section examines how information diffusion (via information disclosure and dissemination) affects price discovery differently in the endogenous equilibrium when information production and speed acquisition are endogenized, compared to the exogenous equilibrium in Section 4.2.

4.3.1. Endogenous Equilibrium

Regarding information production and speed acquisition decisions, each trader can pay a cost, C_F or C_I, to become a fast or informed trader, respectively. Traders who choose neither to acquire speed nor to produce information are slow traders. A trader chooses to become fast or informed only when his expected payoff from the trading is sufficient to cover the cost of acquiring the speed or information, compared to the slow trader. In equilibrium, by taking into account the costs, a trader is indifferent between being fast by acquiring speed (or being informed by producing information) and being slow.

Definition 4.3: *For given speed cost C_F and information cost C_I, market equilibrium fractions of fast and informed traders (I_F^*, I_I^*) are determined by $\Gamma_i = C_i$ for $I_i^* > 0$ or $\Gamma_i \leq C_i$ for $I_i^* = 0$, where $i = F, I$.*

The cost of information (speed) measures the ease of producing information (acquiring speed). A proliferation of sources of information about the firm (say, abundant disclosure, large analyzt/media coverage, and advanced communication technologies) makes information easier to access, reducing information cost (Fishman and Hagerty, 1992, Kim and Verrecchia, 1994). Similarly, a significant improvement in trading technology (say, investments in computational hardware and fast connection to exchange servers) reduces the costs to trade fast. In interior equilibrium, $\Gamma_F(I_F, I_I, n) = C_F$ and $\Gamma_I(I_F, I_I, n) = C_I$. The following corollary shows how the equilibrium fractions of fast and informed traders are affected by the costs.

Corollary 4.6: *When both informed and fast traders are active in market equilibrium,*

(i) with an increase in the cost of being fast, the fraction of fast traders decreases (i.e., $\frac{\partial I_F}{\partial C_F} < 0$) while the fraction of informed traders increases (i.e., $\frac{\partial I_I}{\partial C_F} > 0$) when $I_I > I_I^c$ but decreases when $I_I < I_I^c$;

(ii) with an increase in the cost of being informed, the fractions of the informed and fast traders decrease (i.e., $\frac{\partial I_I}{\partial C_I} < 0$ and $\frac{\partial I_F}{\partial C_I} < 0$).

Corollary 4.6 confirms the standard substitution effect in information production and speed acquisition with respect to their respective costs. More importantly, Corollary 4.6 shows that reducing the cost of being informed increases traders' willingness to become fast, while reducing the cost of speed can increase the willingness of being informed in a market with less informed traders (but decrease such willingness otherwise). This is consistent with the literature. For example, Dugast and Foucault (2018) show that a decrease in the cost of the raw signal (being fast in our model) can increase the equilibrium demand for the processed signal (being informed in our model) when the reliability of the raw signal is relatively low. By bundling information and speed together, they do not examine the effect of cross-technologies in information and speed. In addressing this effect, Huang and Yueshen (2018) find that as one technology increases, given the other, the aggregate speed and information acquisition of investors are initially complements but eventually substitutes. Similar to Huang and Yueshen (2018), we separate speed acquisition and information production and find the non-monotonically cross-technology effect. Different from Huang and Yueshen (2018), in our model, such nonlinear cross-technology effect only exists for the speed technology, but not the information technology. In other words, a trader's willingness to become informed increases initially but reduces eventually with decreasing cost of speed. In contrast, reducing the cost of being informed always increases traders' willingness to become fast.

4.3.2. Information Disclosure and Dissemination

We now examine how information disclosure and dissemination affect endogenous speed acquisition and information production. We focus on analyzis for an interior solution (and provide the analyzis for a corner solution in Appendix E9). From the equilibrium condition, $\Gamma_F(I_F, I_I, \theta, \delta) = C_F$, the optimal response function (on the fraction of fast traders) I_F^* for a given value of I_I^* (and θ, δ) can be described as $I_F^* = \Upsilon_F(I_I^*, \theta, \delta)$. Similarly, given the fraction I_F^*, the optimal response function for the fraction of informed traders can be described as $I_I^* = \Upsilon_I(I_F^*, \theta, \delta)$. Both response functions jointly determine the endogenous equilibrium of information production and speed acquisition, I_F^* and I_I^*, by a fixed-point problem.

<u>*Speed Acquisition:*</u> We first consider how information diffusion affects the equilibrium fraction of fast trading. We substitute the equilibrium condition for an informed trader to the equilibrium condition for a fast trader $I_F^* = \Upsilon_F(\Upsilon_I(I_F^*, \theta, \delta), \theta, \delta)$. Taking a total differentiation with respect to $Q = \{\theta, \delta\}$ yields:

$$\frac{dI_F^*}{dQ} = \frac{\overbrace{\frac{\partial \Upsilon_F(I_I^*,\theta,\delta)}{\partial Q}}^{Direct\ effect} + \overbrace{\frac{\partial \Upsilon_F(I_I^*,\theta,\delta)}{\partial I_I}\frac{\partial \Upsilon_I(I_F^*,\theta,\delta)}{\partial Q}}^{Indirect\ effect\ (-)}}{1 - \frac{\partial \Upsilon_F(I_I^*,\theta,\delta)}{\partial I_I}\frac{\partial \Upsilon_I(I_F^*,\theta,\delta)}{\partial I_F}} < 0. \tag{4-15}$$

In (4-15), the terms $\frac{\partial \Upsilon_F(I_I^*,\theta,\delta)}{\partial Q}$ and $\frac{\partial \Upsilon_F(I_I^*,\theta,\delta)}{\partial I_I}\frac{\partial \Upsilon_I(I_F^*,\theta,\delta)}{\partial Q}$ capture the direct and indirect effects, respectively, of information diffusion. We know from Proposition 4.3 that acquiring speed is a complement to producing information (i.e., $\frac{\partial \Upsilon_F(I_I^*,\theta,\delta)}{\partial I_I} > 0$). We also know from Corollary 4.5 that both faster dissemination and higher disclosure reduce the information advantage of informed traders and hence their equilibrium fraction (i.e., $\frac{\partial \Upsilon_I(I_F^*,\theta,\delta)}{\partial Q} < 0$). Therefore, information diffusion tends to demote speed acquisition through the indirect effect (i.e., $\frac{\partial \Upsilon_F(I_I^*,\theta,\delta)}{\partial I_I}\frac{\partial \Upsilon_I(I_F^*,\theta,\delta)}{\partial Q} < 0$).

Furthermore, from Corollary 4.5, the direct effect of information disclosure ($\frac{\partial \Upsilon_F(I_I^*,\theta,\delta)}{\partial \theta}$) is negative. Thus, higher disclosure crowds out fast traders. Instead, from Corollary 5, the direct effect of dissemination ($\frac{\partial \Upsilon_F(I_I^*,\theta,\delta)}{\partial \delta}$) can be either positive or negative, depending on the fractions of both informed and slow traders. Therefore, the overall impact depends on the trade-off between the direct and indirect effects. Nevertheless, we show (in Appendix E10) that the indirect effect always dominates the trade-off; thus, both faster dissemination and higher disclosure crowd out fast traders.

Information Production: We next examine the effect of information diffusion on information production in equilibrium. Similarly, we have $I_I^* = \Upsilon_I(\Upsilon_F(I_I^*,\theta,\delta),\theta,\delta)$. Taking a total differentiation with respect to $Q = \theta, \delta$ yields:

$$\frac{dI_I^*}{dQ} = \overbrace{\frac{\partial \Upsilon_I(I_F^*,\theta,\delta)}{\partial Q}}^{Direct\ effect} + \overbrace{\frac{\partial \Upsilon_I(I_F^*,\theta,\delta)}{\partial I_F^*}\frac{dI_F^*}{dQ}}^{Indirect\ effect}. \tag{4-16}$$

Corollary 4.5 shows that the direct effect is negative ($\frac{\partial \Upsilon_I(I_F^*,\theta,\delta)}{\partial Q} < 0$); thus information diffusion decreases directly informed traders' information rent and investors' willingness to become informed. The above analyzis on (4-15) shows that information diffusion crowds out fast traders ($\frac{\partial I_F^*}{\partial Q} < 0$). Therefore, the total effect depends on the strategic trading of informed and fast traders, characterized in Proposition 3. When producing information is a complement to acquiring speed ($\frac{\partial \Upsilon_I(I_F^*,\theta,\delta)}{\partial I_F^*} > 0$), information diffusion demotes information acquisition indirectly ($\frac{\partial Y_I(I_F^*,\theta,\delta)}{\partial I_F^*}\frac{dI_F^*}{dQ} < 0$); therefore both faster dissemination and higher disclosure crowd out informed traders.

However, when producing information is a substitute to acquiring speed (i.e., $\frac{\partial Y_I(I_F^*,\theta,\delta)}{\partial I_F^*} < 0$), information diffusion tends to promote information acquisition indirectly ($\frac{\partial Y_I(I_F^*,\theta,\delta)}{\partial I_F^*}\frac{dI_F^*}{dQ} > 0$). Then the overall effect depends on the trade-off between the direct and indirect effects. Proposition 4.3 suggests that only when the fraction of informed traders is sufficiently high will information production be a substitute to speed acquisition. However, with more informed traders, the direct effect also becomes more significant. Thus, the direct effect dominates the indirect effect and therefore the equilibrium fraction of informed traders decreases, crowding out informed traders.

Proposition 4.4: *In the endogenous equilibrium, faster dissemination and higher disclosure reduce both information production ($\frac{dI_I^*}{d\theta} < 0$, $\frac{dI_I^*}{d\delta} < 0$) and speed acquisition ($\frac{dI_F^*}{d\theta} < 0$, $\frac{dI_F^*}{d\delta} < 0$).*

Proposition 4.4 shows that both faster dissemination and higher disclosure crowd out fast traders; however, their mechanisms are different. Higher disclosure intensifies the intra-temporal competition among faster traders, reducing the value of speed. However, faster dissemination reduces informed traders and affects the inter-temporal competition. Following the discussion on (4-15), due to the dominance of the negatively indirect effect from informed trading, faster dissemination reduces the value of speed. In the following we show that these different mechanisms affect price discovery differently.

4.3.3. Price Discovery

We know from Lemma 4.1 that price discovery depends on traders' trading intensities. We first examine price discovery in the short run, which solely depends on fast trading intensity,

$$\Phi_F = \sqrt{\theta I_F N}\frac{\sigma_z}{\sigma} \tag{4-17}$$

based on Proposition 4.1. Following Proposition 4.4, faster dissemination crowds out fast traders. From (4-17), this means that fast trading and hence price discovery are reduced in the short run.[84] Instead, higher disclosure brings more information to fast traders and (we show in Appendix E11) in general promotes fast trading due to the dominance of the information effect, improving price discovery.[85] Therefore faster dissemination impedes while higher disclosure improves price discovery in the short run.

The long run price discovery depends on the trading intensities of all traders, in particular, trading intensity Φ_L in the late period. Let Q be either δ or θ, then we have:

$$\frac{d\Phi_L}{dQ} = \underbrace{\frac{\partial\Phi_L}{\partial I_F}\frac{\partial I_F}{\partial Q}}_{Indirect\ /\ fast} + \underbrace{\frac{\partial\Phi_L}{\partial I_I}\frac{\partial I_I}{\partial Q}}_{indirect\ /\ informed} + \underbrace{\frac{\partial\Phi_L}{\partial Q}}_{direct}, \qquad (Q=\theta,\delta). \tag{4-18}$$

From the terms $\frac{\partial\Phi_L}{\partial I_F}$ and $\frac{\partial\Phi_L}{\partial I_I}$ in Corollary 4.2, $\frac{\partial\Phi_L}{\partial Q}$ in Corollary 4.4, $\frac{\partial I_F}{\partial Q}$ and $\frac{\partial I_I}{\partial Q}$ in Proposition 4.4, the direct effect is positive, while the indirect effect is negative via informed trading but ambiguous via fast trading. The complexity of the overall equilibrium precludes a fully analytical characterization of these outcomes, so we illustrate the results numerically in Figure 4-3.

The upper panels in Figure 4-3 show opposite impacts of information dissemination and disclosure on the long run price discovery; faster dissemination impedes while higher disclosure improves price discovery (in both periods). The results suggest that improvements in trading technologies and changes in information environments jointly influence how information is produced and disseminated in financial markets. This observation provides a new insight to the findings of Bai et al. (2016) who find that stock price informativeness about short horizon earnings has decreased over time, while, at least for stocks in the S&P 500 index, price informativeness about long horizon earnings has increased. Our results indicate that this observation may be the result of the

[84] Faster dissemination reduces information rent and hence intra-temporal competition during the later period directly. It also crowds out faster traders, reduces the speed value indirectly, and hence facilitates fast trading.

[85] Higher disclosure brings more information to faster traders and intensifies the intra-temporal competition between faster traders, crowding out fast traders. It also makes fast traders impound more information into their trading volume and trade more aggressively, dominating the crowding-out effect and hence improving price discovery.

Table 4-1
Scenario specifications

	Speed technology	Observing early equilibrium price	Trading intensity of informed and slow traders	Long run price discovery
FM	Yes	Yes	Increase	Increase
NL	Yes	No	Decrease	Increase
NS	No	No	Decrease	Decrease

dominance of increasing information dissemination in the short run and information disclosure in the long run.

Though both faster dissemination and higher disclosure crowd out fast traders, the underlying mechanisms are different. Therefore, their impacts on price discovery are also different, with differences observed in both long run and short run price discovery. These differences depend not only on the population of informed and fast traders, but more importantly on: a) how aggressively fast traders, as a group, trade on their information; b) how much of fundamental information "leaks" to slow traders through fast trading; and c) how informed trading reacts to changes in fast trading.

To further explore the channel of strategic complementary of late to early trading, we consider two scenarios by switching off several model elements in the full model (FM). Specifically, we first consider a no learning (NL) scenario where informed and slow traders do not observe early period equilibrium price. We then consider a no speed (NS) scenario where investors can only trade in the late period. Details of these two scenarios are analyzed in Appendices E9 and E12, while their connections to the full model (FM) are summarized in Table 4-1.

Firstly, comparing FM with NL shows the effect of learning from early period price. Not observing the early price impedes not only strategic complementarities in informed and fast trading but also slow traders' learning. Therefore, with higher information disclosure, aggregate trading in the late period becomes less aggressive, as illustrated in the bottom left panel in Figure 4-3. However, the long run price discovery still increases due to the significant improvements in short run price discovery. Secondly, comparing NL with NS shows the effect of speed technology. Similarly, without additional

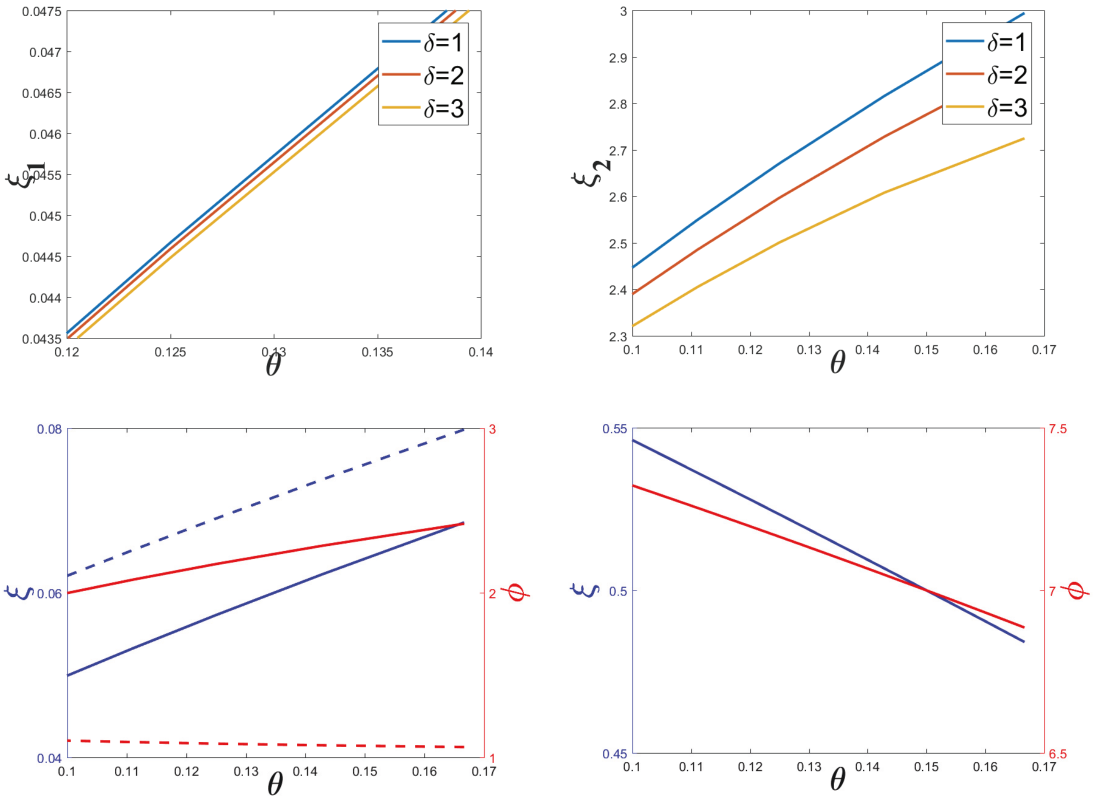


Figure 4-3

The effect of information diffusion on price discovery. The upper panels show equilibrium short (left panel) and long (right panel) run price discovery with respect to information disclosure for three information dissemination rates ($\delta = 1, 2, 3$), while the bottom panels show equilibrium price discovery (blue line) and trading intensity (red line) with respect to the information disclosure for without learning from early price (left panel; $C_I = 0.01, C_F = 1, \delta = 2$) and without speed (right panel; $C_I = 0.2$); in the left bottom panel, the solid and dashed line represents the first and second period, respectively; otherwise, the number of speculators $N = 100$, volatility of the fundamental value $\sigma = 10$, volatility of the order flow of the liquidity traders $\sigma_z = 10$, information production cost $C_I = 0.1$, and speed acquisition cost $C_F = 1.5$.

information from the early period equilibrium price, the aggregate trading of informed and slow traders decreases, illustrated in the bottom right panel in Figure 4-3. Furthermore, under this situation, higher disclosure not only reduces the aggregate trading of informed and slow traders but also impedes price discovery. This analysis shows how endogenous speed acquisition and information production re-shape the trade-off in the exogenous equilibrium and leads to an opposite effect of information disclosure compared to dissemination on price discovery.

4.4 Implications for Price and Trade Patterns

This section provides implications and empirical predictions of information diffusion for trading patterns based on traders' equilibrium trading behaviour. We first analyze trade volume induced by traders' equilibrium trading behaviour and provide empirical predictions of information disclosure and dissemination regarding the trading behaviour of different market participants in the short and long run.

Trading volume: Trading volume provides an important indicator of the trading activity of market participants. The total trading volume of different types of traders can be measured by the variance of their aggregate order flow:

$$\eta_F = VAR\left(\frac{I_F N}{n}\sum_{i=1}^{n} x_F(v_i)\right) = \left(\frac{I_F N}{n}\beta_F\right)^2 \sigma^2; \tag{4-19}$$

$$\eta_I = VAR\big(I_I N x_I(V, p_1)\big) = (I_I N \beta_I)^2\left(\left(1-\lambda_1\frac{I_F N}{n}\beta_F\right)^2 \sigma^2 + \lambda_1^2\sigma_z^2\right); \tag{4-20}$$

$$\begin{aligned}\eta_s &= VAR\left(\frac{(1-I_F-I_I)N}{n}\sum_{i=1}^{n} x_S(v_i, p_1)\right)\\ &= \left(\frac{(1-I_F-I_I)N}{n}\right)^2 [(\beta_S\delta + (\gamma_S-\beta_S)\lambda_1 I_F N\beta_F)^2\sigma^2\\ &\quad + (n(\gamma_S-\beta_S)\lambda_1)^2\sigma_z^2].\end{aligned} \tag{4-21}$$

The above are the total trading volumes for fast, informed, and slow traders, respectively.

Figure 4-4 plots the trading volume patterns with respect to information disclosure and dissemination. The top left panel shows that the trading volume monotonically increases in information disclosure for fast traders but decreases for informed traders. Intuitively, an increase in information disclosure makes it easier for fast traders to access information and improve their signal precision. Therefore, they trade more aggressively. This also renders slow traders more informative (based on the equilibrium trading volume in the early period) and therefore reduces the information advantage and hence the trading volume for the informed traders. The top right panel shows the proportional trading volume of fast and informed traders, relative to the total trading volume. Consistent with the top left panel, the proportional trading volume increases for fast traders but decreases for informed traders in information disclosure.

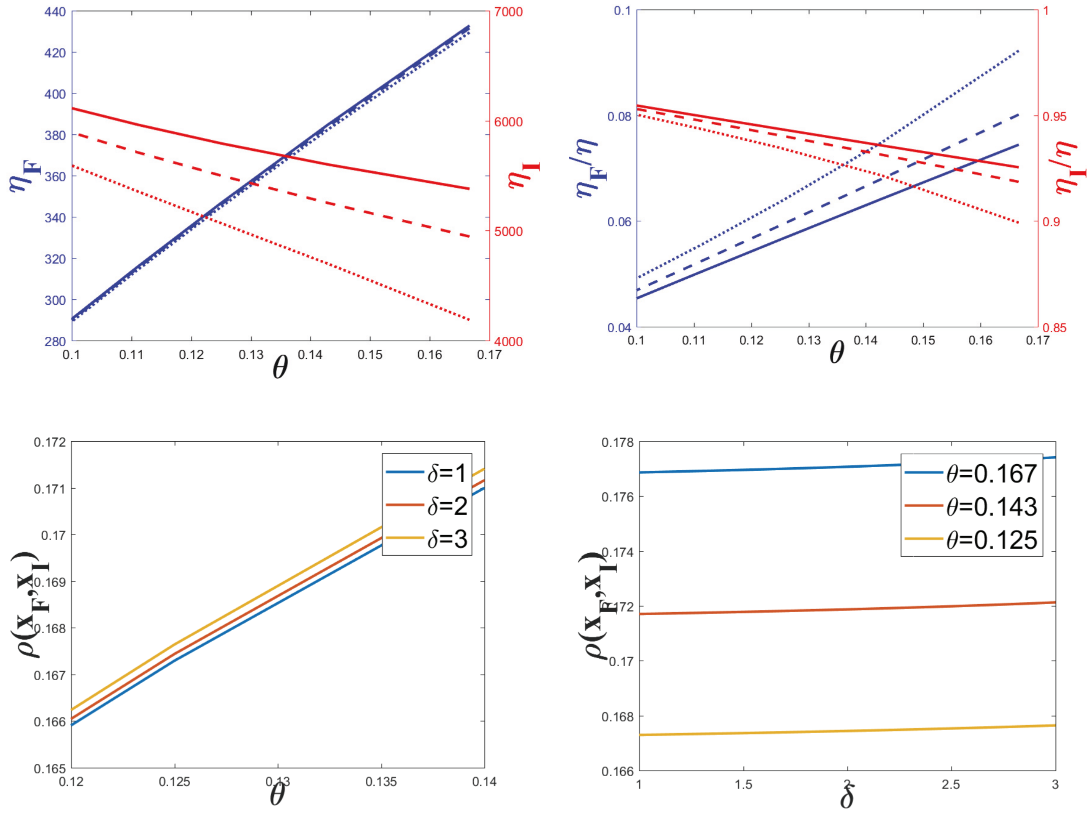


Figure 4-4

Price and trade patterns. The upper panels show the aggregate trading volumes (left) and proportional trading volumes (right) for fast (blue lines) and informed (red lines) traders with respect to information disclosure, while the solid, dashed and dotted lines correspond to the information diffusion speed $\delta = 1,2,3$, respectively; the bottom panels show the correlation coefficient between trades in the early and late periods with respect to information disclosure for three information dissemination rates (left) and information dissemination for three information disclosure levels (right); information production cost $C_I = 0.1$, speed acquisition cost $C_F = 1.5$ and other parameters are the same as in Figure 4-3.

For information dissemination, the top panels in Figure 4-4 show that the trading volume for informed traders also decreases with dissemination due to the dominating crowding out effect. Interestingly, trading volume for fast trading also decreases under this situation. This is because that, different form disclosure, faster dissemination does not bring information to fast traders but reduce trader's willingness to become fast through complementarity of speed acquisition to information production. As for the proportional trading volume, informed decreases but fast increases with dissemination. As we have shown, faster dissemination mainly crowds out faster traders indirectly, channelled by informed trading. Thus, this effect becomes more significant for informed trading but less significant for fast trading.

Trading correlation: The trading relation between the fast and informed traders in equilibrium is measured by the correlation in individual trading volume between fast and informed traders:

$$\rho\big(x_F(v_i), x_I(V, p_1)\big) = \left(1 - \lambda_1 \frac{I_F N}{n} \beta_F\right) \sqrt{\frac{n\beta_F \sigma^2}{(n - \lambda_1 I_F N \beta_F)^2 \sigma^2 + (n\lambda_1)^2 \sigma_z^2}}. \tag{4-22}$$

The bottom panels in Figure 4-4 plot the changes in the correlation coefficient with respect to information dissemination and disclosure. They show that the correlation coefficient is always positive and increasing in information disclosure and dissemination. This is consistent with Dugast and Foucault (2018). In their model, traders can trade either during the early period with a raw signal or the late period with a processed signal. When the raw signal is valid, the processed and raw signals command trades in the same direction. With information diffusion in our model, a fast trader receives a partial but valid signal, which is always positively correlated with the fundamental payoff. With increasing information disclosure, fast traders receive more precise information so their trading contains more fundamental information. Given that informed traders trade on the full information, fast and informed trading are more correlated. More interestingly, given the information disclosure level, though faster dissemination crowds out faster traders, individual faster traders still trade more aggressively due to less intra-temporal competition. Therefore, the correlation also increases in information dissemination.

4.5 Conclusion

Speed hierarchy not only motivates fast trading on low precision signals but also makes slower trades more informative. This chapter examines how improvements in fast trading technologies influence information production and spread. Given that investors can process only a subset of publicly available information, we consider heterogeneous information diffusion among different types of investors. Fast trading can aggregate heterogeneous and imprecise early signals so that equilibrium prices and orders in the early period can render slow traders more informative. Thus, slower traders' vantage point might allow them to extrapolate the informational content of previous orders and to anticipate the future behaviour of prices. In this chapter, we theoretically analyze the implications of these features of speed acquisition and information production in financial markets. Our tractable approach permits closed-form expressions to analyse trading intensities, market quality, and information and speed values. By characterizing both information disclosure level and dissemination rate into a unified information diffusion process, we show that they can have opposite impacts on asset price discovery. This phenomenon identifies a natural economic force that attenuates the information economics problem in financial markets and generates several new empirical predictions and policy implication for market regulators.

Chapter 5

Concluding Remarks

Financial markets have been undergoing a dramatic and an inescapable evolution of algorithmic trading and witnessed more episodes of extreme price movements and liquidity breaks within very short periods. The speed of trading has considerably increased in recent years, due to rapid development of information technologies and automation of the trading process. The increase in the "speed" at which trading takes place in financial markets is a defining characteristic of high frequency trading firms, such as Citadel, Virtu, Flow Traders or Jump Trading. This raises many questions about the effects of trading speed.

Markets are evolving and vulnerable to panicky plunges long before trading algorithms and fast trading technology emerged. Though there are many other potential culprits to blame for accelerating recent liquidity breaks, it is in practice almost impossible to fully dissect and diagnose the myriad triggers and drivers. However, HFTs do appear to contribute the increasing vulnerability of abrupt dislocation and be catalyst during these more recent extreme price movements and liquidity breaks, raising some concerns from policy makers and market regulators.

These concerns have led to several proposals to slow down markets and calls for regulatory interventions. For instance, some trading platforms (e.g., IEX or Alpha) use the so called "speed bumps" to delay the time elapsed between the moment at which an order arrives to a trading platform and the moment in which the order enters into the trading platform's matching engine. This thesis sheds light into the influence of the most important technological innovation, namely HFT, on liquidity evaporation, speed acquisition, and information disclosure and provides some policy implications for market regulators. This chapter summarizes the main contributions of the thesis, discusses its implications, and outlines avenues for future research in finance and information systems.

This thesis contributes to the current literature from three different aspects. Firstly, we have proposed several theoretical frameworks to examine HFT related research issues, such as liquidity breakers, HFT arms race, speed bumps, and information diffusion. More explicitly, Chapter 2 first proposes a nonlinear rational expectations equilibrium model of high-frequency endogenous liquidity provision to explore fragile liquidity. Then the interventions of slowing down trading in financial markets are justified in Chapter 3 through a theoretical model, which takes both speed acquisition of speculators and trading platforms into consideration. Finally, by characterizing both information disclosure level and dissemination rate into a unified information diffusion process, Chapter 4 shows that they can have opposite impacts on price discovery. The thesis identifies natural economic forces that attenuate the information economics problem in financial markets and generate several new empirical predictions and policy implication for market regulators. In general, these frameworks provide a systematic approach to examine the influence of high frequency trading in financial market in terms of speed, information and learning.

Secondly, our models generate several empirical and policy implications for not only policy maker but also investors and institutions in financial markets. These implications involve, among many others, how to regulate trading platform trading speed, how company should manage its information under high frequency trading environment, what is the trade-off between benefit from HFTs and possible market instabilities. Additionally, trading venues are equally interested in these topics in order to understand the role of HFT for their business models, to overcome market frictions, and to prevent market failure by regularly adapting market structure to a dynamic environment.

Specifically, Chapter 2 delivers three key implications: Firstly, markets can experience breaks regardless of the value of the underlying asset. Without changing the fundamental value, shocks to the ELP's signals can trigger dramatic price changes. That is, the equilibrium price function is discontinuous and has a jump, either increasing or decreasing, with respect to the ELP's signal when the ELP switches from being a liquidity maker to a liquidity taker. Secondly, as the information for the informed trader and the ELP becomes more similar, liquidity under the taker regime decreases while liquidity under the maker regime increases. This stronger liquidity break should raise the concern of market regulators and investors, especially as markets become more integrated. Lastly, with either decreasing in ELP's information speed, or increasing in information latency in a segmented market, liquidity is improved under both taker and maker regimes. This suggests some benefits to slow ELPs, echoing recent regulator concern about the superior trading capabilities of high-frequency traders.

While Chapter 3 focuses on the interaction of the speeds between speculators and trading platforms, leading to several key implications. Firstly, given the speed of the exchange, an endogenous speed acquisition of speculators leads to an excessive investment of speculators on the speed. This, however, does not internalize the negative externality of the speed investment on other market participants. When the speed supply from trading platforms becomes endogenous, market forces would not lead to an optimal speed level for the exchange. A welfare loss is then generated from the mismatching in speed demand and supply. Secondly, on the endogenous trading platforms' speed, policy makers, who care about long-run price discovery and simultaneously want to limit the deadweight loss (resulting from the speed investment), have a different objective than trading platforms that simply maximize operation profits. When policy maker only cares about deadweight loss, its desired speed is always higher than the endogenous equilibrium speed of the platforms, i.e., regulators desires faster speed from the exchange. When policy maker combines both objectives, the optimal speed from policy maker's perspective depends on the welfare weight on market quality. When policy maker focuses more on market quality, a lower exchange speed is desirable. Our analysis hence reflects that the intervention of slowing down trading platforms is desirable when policy makers aim to improve market quality. The findings echo recent concern from market regulators about the superior trading capabilities of fast trading.

Furthermore, based on a model of information diffusion and speed competition, Chapter 4 makes several testable predictions about price and trade patterns. First, trading volume from fast traders decreases in information dissemination but increases in information disclosure, though both positively contribute to the proportion of fast trading volume. Second, the model predicts an increasing order flow correlation between fast and informed traders in either information dissemination or disclosure. By allowing fast traders to trade in both periods, we further show that, with either an increase in dissemination or disclosure, fast traders trade “slower” in the sense that they tend to trade less in the early period and more in the late period.

Lastly, this thesis also novelty applies the reinforcement learning to solve a nonlinear Bayesian updating problems, which contributes theoretical market microstructure literature from methodology perspective and provides potential solution for exploring nonlinear equilibrium. More explicitly, Chapter 2 demonstrates how machine learning techniques can be used to analyse nonlinear equilibria in complex models of price formation. When market participants are diversified across numerous risky assets and where market structures are more complex, being able to accurately characterize how innovative and comprehensive trading behaviours affect price and liquidity dynamics is very important for investors and regulators. Such need for richer models poses significant technical challenges in the traditionally linear rational expectations equilibrium models. Chapter 2 shows how machine learning techniques, combined with model structures from microstructure theory, can provide insights to important market microstructure questions. Machine learning’s ability to handle complex problems with non-parametric characteristics, without a pre-specified functional form, is well-suited to microstructure research. Instead of relying on some parameterization of price adjustment rules, asset demand, or use of the projection method in the literature, our solution technique is more flexible. It can be applied to more complicated strategic interaction among different market participants in continuous state space. It can also be extended to study nonlinear rational expectations equilibrium models in the presence of other frictions and provide some insights into how to make financial markets better serve to our economy and society.

Appendix A: Proofs for Chapter 2

This appendix explores the properties of nonlinear equilibrium for those scenarios considered in Sections 2.1 to 2.4 and provides the discussion about the properties of linear approximation and exogenous liquidity provision considered in Sections 2.3 and 2.4.. The proofs of these properties follow a similar approach but focusing on different aspects of parameter choices with respect to the noise trading volatility, limited market participation, and number of ELPs in the market, summarized in Table A1. Accordingly, we first derive the expectation functions, price functions and threshold values for the most general scenario of parameter choices of more than two ELPs ($N \geq 2$), limited participation parameter $\tau = \{0,1\}$, and aggregate liquidity trading parameter σ_z^2. We then apply to those specific parameter choices and provide the proofs of lemmas and propositions in Table A1. We also provide supplement to include discretionary liquidity trader's decision.

A1. Expectation and Demand Functions

We first derive the equilibrium trading strategies given ELP's optimal threshold value for the cut-off endogenous liquidity provision decision and price functions. Considering symmetry information structure among N ELPs, their optimal threshold values should be the same, denoted by θ^*. Depending on the combinations of liquidity provision decision of ELPs, there are $N + 1$ equilibrium regimes: a 'taker regime' in which all ELPs scale back of liquidity provision and consume liquidity simultaneously; and N 'maker regimes' where at least one of the ELP provides liquidity.

Table A1
Characteristics of Lemmas and Propositions for Various Scenarios

		Nonlinear / linear	No. of ELP	Limited participation	Result/Property
Basic setup	Lemma 2.1	Nonlinear	1	$\tau = \{0,1\}$	Expectation functions
	Proposition 2.1	Nonlinear	1	$\tau = \{0,1\}$	Price functions
	Proposition 2.2	Nonlinear	1	$\tau = \{0,1\}$	Threshold value
	Proposition 2.3	Linear (linear approximation)	1	$\tau = \{0,1\}$	Full equilibrium
Circuit breaker with discretionary trader	Lemma 2.2	Nonlinear	1	$\tau = 0$	Price functions in early period
	Proposition 2.4	Nonlinear	1	$\tau = 0$	Price functions in later period
	Proposition 2.5	Nonlinear	1	$\tau = 0$	Discretionary trader's decision
	Lemma 2.3	Linear (exogenous liquidity)	1	$\tau = 0$	Full equilibrium
Multi-ELPs with information overlap	Lemma 2.4	Nonlinear	N	$\tau = 0$	Expectation functions
	Proposition 2.6	Nonlinear	N	$\tau = 0$	Price functions
	Proposition 2.7	Nonlinear	N	$\tau = 0$	Threshold value
	Lemma 2.5	Linear (exogenous liquidity)	N	$\tau = 0$	Full equilibrium

ELPs' market order: We first consider the endogenous liquidity provision strategies for a specific ELP i. When receiving an extreme private information signal $\tilde{\theta}_i$ (i.e., $|\tilde{\theta}_i| > \theta^*$), ELP i submits a market order to consume liquidity with an uncertainty of $j = 0, 1, \ldots, N-1$ possible equilibrium regimes with j ELPs providing liquidity, ranging from no ELP supplying liquidity to all the rest of ELPs providing liquidity. Conditional on the equilibrium price, applying Bayesian rule and calculating the conditional expectation lead to ELP i's market order,

$$x_T(\tilde{\theta}_i) = \gamma E(\tilde{V} - \tilde{p}|I_E) = \gamma \frac{\sigma_f^2 + \sigma_0^2}{\sigma_f^2 + \sigma_0^2 + \sigma_{\varepsilon,\theta}^2} \tilde{\theta}_i - \gamma f_{p,E}(\tilde{\theta}_i), \tag{A-1}$$

where the expectation function

$$f_{p,E}(\tilde{\theta}_i) = \underbrace{Pr(j = 0|\tilde{\theta}_i)\, E(\tilde{p}_T|\tilde{\theta}_i, |\tilde{\theta}_{l\neq i}| > \theta^*)}_{taker\ regime} + \underbrace{\sum_{k=1}^{N-1} Pr(j = k|\tilde{\theta}_i)\, E(\tilde{p}_M^k|\tilde{\theta}_i, j = k)}_{maker\ regimes}. \tag{A-2}$$

The expectation of ELP i in (A2) characterizes the uncertainty of the equilibrium regime when submitting market order, with $N-1$ ELPs providing liquidity at most. Also, due to information overlapping, ELP i can use his private signal to predict the equilibrium regime. The unconditional probability of $j = k$ ELPs providing liquidity can be represented by the joint probability that k out of N ELPs' signals are within the threshold value,

$$Pr(j = k) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} C_N^k \psi(x, y; \theta^*, \rho)^k (1 - \psi(x, y; \theta^*, \rho))^{N-k} f_{\tilde{\delta}_{fs}}(x) dx\, f_{\tilde{\delta}_0}(y) dy;$$

$$\psi(x, y; \theta^*, \rho) = \int_{-\infty}^{+\infty} \left[F_{\tilde{\delta}_{fi}}(\theta^* - x - y - z) - F_{\tilde{\delta}_{fi}}(-\theta^* - x - y - z)\right] f_{\tilde{\varepsilon}_\theta}(z) dz.$$

ELPs' limit order: By contrast, when receiving a moderate signal (i.e., $|\tilde{\theta}_i| \leq \theta^*$), ELP i submits a limit order to supply liquidity, conditional on his private information signal $\tilde{\theta}_i$ and the aggregate market order $\tilde{u}$. The inference problem of ELP i can be summarized by using the ELP's expectation function $f_{V,E}(\tilde{\theta}_i, \tilde{u}, \tilde{p})$ on the fundamental value. Accordingly, the ELP i's limit order is given by

$$x_M(\tilde{\theta}_i; \tilde{u}, \tilde{p}) = \gamma f_{V,E}(\tilde{\theta}_i, \tilde{u}, \tilde{p}) - \gamma \tilde{p}. \tag{A-3}$$

Similarly, ELP i also faces the uncertainty about equilibrium regime when submitting limit order and he can use his private signal to predict the equilibrium regime. However, the difference when ELP i submit a limit order is that the observed equilibrium price and aggregate order flow provide him additional information. Since DMM has no private information, when ELP i is the only liquidity supplier except the DMM, the equilibrium price only contains the information from the aggregate order flow and private information form ELP i. This means that ELP i can infer from the equilibrium price whether he is the only ELP supplying liquidity,

$$\varpi = \tilde{p} - \frac{\tilde{u}}{(1+\tau)\gamma} - \frac{E\left(\tilde{V}\middle|\tilde{\theta}_i, \tilde{u}, \left|\tilde{\theta}_{l\neq i}\right| > \theta^*\right)}{\tau + j} - \sum_{k=1}^{N} \frac{Pr(j = k|\tilde{u}, \psi \neq 0)\, E\left(\tilde{V}|\tilde{u}, \tilde{p}, j = k, \psi \neq 0\right)}{\tau + j};$$

The intuition of expression ϖ becomes clear in Appendix A2. Essentially, after receiving all the information, if $\varpi = 0$, ELP i is the only ELP supplying liquidity; otherwise ($\varpi \neq 0$), there are other ELPs (so that $j \geq 2$) providing liquidity as well. Therefore, the ELP's expectation function $f_{V,E}(\tilde{\theta}_i, \tilde{u}, \tilde{p})$ on the fundamental value can be described by

$$f_{V,E}\left(\tilde{\theta}_i, \tilde{u}, \tilde{p}\right) = E\left(\tilde{V}\middle|\tilde{\theta}_i, \tilde{u}, \left|\tilde{\theta}_{l\neq i}\right| > \theta^*\right)\mathbf{1}_{\varpi=0} + \sum_{k=2}^{N} Pr\left(j = k\middle|\tilde{\theta}_i, \tilde{u}, \varpi \neq 0\right) E\left(\tilde{V}\middle|\tilde{\theta}_i, \tilde{u}, \tilde{p}, j = k, \varpi \neq 0\right) \mathbf{1}_{\varpi\neq 0}; \tag{A-4}$$

where the unconditional probability $Pr(j = k)$ is given by (A-2).

***Inform trader's demand*:** Different from the ELPs, informed trader can only submit market order, but also faces the uncertainty about equilibrium regimes. Conditional on the equilibrium price, applying Bayesian rule and calculating the conditional expectation lead to informed reader's market order,

$$x_I(\tilde{s}) = \gamma\tilde{s} - \gamma f_{p,I}(\tilde{s}), \tag{A-5}$$

in which the expectation function of the informed about the price

$$f_{p,I}(\tilde{s}) = \underbrace{Pr(j = 0|\tilde{s})\, E\left(\tilde{p}_T|\tilde{s}, \left|\tilde{\theta}_l\right| > \theta^*\right)}_{taker\ regime} + \underbrace{\sum_{k=1}^{N} Pr(j = k|\tilde{s})\, E(\tilde{p}_M^k|\tilde{s}, j = k)}_{maker\ regimes}, \tag{A-6}$$

in which the unconditional probability $Pr(j = k)$ is given by (A-2). Different from the expectation function of the ELP in equation (A-2), for the informed trader, there can be N instead of $N - 1$ ELPs providing liquidity at most.

DMM's liquidity provision: The DMM's limit order is described by

$$x_D(\tilde{u}, \tilde{p}) = \gamma f_{V,D}(\tilde{u}, \tilde{p}) - \gamma\tilde{p}, \tag{A-7}$$

in which $f_{V,D}(\tilde{u})$ is DMM's expectation about the fundamental value. With limited market participation parameter $\tau = 0$, the DMM is completely crowded out by ELPs in the maker regimes, and is the only liquidity providers in the taker regime. Otherwise, for $\tau = 1$, ELPs and DMM share the market orders when providing liquidity.

When $\tau = 0$, the DMM knows that she can participate in the market only when all ELPs consuming liquidity. Accordingly, her expectation function can be described as

$$f_{V,D}(\tilde{u}) = E(\tilde{V}|\tilde{u}, |\tilde{\theta}_l| > \theta^*),$$

which only depends on $\tilde{u}$ not on $\tilde{p}$. This is because DMM has no private information and, when she is the only liquidity provider, $\tilde{u}$ is equivalent to $\tilde{p}$.

However, when $\tau = 1$, DMM also receive order even when ELPs provide liquidity. The DMM can also infer from equilibrium price whether she is the only liquidity provider,

$$\psi = \tilde{p} - \frac{\tilde{u}}{\gamma} - E(\tilde{V}|\tilde{u}, |\tilde{\theta}_l| > \theta^*).$$

The intuition of expression ψ becomes clear in Appendix A2. Essentially, after receiving all the information, if $\psi = 0$, DMM is the only liquidity provider supplying liquidity; otherwise ($\psi \neq 0$), there are at least one ($j = 1, 2, \ldots, N$) ELPs providing liquidity simultaneously. The expectation function is therefore described by

$$f_{V,D}(\tilde{u}, \tilde{p}) = E(\tilde{V}|\tilde{u}, |\tilde{\theta}_l| > \theta^*)\mathbf{1}_{\psi=0} + \sum_{k=1}^{N} Pr(j = k|\tilde{u}, \psi \neq 0)\, E(\tilde{V}|\tilde{u}, \tilde{p}, j = k, \psi \neq 0)\, \mathbf{1}_{\psi\neq 0};$$

Different from the ELP's expectation function $f_{V,E}(\tilde{\theta}_i, \tilde{u}, \tilde{p})$ on the fundamental value in (A4), once $\psi \neq 0$, parameter k starts with one instead of two and price contains the information other than the aggregate order flow. Hence, the DMM's expectation function also depends on equilibrium price. Combining the above two scenarios, the DMM's expectation function is given by

$$\begin{aligned} f_{V,D}(\tilde{u}, \tilde{p}) = {} & \big((1-\tau) + \tau\mathbf{1}_{\psi=0}\big)E(\tilde{V}|\tilde{u}, |\tilde{\theta}_l| > \theta^*) \\ & + \tau\mathbf{1}_{\psi\neq 0} \sum_{k=1}^{N} Pr(j = k|\tilde{u}, \psi \neq 0)\, E(\tilde{V}|\tilde{u}, \tilde{p}, j = k, \psi \neq 0); \end{aligned} \tag{A-8}$$

In summary, (A-1), (A-3), (A-5) and (A-7) describe the demand functions while (A-2), (A-4), (A-6) and (A-8) describe the expectation functions of the ELPs, informed, and DMM, respectively.

A2. Price Functions and Optimal Threshold Value

Based on the expectation functions and demand functions in Table A1, we can derive price functions. Given the optimal threshold value, we substitute the demand functions into the market clearance condition under taker and maker regimes, respectively. Under the taker regime, from (A-7), the DMM is the only liquidity provider, therefore,

$$\tilde{u} + \gamma f_{V,D}(\tilde{u}, \tilde{p}) - \gamma \tilde{p} = 0.$$

Therefore, the equilibrium price under taker regime is given by

$$\tilde{p}_T(\tilde{u}) = \frac{\tilde{u}}{\gamma} + f_{V,D}(\tilde{u}), \tag{A-9}$$

showing that price under taker regime only depends on the aggregate order flow. In this case, $\psi = 0$, meaning that DMM infers that he is the only liquidity provider in the market.

Instead, under the maker regime, both the DMM and ELPs provide liquidity. When there are j ELPs supplying liquidity, from (A3), market clearance condition leads to

$$\tilde{u} + \gamma \sum_{i=1}^{j} f_{V,E}\left(\tilde{\theta}_i, \tilde{u}, \tilde{p}\right) + \tau\gamma f_{V,D}(\tilde{u}, \tilde{p}) - (\tau + j)\gamma\tilde{p} = 0.$$

This gives the equilibrium price under maker regime.

$$\tilde{p}_M^j\left(\tilde{\theta}_1, \ldots, \tilde{\theta}_j, \tilde{u}\right) = \frac{\tilde{u}}{(\tau + j)\gamma} + \sum_{i=1}^{j} \frac{f_{V,E}\left(\tilde{\theta}_i, \tilde{u}, \tilde{p}\right)}{\tau + j} + \frac{\tau f_{V,D}(\tilde{u}, \tilde{p})}{\tau + j} \tag{A-10}$$

This equilibrium price function under maker regime provide the connection to expression ϖ.When ELP i plugs his information into expression ϖ, $\varpi = 0$ indicates that ELP i is the only ELP providing liquidity in the market. We insert $j = 1$ into equation (A10); and at the same time $\psi \neq 0$ due to there is one ELP providing liquidity in the market. This leads to the expression of ϖ.

We now characterize the endogenous equilibrium by determining the endogenous threshold value. Unlike the DMM, ELPs have no obligation to supply liquidity. Thus, ELPs endogenously provide liquidity, depending on their information and market conditions. The equilibrium threshold signal value makes ELP's expected payoffs under both the taker and maker regimes equal. Note that both the demand and expected payoff functions for ELPs can be expressed as functions of the exogenous threshold value. When receiving a signal θ, ELP's payoff in taker regime is given by

$$U_{E,T} = \gamma\left(\frac{\sigma_f^2 + \sigma_0^2}{\sigma_f^2 + \sigma_0^2 + \sigma_{\varepsilon,\theta}^2}\theta - f_{p,E}(\theta)\right)^2. \tag{A-11}$$

The payoffs under maker regimes are more complicated. After receiving all the information, if $\varpi = 0$, ELP i is the only ELP supplying liquidity; otherwise, there are more than two ELPs providing liquidity at the same time. Thus, if $\varpi = 0$, the ELP's expected payoff becomes

$$U_{E,M}^{\varpi=0} = \gamma E\left[\left(\frac{\tau E(\tilde{V}|\tilde{u},|\tilde{\theta}_l| > \theta^*)}{\tau+1} - \frac{\tilde{u}}{(\tau+1)\gamma} - \frac{\tau\sum_{k=1}^{N} Pr(j=k|\tilde{u},\psi\neq 0)\, E(\tilde{V}|\tilde{u},\tilde{p},j=k,\psi\neq 0)}{\tau+1}\right)^2 \middle| \tilde{\theta}_i = \theta\right].$$

Otherwise, the ELP's expected payoff is given by

$$U_{E,M}^{\varpi\neq 0} = E\left[\sum_{k=2}^{N} Pr(j=k|\tilde{\theta}_i,\tilde{u},\varpi \neq 0)\, E\left[\left(\frac{\tau+j-1}{\tau+j}f_{V,E}(\tilde{\theta}_i,\tilde{u},\tilde{p}) - \frac{\tilde{u}}{(\tau+j)\gamma} - \sum_{l\neq i}\frac{f_{V,E}(\tilde{\theta}_l,\tilde{u},\tilde{p})}{\tau+j} - \frac{\tau f_{V,D}(\tilde{u},\tilde{p})}{\tau+j}\right)^2 \middle| \tilde{\theta}_i, j=k,\varpi \neq 0\right]\middle| \tilde{\theta}_i = \theta, j\geq 2\right],$$

which is the expected payoff over $N-1$ market regimes. ELP can also use the signal to predict the equilibrium regimes. Therefore, the payoff in maker regimes is given by

$$U_{E,M} = \gamma\, Pr(\mathbf{1}_{\varpi=0}|\theta)\, U_{E,M}^{\varpi=0} + \gamma Pr(\mathbf{1}_{\varpi\neq 0}|\theta) U_{E,M}^{\varpi\neq 0}; \tag{A-12}$$

By equalizing the expected payoffs for the ELP in the taker and maker regimes, we then determine the optimal threshold value endogenously.

In equilibrium, the beliefs of all traders are consistent with the jointly conditional probability distribution. This implies that, given the four expectation functions, we can calculate the following intermedium variables as

$$\boldsymbol{A}(k) = \gamma\tilde{s} - \gamma\boldsymbol{f_{p,I}}(\tilde{\boldsymbol{s}}) + \gamma\frac{\sigma_f^2 + \sigma_0^2}{\sigma_f^2 + \sigma_0^2 + \sigma_{\varepsilon,\theta}^2}\tilde{\theta} - \gamma\boldsymbol{f_{p,E}}(\tilde{\boldsymbol{\theta}}) + \sum_{i=1}^{N-k-1}\left(\gamma\frac{\sigma_f^2 + \sigma_0^2}{\sigma_f^2 + \sigma_0^2 + \sigma_{\varepsilon,\theta}^2}\tilde{\theta}_i - \gamma\boldsymbol{f_{p,E}}(\tilde{\boldsymbol{\theta}}_i)\right) + \tilde{z};$$

$$\boldsymbol{B}(k) = \gamma\tilde{s} - \gamma\boldsymbol{f_{p,I}}(\tilde{\boldsymbol{s}}) + \sum_{i=1}^{N-k}\left(\gamma\frac{\sigma_f^2 + \sigma_0^2}{\sigma_f^2 + \sigma_0^2 + \sigma_{\varepsilon,\theta}^2}\tilde{\theta}_i - \gamma\boldsymbol{f_{p,E}}(\tilde{\boldsymbol{\theta}}_i)\right) + \tilde{z};$$

$$\boldsymbol{C}(k) = \gamma\tilde{s} - \gamma\boldsymbol{f_{p,I}}(\tilde{\boldsymbol{s}}) + \sum_{i=1}^{N-k}\left(\gamma\frac{\sigma_f^2 + \sigma_0^2}{\sigma_f^2 + \sigma_0^2 + \sigma_{\varepsilon,\theta}^2}\tilde{\theta}_i - \gamma\boldsymbol{f_{p,E}}(\tilde{\boldsymbol{\theta}}_i)\right) + \tilde{z}.$$

Table A2

Parameters to Lemmas and Propositions in Nonlinear Equilibrium Model

		No of ELPs	Limited participation	Liquidity trading
Lemma 2.1 Proposition 2.1 Proposition 2.2		1	$\tau = \{0,1\}$	σ_z^2
Lemma 2.2	Split	1	$\tau = 0$	$\sigma_z^2 + \sigma_l^2$
	Concentrate	1	$\tau = 0$	$\sigma_z^2 + 4\sigma_l^2$
Lemma 2.4 Proposition 2.6 Proposition 2.7		N	$\tau = 0$	σ_z^2

Thus, in the equilibrium, the expectation functions satisfy the following "fixed-function" problem.

$$\boldsymbol{f_{p,E}}(\widetilde{\boldsymbol{\theta}}) = Pr\big(j = 0\big|\tilde{\theta}\big)\, E\left(\frac{\tilde{u}}{\gamma} + f_{V,D}(\tilde{u})\Big|\tilde{\theta}, \boldsymbol{A}(0)\right)$$
$$+\sum_{k=1}^{N-1} Pr\big(j = k\big|\tilde{\theta}\big)\, E\left(\frac{\tilde{u}}{(\tau + j)\gamma} + \sum_{i=1}^{j}\frac{f_{V,E}\big(\tilde{\theta}_i, \tilde{u}, \tilde{p}\big)}{\tau + j} + \frac{\tau f_{V,D}(\tilde{u}, \tilde{p})}{\tau + j}\big|\tilde{\theta}, \boldsymbol{A}(k)\right);$$

$$\boldsymbol{f_{V,E}}(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = E\big(\tilde{V}\big|\tilde{\theta}, \boldsymbol{B}(1), \big|\tilde{\theta}_{l\neq i}\big| > \theta^*\big)\mathbf{1}_{\varpi=0}$$
$$+\sum_{k=2}^{N} Pr(j = k|\tilde{\theta}, \tilde{u}, \varpi \neq 0)\, E\big(\tilde{V}\big|\tilde{\theta}, \boldsymbol{B}(k), \tilde{p}, j = k, \varpi \neq 0\big)\, \mathbf{1}_{\varpi\neq 0};$$

$$\boldsymbol{f_{p,I}}(\tilde{\boldsymbol{s}}) = Pr(j = 0|\tilde{s})\, E\left(\frac{\tilde{u}}{\gamma} + f_{V,D}(\tilde{u})|\tilde{s}, \boldsymbol{C}(0), \big|\tilde{\theta}_l\big| > \theta^*\right)$$
$$+\sum_{k=1}^{N} Pr(j = k|\tilde{s})\, E\left(\frac{\tilde{u}}{(\tau + j)\gamma} + \sum_{i=1}^{j}\frac{f_{V,E}\big(\tilde{\theta}_i, \tilde{u}, \tilde{p}\big)}{\tau + j} + \frac{\tau f_{V,D}(\tilde{u}, \tilde{p})}{\tau + j}|\tilde{s}, \boldsymbol{C}(k), j = k\right);$$

$$\boldsymbol{f_{V,D}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \gamma\big((1 - \tau) + \tau\mathbf{1}_{\psi=0}\big)E\big(\tilde{V}\big|\boldsymbol{C}(0), \big|\tilde{\theta}_l\big| > \theta^*\big)$$
$$+\gamma\tau\mathbf{1}_{\psi\neq 0}\sum_{k=1}^{N} Pr(j = k|\tilde{u}, \psi \neq 0)\, E\big(\tilde{V}\big|\boldsymbol{C}(k), \tilde{p}, j = k, \psi \neq 0\big).$$

The above four expressions describe the 'fixed-function' problem that characterizes the threshold value of the endogenous equilibrium.

A3. Connections to Lemmas and Propositions for Nonlinear Equilibrium Model

With specific parameter choices, the results in A1 and A2 provide connections to the lemmas and propositions of different scenarios in nonlinear equilibrium model summarized in Table A2. The circuit breaker with discretionary liquidity traders in Section 2.3 involves several sub-scenarios (split or concentrate the trade), which are examined in A4 separately.

To better clarify our results, it is worth emphasizing that, in section 2.1, there is only one ELP in the market. When $\tau = 1$, DMM competes with ELP for providing liquidity. Although DMM is not sure about ELP's endogenous liquidity provision decision, DMM can infer the equilibrium regime from the equilibrium price and aggregate order flow, i.e., either DMM is the only market maker (when the equilibrium price is $\tilde{p}_T(\tilde{u})$) or ELP also makes market (when the equilibrium price is $\tilde{p}_M(\tilde{\theta}, \tilde{u})$). Hence, when $\psi \neq 0$, DMM knows that ELP is also providing liquidity and learns ELP's information from equilibrium price. This means that DMM behaves the same as ELP does. This implies

$$\sum_{k=1}^{N} Pr(j = k|\tilde{u}, \psi \neq 0)\, E(\tilde{V}|\tilde{u}, \tilde{p}, j = k, \psi \neq 0)\Bigg|_{N=1} = f_{V,E}(\tilde{\theta}, \tilde{u}),$$

resulting Lemma 2.1.

A4. Supplement Proofs for Circuit Breaker with Discretionary Liquidity Traders

This part provides additional proofs for Section 2.3 on the equilibrium in later period (Proposition 2.4) and discretionary liquidity traders' decision (Proposition 2.5). In the later period, DMM 2's limit order is described as $x_{D1'} = \gamma f_{V,D1'} - \gamma \tilde{p}_{1'}$. Based on the market clearance condition, the equilibrium price in the later period solely on DMM_l' expectation function,

$$\tilde{p}_{1\prime} = \frac{1}{\gamma}\tilde{u}_{1\prime} + f_{V,D1\prime}.$$

Depending on discretionary liquidity traders' decision, there are two types of equilibrium in the later period: concentrate or split. On the concentrate equilibrium, the discretionary liquidity trader chooses to concentrate her trading in early period, while the order flow in later period is pure noise, containing no information about the fundamental and equilibrium price in early period. Therefore, DMM_l has the same expectation (as either ELP or DMM_e, depending on the equilibrium regime) as in the early period. On the equilibrium in the later period, by observing the equilibrium order imbalance $\tilde{u}_1$ and price $\tilde{p}_1^k$ in early period, DMM_l knows the equilibrium regime in the early period (it must be in the taker regime when $\tilde{p}_1^k - \tilde{u}_1/\gamma - f_{V,D}^k(\tilde{u}_1) = 0$ or the maker regime otherwise) and bids based on the order flow $\tilde{u}_2$ in later 2 and equilibrium information in early period. When the ELP makes liquidity, observing equilibrium order imbalance $\tilde{u}_1$ and price $\tilde{p}_{1,M}^k$ mean that DMM_l can extract the value of $\tilde{\theta}$ according to Lemma 2.2. Therefore, DMM_l's expectation function is the same as ELP, i.e., $f_{V,D1\prime} = f_{V,E}^c(\tilde{\theta}, \tilde{u}_1)$. When DMM_e makes liquidity, DMM_l's expectation function is the same as DMM_e, i.e., $f_{V,D1\prime} = f_{V,D}^c(\tilde{u}_1)$.

By contrast, when the discretionary liquidity trader chooses to split her trading, DMM_l can use the order flow in later period to infer information form the order flow in early period. The order flow $\tilde{u}_2$ can only be useful when the discretionary liquidity trader chooses to split her trading across two periods. In this case, $\tilde{u}_{1\prime}$ can provide information to predict the demand of the discretionary liquidity trader in the early period so that the precision of the order flow in early period can be improved.

For example, assuming ELP takes liquidity, then the equilibrium order imbalance and price are equivalent since the equilibrium order imbalance $\tilde{u}_1$ contains information about signals $\tilde{s}$ and $\tilde{\theta}$. In addition, DMM_l can use the order flow $\tilde{u}_{1\prime}$ to predict the noise order flow (i.e., the discretionary liquidity trading $\tilde{l}$) in the early period. By Bayes' rule the two signals combine to generate the following signal in predicting the fundamental $\tilde{V}$:

$$\tilde{y} \equiv \tilde{u}_1 - \frac{\sigma_l^2}{\sigma_z^2 + \sigma_l^2}\tilde{u}_{1\prime} = x_I(\tilde{s}) + x_T(\tilde{\theta}) + \tilde{z}_1 + \tilde{l} - E(\tilde{z}_1 | \tilde{u}_{1\prime} = \tilde{z}_{1\prime} + \tilde{l})$$

$$= x_I(\tilde{s}) + x_T(\tilde{\theta}) + \tilde{z}_1 + \frac{\sigma_z^2}{\sigma_z^2 + \sigma_l^2}\tilde{l} - \frac{\sigma_l^2}{\sigma_z^2 + \sigma_l^2}\tilde{z}_{1\prime} = x_I(\tilde{s}) + x_T(\tilde{\theta}) + \tilde{n}_y,$$

which summarizes the overall information that DMM_l can extract after observing the aggregate order imbalances in both periods, in which the noise term $\tilde{n}_y$ follows

$$\tilde{n}_y \sim N\left(0, \sigma_z^2 + \frac{\sigma_z^2\sigma_l^2}{\sigma_z^2 + \sigma_l^2}\right).$$

When ELP makes liquidity, Bayes' rule delivers a similar result about the combination of the two signals, leading to the equilibrium in period 2. Therefore, when DMM_e makes liquidity, DMM_l's expectation function is similar to DMM_e but with different volatility of noise trading, i.e., $f_{V,D1'} = f_{V,D}^{y}(\tilde{y})$. When the ELP makes liquidity, observing equilibrium order imbalance $\tilde{u}_1$ and price $\tilde{p}_{1,M}^{k}$ means that DMM_l can extract the value of $\tilde{\theta}$ according to Lemma 2.2. Therefore, DMM_l's expectation function is similar to ELP but with different volatility of noise trading $f_{V,E}^{y}(\tilde{\theta}, \tilde{y})$.

We now provide the details about discretionary liquidity traders' decision (Proposition 2.5). By concentrating the order into the early period, the discretionary liquidity trader would not experience the cost of not being able to trade. Meanwhile, she will have to hold $2\tilde{l}$ units of the risky asset, which yields an expected revenue of $E[(\tilde{p}_1 - \tilde{V})2\tilde{l}]$: if $\tilde{l} > 0$, she then buys $\tilde{l}$ units for $\tilde{p}_1$ in early period and sells in later period, receiving a payoff of $\tilde{V}$; if $\tilde{l} < 0$, she then shorts $\tilde{l}$ units of the risky asset, which then generates a payoff of $\tilde{p}_1$ per unit in early period and an obligation of $\tilde{V}$ per unit in later period. Since liquidity traders are trading against speculators who have superior information, the expected revenue $E[(\tilde{p}_1 - \tilde{V})2\tilde{l}]$ is negative, representing the expected cost for liquidity trader to meet her liquidity demand.

Therefore, based on the price function in Lemma 2.2 and the given order size $2\tilde{l}$ for the discretionary liquidity trader, the expected trading loss of concentrating trades in early period is given by

$$C^c = E[(\tilde{p}_1^c - \tilde{V})2\tilde{l}] = \int_{-\infty}^{+\infty} \left\{\mu\left[\frac{2}{\gamma}\tilde{l} + f_{V,D}^{c}(2\tilde{l})\right] + (1-\mu)\left[\frac{2}{\gamma}\tilde{l} + f_{V,E}^{c}(\tilde{\theta}, 2\tilde{l})\right]\right\} \times 2\tilde{l} \times f_{\tilde{l}} d\tilde{l}.$$

Due to the endogenous liquidity provision of the ELP, the discretionary liquidity trader is uncertainty about the equilibrium regimes. Alternatively, based on the price function in Proposition 2.4 and the given order size $\tilde{l}$ for the discretionary liquidity trader in each period, the expected trading loss of splitting trades in the early period is given by

$$C_1^s = E[(\tilde{p}_1^s - \tilde{V})\tilde{l}] = \int_{-\infty}^{+\infty} \left\{\mu\left[\frac{1}{\gamma}\tilde{l} + f_{V,D}^{s}(\tilde{l})\right] + (1-\mu)\left[\frac{1}{\gamma}\tilde{l} + f_{V,E}^{s}(\tilde{\theta}, \tilde{l})\right]\right\} \times \tilde{l} \times f_{\tilde{l}} d\tilde{l}.$$

The discretionary liquidity trader also faces the uncertainty about ELP trading behaviour; however, with a different trading size. Similarly, the expected trading loss of splitting trades in later period is given by

$$C_{1'}^{s} = E\left[\left(\tilde{p}_{1'}^{s} - \tilde{V}\right)\tilde{l}\right]$$
$$= \int_{-\infty}^{+\infty}\left\{\mu\left[\frac{1}{\gamma}\tilde{l} + f_{V,D}^{y}\left(\frac{\sigma_z^2}{\sigma_z^2+\sigma_l^2}\tilde{l}\right)\right] + (1-\mu)\left[\frac{1}{\gamma}\tilde{l} + f_{V,E}^{y}\left(\tilde{\theta},\frac{\sigma_z^2}{\sigma_z^2+\sigma_l^2}\tilde{l}\right)\right]\right\} \times \tilde{l} \times f_{\tilde{l}}\, d\tilde{l}.$$

In the presence of the circuit breaker, to calculate the expected loss for splitting orders, the discretionary trader factors in the probability that the trading in later may not go through. Consequently, the total trading loss of splitting order can be described by

$$C^{s} = C_1^{s} + P(|\tilde{p}_1^{s}| \le p^{*})C_{1'}^{s} + P(|\tilde{p}_1^{s}| > p^{*})K^{*}.$$

The trading in early period will go through no matter whether the price in early period moves outside the bounds or not; however, trading might be halted in later period with probability $P(|\tilde{p}_1^{s}| > p^{*})$.

A5. Proof of Proposition 2.3: Linear Approximation

The linear approximation helps to obtain a closed-form linear solution, which is not the only way to deal with DMM's inference problem. However, this assumption is crucial for us to have an analytical solution. The main model intuition is not change if we adopt other way to explore DMM's inference problem (such as reinforcement learning), as long as the mechanism, which triggers ELP switching trading behaviour, remains the same. Besides, the assumption that there is no information overlap between informed trader and ELP is also crucial for deriving closed form solution.

We first derive equilibrium trading strategies and price functions. In the model, a trader's strategy depends on the trading strategies of the others and the market maker's pricing rule, which in turn determines traders' trading strategies. This endogenous feedback where one agent's strategy affects another's, which in turn affects their own strategy (e.g. see Foster and Viswanathan, (1996)) can be characterized as a fixed-point problem, which can complicate the analysis in general. Based on Assumption 2.1, we first conjecture an equilibrium with linear price and demand functions. We specify the linear price function as

$$\tilde{p} = \tilde{p}_T \mathbf{1}_{|\tilde{\theta}|>\theta^*} + \tilde{p}_M \mathbf{1}_{|\tilde{\theta}|\leq\theta^*} \text{ with } \tilde{p}_T = \hat{\lambda}_T \tilde{u}_T \text{ and } \tilde{p}_M = \hat{\lambda}_{\theta,M}\tilde{\theta} + \hat{\lambda}_M \tilde{u}_M,$$

where $\tilde{u}_T$ and $\tilde{u}_M$ are defined in (2-4) and (2-5) and coefficients $\hat{\lambda}_T$ and $\hat{\lambda}_M$ are determined in the equilibrium. This price function reflects the idea that ELP consumes liquidity when receiving extreme signals and provides liquidity when receiving moderate signals.

Trader i chooses the demand function x_i to maximize the expected utility conditional on his belief. The procedure to derive the equilibrium is similar to Appendices A1 and A2. The differences are that we can use linear form to express the expectation functions in (A-2), (A-4), (A-6) and (A-8) explicitly. Hence, traders' strategies can be represented linearly by

$$x_I = \hat{\beta}_I \tilde{\delta}_d; \qquad x_T = \hat{\beta}_E \tilde{\theta};$$

$$x_M = \gamma(\hat{\phi}_{E,\theta}\tilde{\theta} + \hat{\phi}_E \tilde{u} - \tilde{p}); \qquad x_D = \gamma\left(\left((1-\tau) + \tau\mathbf{1}_{\psi=0}\right)\hat{\phi}_D\,\tilde{u} + \tau\mathbf{1}_{\psi\neq0}(\hat{\phi}_{E,\theta}\tilde{\theta} + \hat{\phi}_E\tilde{u}) - \tilde{p}\right);$$

We then derive the coefficients $(\hat{\beta}_I, \hat{\beta}_E, \hat{\phi}_{E,\theta}, \hat{\phi}_E, \hat{\phi}_D)$ accordingly. We first consider the endogenous liquidity provision strategies for ELP. When receiving an extreme signal as private information $\tilde{\theta}$ (i.e., $|\tilde{\theta}| > \theta^*$), the ELP submits a market order to consume liquidity. Also, there is only one ELP in the market, hence ELP knows the equilibrium regime is the taker when submitting market order. Conditional on the equilibrium price, applying Bayesian rule and calculating the conditional expectation, we obtain the market order of the ELP,

$$x_T(\tilde{\theta}) = \gamma E(\tilde{V} - \tilde{p}|I_E) = \gamma \frac{\sigma_f^2}{\sigma_f^2 + \sigma_{\varepsilon,\theta}^2}\tilde{\theta} - \gamma\hat{\lambda}_T\hat{\beta}_E\tilde{\theta}. \tag{A-13}$$

This indicates the trading intensity of ELP's market order is given by

$$\hat{\beta}_E = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_{\varepsilon,\theta}^2}\,\frac{1}{\frac{1}{\gamma} + \hat{\lambda}_T}. \tag{A-14}$$

By contrast, when receiving a moderate signal (i.e., $|\tilde{\theta}| \leq \theta^*$), the ELP i submits a limit order to supply liquidity, conditional on their private information $\tilde{\theta}$ and the aggregate market order $\tilde{u}$. Also, there is only one ELP in the market, hence ELP knows the equilibrium regime is maker when submitting limit order. Accordingly, the ELP i's limit order is described as

$$x_M(\tilde{\theta}_i; \tilde{u}, \tilde{p}) = \gamma(\hat{\phi}_{E,\theta}\tilde{\theta} + \hat{\phi}_E\tilde{u}) - \gamma\tilde{p}. \tag{A-15}$$

Because there is no information overlap between informed trader and ELP, the aggregate order flow is independent of ELP's signal. Based on Bayesian updating, we have trading intensity of ELP's limit order

$$\hat{\phi}_{E,\theta} = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_{\varepsilon,\theta}^2}; \qquad \hat{\phi}_E = \frac{\hat{\beta}_I \sigma_d^2}{\hat{\beta}_I^2 \sigma_d^2 + \sigma_Z^2}; \tag{A-16}$$

Second, different from the ELPs, informed traders only submits market order. Different form ELPs informed trader also has uncertainty about equilibrium regimes. However, due to there is no information overlap between informed trader and ELP, informed trader cannot use his signal to predict the probability ELP takes liquidity, which is described as

$$\mu = \Pr\left(\left|\tilde{\theta}\right| > \theta^*\right) = 2N\left(-\theta^* / \sqrt{\sigma_f^2 + \sigma_{\varepsilon,\theta}^2}\right) \tag{A-17}$$

Conditional on the equilibrium price, applying Bayesian rule and calculating the conditional expectation, we obtain the market order of the ELP,

$$x_I(\tilde{s}) = \gamma\tilde{s} - \gamma[\mu\hat{\lambda}_T\left(\hat{\beta}_I\tilde{\delta}_d\right) + (1-\mu)\hat{\lambda}_M(\hat{\beta}_I\tilde{\delta}_d)]. \tag{A-18}$$

This indicates the trading intensity of ELP's market order is

$$\hat{\beta}_I = \frac{1}{\frac{1}{\gamma} + \left(\mu\hat{\lambda}_T + (1-\mu)\hat{\lambda}_M\right)}. \tag{A-19}$$

Different from equation (B2), informed trader is affected by the uncertainty about ELP's endogenous liquidity provision.

On DMM's strategies, when limited market participation parameter $\tau = 0$, ELP completely crowds out the DMM, who encounters limit market participation, and is the only liquidity providers in the taker regime. However, when $\tau = 1$, the ELPs and the DMM share the market orders.

When $\tau = 0$, DMM knows that she can participate in the market only when the ELP consumes liquidity. The demand function of ELP is given by

$$x_D = \gamma\left(\frac{COV\left(\tilde{V}, \tilde{u}_T \middle| \tilde{u}_T, \left|\tilde{\theta}\right| > \theta^*\right)}{VAR\left(\tilde{u}_T \middle| \tilde{u}_T,\ \left|\tilde{\theta}\right| > \theta^*\right)}\tilde{u} - \tilde{p}\right)$$

The expectation function only depends on $\tilde{u}$ not on $\tilde{p}$. This is because DMM has no private information and, when she is the only liquidity provider in market, $\tilde{u}$ is equivalent to $\tilde{p}$. Besides, DMM takes the fact that ELP receives extreme signal into consideration.

However, when $\tau = 1$, DMM also receives order even though ELP is also providing liquidity. However, DMM can infer whether she is the only liquidity provider supplying liquidity from equilibrium price.

$$\psi = \tilde{p} - \hat{\lambda}_T \tilde{u}_T;$$

The intuition of expression ψ is straightforward. After receiving all the information, if $\psi = 0$, DMM is the only liquidity provider supplying liquidity; otherwise, ELP is also providing liquidity at the same time. More explicitly, DMM knows which equilibrium regime in the first period (it must be in the taker regime when $\tilde{p} - \hat{\lambda}_T \tilde{u}_T = 0$ or the maker regime otherwise). If the ELP makes liquidity, observing equilibrium order imbalance and price means that DMM can extract the value of $\tilde{\theta}$. Therefore, the expectation function is described

$$f_{V,D}(\tilde{u}, \tilde{\theta}) = \frac{COV(\tilde{V}, \tilde{u}_T | \, \tilde{u}_T, |\tilde{\theta}| > \theta^*)}{VAR(\tilde{u}_T | \, \tilde{u}_T, \; |\tilde{\theta}| > \theta^*)} \, \tilde{u} \mathbf{1}_{\psi=0} + (\hat{\phi}_{E,\theta} \tilde{\theta} + \hat{\phi}_E \tilde{u}) \mathbf{1}_{\psi \neq 0};$$

Combining the above two situations together, the DMM's demand function becomes

$$x_D = \gamma \left(\left((1-\tau) + \tau \mathbf{1}_{\psi=0} \right) \hat{\phi}_D \, \tilde{u} + \tau \mathbf{1}_{\psi \neq 0} \left(\hat{\phi}_{E,\theta} \tilde{\theta} + \hat{\phi}_E \tilde{u} \right) - \tilde{p} \right);$$

with the trading intensity

$$\hat{\phi}_D = \frac{COV(\tilde{V}, \tilde{u}_T | \, \tilde{u}_T, |\tilde{\theta}| > \theta^*)}{VAR(\tilde{u}_T | \, \tilde{u}_T, \; |\tilde{\theta}| > \theta^*)} = \frac{\hat{\beta}_E \bar{\sigma}_f^2 + \hat{\beta}_I \sigma_d^2}{\hat{\beta}_E^2 \bar{\sigma}_f^2 + \hat{\beta}_E^2 \bar{\sigma}_{\varepsilon,\theta}^2 + \hat{\beta}_I^2 \sigma_d^2 + \sigma_z^2}. \tag{A-20}$$

When updating posterior fundamental value, DMM takes the ELP's cut-off strategy about endogenous liquidity provision into consideration. Also, in order to have an analytical form, we assume that, when calculating posterior expectation about fundamental value, DMM applies projection theorem by plugging the posterior variance. In other word, DMM can extract information about the posterior distribution of fundamental volatility (σ_f^2) and signal precision ($\sigma_{\varepsilon,\theta}^2$) based on different equilibrium regime. We use the parameters to represent these posterior variance

$$\bar{\sigma}_f^2 = VAR\left[\tilde{\delta}_f \, \middle| |\tilde{\theta}| > \theta^*\right]; \bar{\sigma}_{\varepsilon,\theta}^2 = VAR\left[\tilde{\varepsilon}_\theta \, \middle| |\tilde{\theta}| > \theta^*\right].$$

We next show how to update the posterior distribution for the fundamental component $\tilde{\delta}_f$ and information noise $\tilde{\varepsilon}_\theta$ through the definition of variance and expectation. Here, we focus on the fundamental component δ_f. The noise part ε_θ follows a similar derivation and is thus omitted. We first focus on the maker regime updating,

$$E\left[\tilde{\delta}_f \, \middle| |\tilde{\theta}| \leq \theta^*\right] = E\left[E[\tilde{\delta}_f | -\theta^* - \varepsilon_\theta < \tilde{\delta}_f < \theta^* - \varepsilon_\theta]\right]. \tag{A-21}$$

In (B9), the outer layer expectation on the right-hand side is to calculate the expectation over noise term ε_θ,

$$E\left[\tilde{\delta}_f \left||\tilde{\theta}| \leq \theta^*\right.\right] = \int_{-\infty}^{\infty} \left\{E\left[\tilde{\delta}_f \middle| -\theta^* - t < \tilde{\delta}_f < \theta^* - t\right]\right\} dF_\varepsilon(t).$$

For the variance, it is more complicated. Based on the definition, we have

$$VAR\left[\tilde{\delta}_f \left||\tilde{\theta}| \leq \theta^*\right.\right] = \underbrace{E\left[VAR\left[\tilde{\delta}_f \middle| -\theta^* - \varepsilon_\theta < \tilde{\delta}_f < \theta^* - \varepsilon_\theta\right]\right]}_{Expacted\ conditional\ variance} + \underbrace{VAR\left[E\left[\tilde{\delta}_f \middle| -\theta^* - \varepsilon_\theta < \tilde{\delta}_f < \theta^* - \varepsilon_\theta\right]\right]}_{Variance\ of\ conditional\ expectation}. \tag{A-22}$$

On the first component (the expected conditional variance) in (B10), we have

$$E\left[VAR\left[\tilde{\delta}_f \middle| -\theta^* - \varepsilon_\theta < \tilde{\delta}_f < \theta^* - \varepsilon_\theta\right]\right] = \int_{-\infty}^{\infty} \left\{VAR\left[\tilde{\delta}_f - \theta^* - t < \tilde{\delta}_f < \theta^* - t\right]\right\} dF_\varepsilon(t).$$

For the second component (the variance of conditional expectation) in (B10), we have

$$VAR\left[E\left[\tilde{\delta}_f \middle| -\theta^* - \varepsilon_\theta < \tilde{\delta}_f < \theta^* - \varepsilon_\theta\right]\right] = \int_{-\infty}^{\infty} \left\{E\left[\tilde{\delta}_f \middle| -\theta^* - t < \tilde{\delta}_f < \theta^* - t\right] - E\left[\tilde{\delta}_f \left||\tilde{\theta}| \leq \theta^*\right.\right]\right\}^2 dF_\varepsilon(t).$$

Therefore, the posterior distribution is given by the truncated normal distribution,

$$E\left[\tilde{\delta}_f \middle| -\theta^* - t < \tilde{\delta}_f < \theta^* - t\right]$$

and

$$VAR\left[\tilde{\delta}_f \middle| -\theta^* - t < \tilde{\delta}_f < \theta^* - t\right].$$

Both expressions have closed form solutions. Consequently, we obtain the posterior distributions about the fundamental component $\tilde{\delta}_f$ and information noise $\tilde{\varepsilon}_\theta$ for maker regime.

We then examine the posterior distribution of the fundamental component $\tilde{\delta}_f$ and information noise $\tilde{\varepsilon}_\theta$ under the taker regime. To updating the posterior mean, similar to the maker regime, we have

$$E\left[\tilde{\delta}_f \left||\tilde{\theta}| > \theta^*\right.\right] = \int_{-\infty}^{\infty} \left\{E\left[\tilde{\delta}_f \middle| \theta < -\theta^* - t\ or\ \theta > \theta^* - t\right]\right\} dF_\varepsilon(t).$$

To updating the posterior variance, based on the truncated normal distribution, we have

$$VAR[\tilde{\delta}_f] = \mu VAR[\tilde{\delta}_f|\tilde{\theta}| > \theta^*] + (1-\mu)VAR\left[\tilde{\delta}_f \left||\tilde{\theta}| \leq \theta^*\right.\right]$$
$$+ \mu\left(E\left[\tilde{\delta}_f \left||\tilde{\theta}| > \theta^*\right.\right] - E[\tilde{\delta}_f]\right)^2$$
$$+ (1-\mu)\left(E\left[\tilde{\delta}_f \left||\tilde{\theta}| \leq \theta^*\right.\right] - E[\tilde{\delta}_f]\right)^2.$$

Solving the above equation and simplifying the expression, we have

$$VAR\left[\tilde{\delta}_f \left||\tilde{\theta}| > \theta^*\right.\right] = \frac{VAR[\tilde{\delta}_f]}{\mu} - \left(\frac{1}{\mu} - 1\right)VAR\left[\tilde{\delta}_f \left||\tilde{\theta}| \leq \theta^*\right.\right]$$
$$- \left(E\left[\tilde{\delta}_f \left||\tilde{\theta}| > \theta^*\right.\right] - E[\tilde{\delta}_f]\right)^2 - (1/\mu - 1)\left(E\left[\tilde{\delta}_f \left||\tilde{\theta}| \leq \theta^*\right.\right] - E[\tilde{\delta}_f]\right)^2.$$

This procedure of updating the posterior variance is used in DMM's inference problem.

To solve the equilibrium, we substitute the demand functions into the market clearance condition under taker and maker regimes, respectively. Under the taker regime, the DMM is the only liquidity provider, therefore, $\gamma\left(\left((1-\tau) + \tau\mathbf{1}_{\psi=0}\right)\hat{\phi}_D\,\tilde{u} + \tau\mathbf{1}_{\psi\neq 0}\left(\hat{\phi}_{E,\theta}\tilde{\theta} + \hat{\phi}_E\tilde{u}\right) - \tilde{p}\right) + u = 0$. Note that $\psi = 0$, it is taker regime. Comparing to the equilibrium price under taker regime $\tilde{p}_T = \lambda_T\tilde{u}$, we have

$$\hat{\lambda}_T = \hat{\phi}_D + \frac{1}{\gamma}. \tag{A-23}$$

Under the maker regime, ELP provides liquidity and DMM supplies liquidity, depending on the market participation condition. Hence, we have the market clearance condition $\tilde{u} + \gamma\left(\left((1-\tau) + \tau\mathbf{1}_{\psi=0}\right)\hat{\phi}_D\,\tilde{u} + \tau\mathbf{1}_{\psi\neq 0}\left(\hat{\phi}_{E,\theta}\tilde{\theta} + \hat{\phi}_E\tilde{u}\right) - \tilde{p}\right) + \gamma\left(\hat{\phi}_{E,\theta}\tilde{\theta} + \hat{\phi}_E\tilde{u} - \tilde{p}\right) = 0$. Note that it is maker regime, so that $\psi \neq 0$. Comparing to the equilibrium price under maker regime $\tilde{p}_M = \lambda_{\theta,M}\tilde{\theta} + \lambda_M\tilde{u}$, we have

$$\hat{\lambda}_M = \frac{1}{\gamma(1+\tau)} + \hat{\phi}_E;\ \hat{\lambda}_{\theta,M} = \hat{\phi}_{E,\theta}. \tag{A-24}$$

Lastly, we characterize the endogenous equilibrium by determining the endogenous threshold value. Unlike the DMM, the ELP has no obligation to supply liquidity. Thus, the ELP endogenously provides liquidity, depending on their information and market conditions. The equilibrium threshold signal value equates the ELP's expected payoffs under both the taker and maker regimes. Note that both the demand and expected payoff functions for the ELP can be expressed as functions of the exogenous threshold value. Assuming the signal ELP receive is θ. Calculating the expected payoff for taker, we have $U_{E,T} = \beta_E^2\theta^2/2\gamma$. Expected payoff maker regimes is given by

$$
\begin{aligned}
U_{E,M} &= E\left[\frac{\gamma}{2}\left(\hat{\phi}_{E,\theta}\theta + \hat{\phi}_E\tilde{u} - \tilde{p}\right)^2\right] = E\left[\frac{\gamma}{2}\left((\hat{\phi}_{E,\theta} - \hat{\lambda}_{\theta,M})\theta + (\hat{\phi}_E - \hat{\lambda}_M)\tilde{u}\right)^2\right] \\
&= \frac{\gamma}{2}\left(\hat{\phi}_{E,\theta} - \hat{\lambda}_{\theta,M}\right)^2\theta^2 + \frac{\gamma}{2}\left(\hat{\phi}_E - \hat{\lambda}_M\right)^2(\beta_I^2\sigma_d^2 + \sigma_z^2) \\
&= \frac{1}{2\gamma}\frac{\beta_I^2\sigma_d^2 + \sigma_z^2}{(1+\tau)^2};
\end{aligned}
$$

Therefore, comparing the two payoffs, the equilibrium condition implies that the threshold value can be represented as

$$
\theta^* = \left(\hat{\lambda}_M - \hat{\phi}_E\right)\sqrt{\frac{\left(\hat{\beta}_I^2\sigma_d^2 + \sigma_z^2\right)}{\beta_E^2 - \left(\phi_{E,\theta} - \lambda_{\theta,M}\right)^2}}.. \tag{A-25}
$$

Therefore (A-14), (A-16), (A-17), (A-19), (A-20), (A-23), (A-24) and (A-25) summarize equilibrium in Proposition 2.3.

A6. Proofs for Lemma 2.4 and Lemma 2.5

Both Lemmas 2.4 and 2.5 are about exogenous liquidity provision. We first discuss Lemma 2.5. The proofs for Lemma 2.4 is similar except focusing on specific parameters choices such as noise trading volatility and number of ELPs in the market. We first derive the expectation functions, price functions and threshold values for Lemma 2.5. We then link these to Lemma 2.4. Lastly, we provide a supplement analysis on discretionary liquidity trader's decision. When the ELP's liquidity provision is given exogenously, the ELP randomly chooses to either consume liquidity with probability μ_i or supply liquidity with probability $(1-\mu_i)$, that is,

$$
x_{E,i} = x_{T,i}(\tilde{\theta})\,\mathbf{1}_{\varphi_i=1} + x_{M,i}(\tilde{\theta};\,\tilde{u},\tilde{p}\,)\mathbf{1}_{\varphi_i=0},
$$

where φ_i is a binomial random variable, independent of others, satisfying $P(\varphi_i = 1) = \mu_i$ and $P(\varphi_i = 0) = 1 - \mu_i$. Hence, the ELP's liquidity provision decision contains no information about the fundamental payoff. Furthermore, we assume that two ELPs' exogenous liquidity provision decisions are independent.

In Lemma 2.5, we consider a case with two ELPs in the financial market, i.e., $N = 2$ to illustrate the influence of increasing information similarity between ELPs. This simplifies the calculation. The market making ELP i only faces two situations, either DMM is the only market maker (the equilibrium price is then given by $\tilde{p}_{i,M}(\tilde{\theta}_i, \tilde{u})$) or both of ELPs make market (when the equilibrium price is given by $\tilde{p}_{M,M}(\tilde{\theta}_1, \tilde{\theta}_2, \tilde{u})$). Hence, when $\varpi \neq 0$, ELP is uncertain if both ELPs are providing liquidity and would learn the other ELP's information from equilibrium price. This means that the ELPs' demand functions do not contain the equilibrium price (use both private information instead).[86] Besides, with the limited market participation parameter i$\tau = 0$, this further simplifies the computation. Similar to Appendix A5, we start with linear forms of demand functions and price functions.

$$\tilde{p} = \tilde{p}_T \mathbf{1}_{\varphi_1=1}\mathbf{1}_{\varphi_2=1} + \tilde{p}_{1,M}\mathbf{1}_{\varphi_1=0}\mathbf{1}_{\varphi_2=1} + \tilde{p}_{2,M}\mathbf{1}_{\varphi_1=1}\mathbf{1}_{\varphi_2=0} + \tilde{p}_{M,M}\mathbf{1}_{\varphi_1=0}\mathbf{1}_{\varphi_2=0}$$

With $\tilde{p}_T = \hat{\lambda}_T \tilde{u}_T, \tilde{p}_{1,M} = \lambda_{1,\theta,M}\tilde{\theta}_1 + \lambda_{1,M}\tilde{u}\,, \tilde{p}_{2,M} = \lambda_{2,\theta,M}\tilde{\theta}_2 + \lambda_{2,M}\tilde{u}$ and $\tilde{p}_{M,M} = \lambda_{M,\theta,M}(\tilde{\theta}_1 + \tilde{\theta}_2) + \lambda_{M,M}\tilde{u}$,

we have four different price functions depends on whether ELPs provide liquidity or take liquidity. Similarly, demand functions can be described as

$$x_I = \hat{\beta}_I \tilde{\delta}_d; \qquad x_{T,i} = \hat{\beta}_{E,i}\tilde{\theta}; \qquad x_D = \gamma(\hat{\phi}_D\, \tilde{u} - \tilde{p});$$

$$x_{M,i} = \gamma\left(\left(\phi_{i,E,\theta}\tilde{\theta} + \phi_{i,E}\tilde{u}\right)\mathbf{1}_{\varpi=0} + \left(\phi_{M,E,\theta}(\tilde{\theta}_1 + \tilde{\theta}_2) + \phi_{M,E}\tilde{u}\right)\mathbf{1}_{\varpi\neq 0} - \tilde{p} \right);$$

We start with DMM demand function. Different from Appendix B1, the ELP's liquidity provision decision contains no information about the fundamental payoff. When $\tau = 0$, DMM knows that she can participate in the market only when all ELPs are consuming liquidity. The demand function of ELP is

$$x_D = \gamma\phi_D u - \gamma p,$$

while the trading intensity for DMM is

$$\hat{\phi}_D = \frac{(\bar{\beta}_I + \bar{\beta}_{1,E} + \bar{\beta}_{2,E})\sigma_0^2 + (\bar{\beta}_{1,E} + \bar{\beta}_{2,E})\sigma_f^2 + \bar{\beta}_I\sigma_d^2}{(\bar{\beta}_I + \bar{\beta}_{1,E} + \bar{\beta}_{2,E})^2\sigma_0^2 + (\bar{\beta}_{1,E} + \bar{\beta}_{2,E})^2\rho\sigma_f^2 + (\bar{\beta}_{1,E}^2 + \bar{\beta}_{2,E}^2)\left((1-\rho)\sigma_f^2 + \sigma_{\varepsilon,\theta}^2\right) + \bar{\beta}_I^2\sigma_d^2 + \sigma_z^2}; \tag{A-26}$$

We then consider ELP's optimal trading strategies. We first consider the limit order. When ELP i the only liquidity supplier except for DMM, there is no private information for DMM and the equilibrium price only contains the information from aggregate order flow and private information form ELP i. This means ELP i can infer whether he is the only ELP supplying liquidity from equilibrium price.

---

[86] Otherwise, when $N \geq 3$, if $\varpi \neq 0$ ELP i would not know whether there are one or more ELPs making market except for himself. Hence, ELP i could infer other ELPs' signal but only depends on price to form the expectation about fundamental value. This, as mentioned before, will leads to a no solution situation for reinforcement learning.

$$\varpi = \tilde{p} - \lambda_{i,\theta,M}\tilde{\theta}_i + \lambda_{i,M}\tilde{u};$$

The intuition of expression ϖ is straightforward. After receiving all the information, if $\varpi = 0$, ELP i is the only ELP supplying liquidity; otherwise, there are two ELPs providing liquidity at the same time. We focus on the situations with $\varpi = 0$ first. In this case, applying Bayesian rule and calculating the conditional expectation, we obtain the trading intensity of the ELP i

$$\phi_{i,E,\theta} = \frac{Cov(V,\tilde{\theta}_i)Var(\tilde{u}) - Cov(\tilde{\theta}_i,\tilde{u})Cov(V,\tilde{u})}{Var(\tilde{\theta}_i)Var(\tilde{u}) - Cov(\tilde{\theta}_i,\tilde{u})^2}; \tag{A-27}$$

$$\phi_{i,E} = \frac{Cov(V,\tilde{u})Var(\tilde{\theta}_i) - Cov(\tilde{\theta}_i,\tilde{u})Cov(V,\tilde{\theta}_i)}{Var(\tilde{\theta}_i)Var(\tilde{u}) - Cov(\tilde{\theta}_i,\tilde{u})^2},$$

here the expressions satisfies

$$\begin{aligned}
&Var(\tilde{\theta}_i)Var(\tilde{u}) - Cov(\tilde{\theta}_i,\tilde{u})^2 \\
&\quad = (\sigma_0^2 + \sigma_f^2 + \sigma_{\varepsilon,\theta}^2)\left((\bar{\beta}_I + \bar{\beta}_{1,E})^2\sigma_0^2 + \bar{\beta}_{1,E}^2\sigma_f^2 + \bar{\beta}_{1,E}^2\sigma_{\varepsilon,\theta}^2 + \bar{\beta}_I^2\sigma_d^2 + \sigma_z^2\right) \\
&\quad - \left((\bar{\beta}_I + \bar{\beta}_{1,E})\sigma_0^2 + \bar{\beta}_{1,E}\rho\sigma_f^2\right)^2;
\end{aligned}$$

$$\begin{aligned}
&Cov(V,\tilde{\theta}_i)Var(\tilde{u}) - Cov(\tilde{\theta}_i,\tilde{u})Cov(V,\tilde{u}) \\
&\quad = (\sigma_0^2 + \sigma_f^2)\left((\bar{\beta}_I + \bar{\beta}_{1,E})^2\sigma_0^2 + \bar{\beta}_{1,E}^2\sigma_f^2 + \bar{\beta}_{1,E}^2\sigma_{\varepsilon,\theta}^2 + \bar{\beta}_I^2\sigma_d^2 + \sigma_z^2\right) \\
&\quad - \left((\bar{\beta}_I + \bar{\beta}_{1,E})\sigma_0^2 + \bar{\beta}_{1,E}\rho\sigma_f^2\right)\left((\bar{\beta}_I + \bar{\beta}_{1,E})\sigma_0^2 + \bar{\beta}_{1,E}\sigma_f^2 + \bar{\beta}_I\sigma_d^2\right);
\end{aligned}$$

$$\begin{aligned}
&Cov(V,\tilde{u})Var(\tilde{\theta}_i) - Cov(\tilde{\theta}_i,\tilde{u})Cov(V,\tilde{\theta}_i) \\
&\quad = \left((\bar{\beta}_I + \bar{\beta}_{1,E})\sigma_0^2 + \bar{\beta}_{1,E}\sigma_f^2 + \bar{\beta}_I\sigma_d^2\right)(\sigma_0^2 + \sigma_f^2 + \sigma_{\varepsilon,\theta}^2) \\
&\quad - \left((\bar{\beta}_I + \bar{\beta}_{1,E})\sigma_0^2 + \bar{\beta}_{1,E}\rho\sigma_f^2\right)(\sigma_0^2 + \sigma_f^2).
\end{aligned}$$

On the situations with $\varpi = 1$, ELP i can infer the private information about another ELP from equilibrium price. Therefore, in equilibrium, both ELPs knows information about each other. Hence, they behave the same. Due to the same information structures for both ELPs, we can use $\tilde{\theta}_1 + \tilde{\theta}_2$ to summarize the aggregate private information. Under this situation, applying Bayesian rule and calculating the conditional expectation, we obtain the trading intensity of the ELP i

$$\phi_{M,E,\theta} = \frac{1}{2}\frac{Cov(V,\tilde{n})Var(\tilde{u}) - Cov(\tilde{n},\tilde{u})Cov(V,\tilde{u})}{Var(\tilde{n})Var(\tilde{u}) - Cov(\tilde{n},\tilde{u})^2}; \tag{A-28}$$

$$\phi_{M,E} = \frac{Cov(V,\tilde{u})Var(\tilde{n}) - Cov(\tilde{n},\tilde{u})Cov(V,\tilde{n})}{Var(\tilde{n})Var(\tilde{u}) - Cov(\tilde{n},\tilde{u})^2}.$$

With the expression satisfy

$$n=\frac{\tilde{\theta}_1+\tilde{\theta}_2}{2}=\tilde{\delta}_0+\delta_{fs}+\frac{\left(\delta_{f1}+\delta_{f2}\right)}{2}+\frac{\varepsilon_{\theta,1}+\varepsilon_{\theta,2}}{2};$$

$$ar(\tilde{n})Var(\tilde{u})-Cov(\tilde{n},\tilde{u})^2=\left(\sigma_0^2+\frac{1+\rho}{2}\sigma_f^2+\frac{\sigma_{\varepsilon,\theta}^2}{2}\right)\left(\bar{\beta}_I^2\sigma_0^2+\bar{\beta}_I^2\sigma_d^2+\sigma_z^2\right)-\left(\bar{\beta}_I\sigma_0^2\right)^2;$$

$$Cov(V,\tilde{n})Var(\tilde{u})-Cov(\tilde{n},\tilde{u})Cov(V,\tilde{u})=\left(\bar{\beta}_I^2\sigma_0^2+\bar{\beta}_I^2\sigma_d^2+\sigma_z^2\right)\sigma_f^2+\sigma_0^2\sigma_z^2;$$

$$Cov(V,\tilde{u})Var(\tilde{n})-Cov(\tilde{n},\tilde{u})Cov(V,\tilde{n})=\left(\bar{\beta}_I\sigma_0^2+\bar{\beta}_I\sigma_d^2\right)\left(\sigma_0^2+\frac{1+\rho}{2}\sigma_f^2+\frac{\sigma_{\varepsilon,\theta}^2}{2}\right)-\left(\bar{\beta}_I\sigma_0^2\right)\left(\sigma_0^2+\sigma_f^2\right).$$

On ELP's market order and informed trader's market order. They both face the uncertainty about equilibrium regime. Conditional on the equilibrium price, applying Bayesian rule and calculating the conditional expectation, we obtain the trading intensities of the ELP and DMM,

$$\begin{aligned}\bar{\beta}_I=\gamma\Bigg[1&-\mu_1\mu_2\lambda_T\left(\bar{\beta}_I+\left(\bar{\beta}_{1,E}+\bar{\beta}_{2,E}\right)\frac{\sigma_0^2}{\sigma_0^2+\sigma_d^2}\right)\\&-\mu_1(1-\mu_2)\left(\lambda_{2,\theta,M}\frac{\sigma_0^2}{\sigma_0^2+\sigma_d^2}+\lambda_{2,M}\left(\bar{\beta}_I+\bar{\beta}_{1,E}\frac{\sigma_0^2}{\sigma_0^2+\sigma_d^2}\right)\right)\\&-\mu_2(1-\mu_1)\left(\lambda_{1,\theta,M}\frac{\sigma_0^2}{\sigma_0^2+\sigma_d^2}+\lambda_{1,M}\left(\bar{\beta}_I+\bar{\beta}_{2,E}\frac{\sigma_0^2}{\sigma_0^2+\sigma_d^2}\right)\right)\Bigg]\\&-(1-\mu_1)(1-\mu_2)\left(\lambda_{M,\theta,M}\frac{2\sigma_0^2}{\sigma_0^2+\sigma_d^2}+\lambda_{M,M}\bar{\beta}_I\right).\end{aligned}\tag{A-29}$$

$$\begin{aligned}\bar{\beta}_{i,E}=\gamma\Bigg[&\frac{\sigma_0^2+\sigma_f^2}{\sigma_0^2+\sigma_f^2+\sigma_{\varepsilon,\theta}^2}-\mu_j\lambda_T\left(\bar{\beta}_I\frac{\sigma_0^2}{\sigma_0^2+\sigma_f^2+\sigma_{\varepsilon,\theta}^2}+\bar{\beta}_{i,E}+\bar{\beta}_{j,E}\frac{\sigma_0^2+\rho\sigma_f^2}{\sigma_0^2+\sigma_f^2+\sigma_{\varepsilon,\theta}^2}\right)\\&-\left(1-\mu_j\right)\Bigg(\lambda_{j,\theta,M}\frac{\sigma_0^2+\rho\sigma_f^2}{\sigma_0^2+\sigma_f^2+\sigma_{\varepsilon,\theta}^2}\\&+\lambda_{j,M}\left(\bar{\beta}_I\frac{\sigma_0^2}{\sigma_0^2+\sigma_f^2+\sigma_{\varepsilon,\theta}^2}+\bar{\beta}_{i,E}\right)\Bigg)\Bigg].\end{aligned}\tag{A-30}$$

Comparing (A-29) and (A-30), informed trader face four different regimes, while liquidity taking ELP only faces uncertainty about another ELP. With all demand functions, plugging the demand functions into market clearance condition, we have

$$\begin{aligned}&\lambda_{i,\theta,M}=\phi_{i,E,\theta},\qquad \lambda_{i,M}=\frac{1}{\gamma}+\phi_{i,E},\\&\lambda_{M,\theta,M}=\phi_{M,E,\theta},\quad \lambda_{M,M}=\frac{1}{2\gamma}+\phi_{M,E}.\end{aligned}\tag{A-31}$$

Equations (A-26) to (A-31) describe equilibrium in Lemma 2.5. Lemma 2.4 is similar to Lemma 2.5 except different parameter choice in Table A2 and is therefore omitted.

Appendix B: Reinforcement Learning for Nonlinear Equilibrium

As outlined in Section 2.1, this appendix provides further details about how to employ reinforcement learning technique to generate nonlinear equilibrium.[87] Reinforcement learning algorithms are often highly non-parametric and do not pre-specify functional forms. They are designed to be adaptive to extract patterns or functions, which may not be recognized by parametric models ex ante, from observations through repeated bootstrapping. As a result, reinforcement learning algorithms not only provides more precise results for more complicated models but also brings new evidence and insights into theoretical microstructure research. In the following, we first outline the basic setup and general procedure for reinforcement learning approach in Appendix B1. Next, we illustrate how to generate observations for updating based on a general situation in Appendix B2. Appendix B3 summarizes the difference among benchmark model, circuit breaker scenario and the general scenario considered in Appendix B2.

B1. Basic Setup and General Procedure

[87] Reinforcement learning techniques can be used to solve the nonlinear rational expectations equilibrium in a more complicated market microstructure setting. To examine strategic trading behaviours of different group of investors with asymmetry information and their impact on price efficiency and market liquidity, most of microstructure models relies on linear rational expectations equilibrium. With nonlinear equilibrium (such as in this chapter), this becomes very challenging. To overcome this challenging, we introduce the reinforcement learning algorithm to solve the nonlinear equilibrium numerically.

Table B1: Function and cut-off strategy set for different scenarios

	Function set	Cut-off strategy set
Benchmark model	$F = \left\{f_{\{p,I\}}, f_{\{V,E\}}, f_{\{p,E\}}, f_{\{V,D\}}, U_{\{E,M\}}, U_{\{E,T\}}\right\}$	$\Theta = \{\hat{\theta}\}$
Two ELPs scenario	$F = \left\{f_{\{p,I\}}, f_{\{V,1E\}}, f_{\{V,2E\}}, f_{\{p,E\}}, f_{\{V,D\}}, U_{\{E,M\}}, U_{\{E,T\}}\right\}$	$\Theta = \{\hat{\theta}_1, \hat{\theta}_2\}$
Circuit breaker scenario	$F = \left\{f_{\{p,I\}}, f_{\{V,E\}}, f_{\{p,E\}}, f_{\{V,D\}}, f_{\{V,D1\prime T\}}\ f_{\{V,D1\prime M\}}, U_{\{E,M\}}, U_{\{E,T\}}\right\}$	$\Theta = \{\hat{\theta}_c, \hat{\theta}_s\}$

The basic setups and procedures for applying reinforcement learning technique and simplifying the expression include four main parts: a) notations, b) discretization, c) inference, and d) convergence. The convergence further includes two parts: the one for determine the financial market equilibrium (given the cut-off strategy for ELP's liquidity provision) and the one for governing ELP's endogenous liquidity provision. The general idea of reinforcement learning technique is to delivering stable and consistent outcomes through gradually adjusting inference and equilibrium parameters with simulations. More explicitly, with the initialization of model parameters and expectations, traders trade accordingly. Based on the outcomes from market clearing mechanism, traders calculate intermediate expectations and then update their expectations repetitively until the convergence criteria are satisfied. We then regard the expectations as the equilibrium of the model and analyze the characteristics of the equilibrium. We now provide a general explanation for each part based on the baseline model.

<u>Setups</u>: Based on Subsection 2.1.4, the nonlinear equilibrium is characterized by two parts. On the one hand, the financial market equilibrium is fully depicted by traders expectation functions. For instance, in the benchmark model, there are four expectation functions, which are f_h with $h \in \left\{\{p,I\};\{p,E\};\{V,E\};\{V,D\}\right\}$. On the other hand, the ELP's liquidity provision decision is determined by the cut-off strategy. Specifically, the optimal threshold value θ^* for the cut-off strategy is determined through comparing ELP's realized profit of taking liquidity, $U_{\{E,T\}}$, and making liquidity, $U_{\{E,M\}}$. Denote F the function set for describing the expectation and utility functions and Θ the cut-off strategy set for governing ELP's endogenous liquidity provision decision. In the benchmark model, $F = \left\{f_{\{p,I\}}, f_{\{V,E\}}, f_{\{p,E\}}, f_{\{V,D\}}, U_{\{E,M\}}, U_{\{E,T\}}\right\}$ and $\Theta = \{\hat{\theta}\}$. For other scenarios, the function and cut-off strategy sets are listed in Table B1.

Reaching the nonlinear equilibrium involves two stages, as illustrated in the upper panel in Figure B1. First, given the threshold value of cut-off strategy (for governing ELP's endogenous liquidity provision decision), we update the expectation functions to generate financial market equilibrium. This process can be treated as exogenous financial market equilibrium with given threshold value. Second, to find the equilibrium threshold value of the cut-off strategy, we start with an initial cut-off strategy set,

$$\Theta^{m=0} = \left\{\hat{\theta}^{m=0} = 2\sqrt{\sigma_0^2 + \sigma_d^2 + \sigma_{\varepsilon,\theta}^2}\right\}.$$

Here m denotes the m^{th}round and $\Theta^m = \{\hat{\theta}^m\}$ the m^{th} updating of threshold value.

In each round, traders' expectation functions are updated to reach financial market equilibrium. The resulting utility functions are then used to update the threshold value. Donate k_m as the sequential of updating of function set in round m for the k_m-th updating and F^{k_m} the $k_m - th$ updating function set. Initially, we choose zero prior $F^{k_0=0} = \left\{f_{\{p,I\}}^{k_0=0} = 0,\ f_{\{V,E\}}^{k_0=0} = 0,\ f_{\{p,E\}}^{k_0=0} = 0,\ f_{\{V,D\}}^{k_0=0} = 0,\ U_{\{E,M\}}^{k_0=0} = 0, U_{\{E,T\}}^{k_0=0} = 0\right\}$ in the benchmark. In each round m of the threshold value updating, k_m^* denotes the required updating times for traders' expectation function set to reach financial market equilibrium (i.e., the times required to meet the convergence criteria). Accordingly, the initial function set, $F^{k_{m+1}=0}$, in the $(m+1) - th$ round of the threshold value updating is determined by

$$F^{k_{m+1}=0} = F^{k_m=k_m^*}.$$

This procedure is illustrated in the upper panel in Figure B1. Denote $m^*$the number of rounds to reach endogenous liquidity provision equilibrium. Hence, the total number of the updating rounds is given by

$$\sum_{m=0}^{m^*-1} k_m^*$$

In the equilibrium, we have the following equilibrium function and cut-off strategy sets

$$F^* = F^{k^*_{m^*-1}}, \Theta^* = \Theta^{m^*}$$

***<u>Discretization</u>*:** To apply reinforcement learning technique for updating purpose, the continuous expectation and utility functions, described by function set F^{k_m} , are approximated by piecewise linear functions. Discretising the state space of the economy allows to explicitly compute investors' posterior beliefs and to solve investors' first-order conditions and the market-clearing condition. This also allows for arbitrary price and

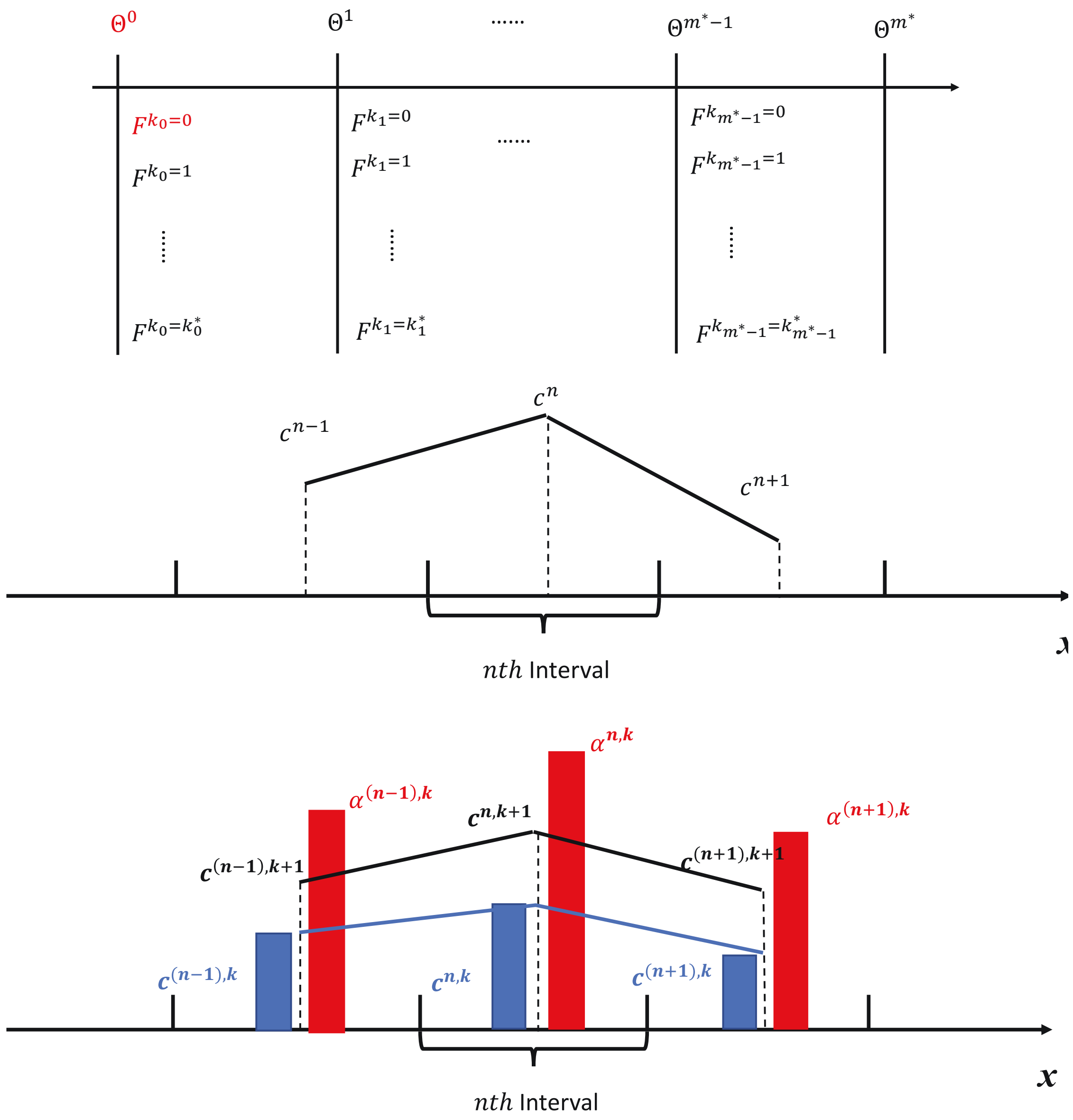


Figure B1: Intuitions for procedure, ddiscretization and updating. This figure illustrate the intuition of a general procedure to generate nonlinear equilibrium (the upper panel), to discrete the expectation and utility functions (the middle panel), and the updating of these functions based on the simulation bootstrap observations (the bottom panel), while the details about how to generate observations are introduced in Appendix B2.

demand functions, that is, there is no need to parameterize (conjecture) these functions. The price we have to pay is that one can only compute the equilibrium for points within the discretization space (though one can make the space between grid points arbitrarily narrow or rely on the interpolation between grind points). We divide the value of independent variable x into N intervals: $N-2$ intervals from $-H$ to H and two additional tail intervals, $(-\infty,-H)$ and $[H,\infty)$. The length of each interval between $-H$ and H is L. We use $n=1,\dots,N$ to indicate different intervals; specifically, $n=1,N$ indicate the left and right tail intervals, respectively, while $n=2,\dots,N-1$ indicate the $n-th$ interval (equivalently $(n-1)-th$ interval starting from $-H$) in between the tail intervals and hence $2H=(N-2)L$.

To approximate the continuous expectation and utility functions, we use piecewise linear functions. For each expectation function $f_h^{k_m}(x)$ or utility function $U_l^{k_m}(x)$, within two tail intervals, the function value is a constant. Within $(N-2)$ intervals between $[-H,H)$, we focus on the value c_h^{n,k_m} (or c_l^{n,k_m}) of the mid-point of each interval for approximating and updating. The disadvantage of this approach is that one can only compute the equilibrium for points within the discretization space (though one can make the space between grid points arbitrarily narrow or rely on the interpolation between grind points). However, due to the constrain of computation power, it is difficult to set arbitrarily narrow space. To overcome this challenge, we use linear functions based on the value of mid-point to improve the estimation results. To illustrate, we drop the subscript k_m, h and l to simplify the expression. Within each interval, the function value is determined through a linear approximation based on two nearest mid-point values. Specifically,

$$f(x)=\begin{cases} c^1 & , \ if\ x<-H; \\ c^{n-1}+(c^n-c^{n-1})\left(\dfrac{x+H}{L}-\left(n-\dfrac{5}{2}\right)\right), & if-H+(n-2)L\le x< -H+\left(n-\dfrac{3}{2}\right)L, n=2,\dots,N-1; \\ c^n+(c^{n+1}-c^n)\left(\dfrac{x+H}{L}-\left(n-\dfrac{3}{2}\right)\right), & if-H+\left(n-\dfrac{3}{2}\right)L\le x< -H+(n-1)L, n=2,\dots,N-1; \\ c^N & , \ if\ x\ge\ H. \end{cases}$$

The middle panel in Appendix Figure 1 illustrates the intuition how to discretise expectation and utility functions and use piecewise linear function to approach continuous functions. Note that for certain expectation function, such as $f_{\{V,E\}}$ in the benchmark model, the state space is two dimensions. However, the intuition and procedure remain the same.

<u>Inference</u>: When updating functions, we use observations from bootstrap to update the function and cut-off strategy sets. In each round of updating, we generate a total number T of observations. The details about how to generate the observations is introduced below in Appendix B2. The observations in the k_m-th simulation in round m of the input variable, x, and updating variable, y, are denoted by

$$\{x_i^{k_m}, y_i^{k_m}\}\big|_{i=1,\dots,T}$$

To simplify the expression, we drop the subscript k_m. Similar to the expectation and utility functions, we group the observations into different interval and summarize the average of the updating values in the k_m-th simulation over the $n-th$ interval by

$$\alpha_{h/l}^{n,k_m} = \begin{cases} \dfrac{\sum_{i=1}^{T} \mathbf{1}_{x_i < -H} \times y_i}{\sum_{i=1}^{T} \mathbf{1}_{x_i < -H}}, & n = 1 \\ \dfrac{\sum_{i=1}^{T} \mathbf{1}_{-H+(n-2)L \le x_i < -H+(n-1)L} \times y_i}{\sum_{i=1}^{T} \mathbf{1}_{-H+(n-2)L \le x_i < -H+(n-1)L}}, & n = 2, \dots, N-1 \\ \dfrac{\sum_{i=1}^{T} \mathbf{1}_{x_i \ge H} \times y_i}{\sum_{i=1}^{T} \mathbf{1}_{x_i \ge H}}, & n = N \end{cases}$$

Here $\alpha_{h/l}^{n,k_m}$ represents either the expectation function $f_h^{k_m}(x)$ or utility function $U_l^{k_m}(x)$. To further simplify the expression, we drop the subscript m, h and l and consider an expectation function. The expectation function is then updated by

$$c^{n,k+1} = \begin{cases} (1-\beta^{n,k})c^{n,k} + \beta^{n,k}\alpha^{n,k}, & t^{n,k} > 0; \\ c^{n,k}, & t^{n,k} = 0; \end{cases}$$

where the weighting function is defined by

$$\beta^{n,k} = max\left(\frac{t^{n,k}}{\sum_{j=1}^{k} t^{n,j}}, \beta_{min}\right)$$

and $t^{n,k}$ is the frequency count of the input variable x located in the n-th interval,

$$t^{n,k} = \begin{cases} \sum_{i=1}^{T} \mathbf{1}_{x_i < -H}, & n = 1 \\ \sum_{i=1}^{T} \mathbf{1}_{-H+(n-2)L \le x_i < -H+(n-1)L}, & n = 2, \dots, N-1 \\ \sum_{i=1}^{T} \mathbf{1}_{x_i \ge H}, & n = N \end{cases}$$

The bottom panel in Figure B1 illustrates the intuition how to updating the function set according to simulated observations. The intuition behind the weighting function $\beta^{n,k}$ is as follows. The first term measures the reliability of the newest estimates based on the number of the random drawing in history. The second term indicates the lowest value β_{min} for $\beta^{n,k}$, which provides a low bound to the weight on the newly simulated data. This prevents the updating from relying too much on the historical results. For $(-\infty, -H)$ and $[H, \infty)$, the frequency of random drawing that locates in these intervals is lower; hence, we rely more on new simulations. Note that the observations located in $(-\infty, -H)$ and $[H, \infty)$ are fewer, which might cause an imprecise estimation over these intervals.

Convergence: The convergence is tested for determination about financial market (given the cut-off strategy for liquidity provision) and for governing endogenous liquidity provision. Firstly, we discuss the converge test for $F^{k_m^*}$ governing exogenous financial market equilibrium to determine the number of simulations k_m^* in the $m-th$ round of the cut-off threshold value updating. The intuition behind this convergence corresponds to the so called "fixed-function" problem.[88] Under linear setup, equilibrium is obtained through a guess-and-verify procedure[89] by solving a fixed-point problem.[90]

For nonlinear equilibrium, paralleled to the fixed point idea in linear equilibrium, when the belief converges in equilibrium, any additional information provided from additional simulations would be insignificant. This means that, after sufficient number of updates, the change in the belief becomes insignificant. We introduce two convergence criteria. For the first criterion, to measure the overall change of belief, we calculate the update error of beliefs in each interval,

$$\Delta^{n,k} = |c^{n,k-1} - \alpha^{n,k}|.$$

This measures the amount of new information from the updating in each interval. We then calculate frequency-weighted errors for each belief by

$$\delta^k = \frac{1}{\sum_{n=1}^{N} t^{n,k}} \sum_{n=1}^{N} \Delta^{n,k} t^{n,k}.$$

[88] To illustrate the nature of the fixed function problem in nonlinear equilibrium, we first review the fixed point problem in linear Kyle model. In equilibrium traders and market makers are strategic in their trading and pricing, as in "*forecast the forecasts of others*" (see, for example, Foster and Viswanathan (1996)). A trader's trading strategy depends on the trading strategies of the others and market maker's pricing rules, which in turn also affect the trader's trading strategy. This endogenous feedback considerably complicates the inference problem in solving the equilibrium. Under some specific models, the equilibrium structure can be linear, which makes the inference problem easier.

[89] More specific, in Kyle (1985), informed trader's demand function is linear $x = \beta v$. Due to strategic trading and pricing, a market maker sets a linear price $p = \lambda u$, with $\lambda = \beta\sigma^2/(\beta^2\sigma^2 + \sigma_z^2)$. The informed trader's strategy depends on the trading strategies of market maker's pricing rules, which in turn also depend on the informed trader's strategy. When informed trader solves the optimization problem, his trading strategy follows $x = v/2\lambda = (\beta^2\sigma^2 + \sigma_z^2)v/2\beta\sigma^2$, creating a fixed point problem, $(\beta^2\sigma^2 + \sigma_z^2)/\ 2\beta\sigma^2 = \beta$, with a solution of $\beta = \sigma^2/\sigma_z^2$.

[90] However, there are two prominent requirements for this to work: the specific equilibrium structure needs to be known in advance and the filter problem can be solved. In our model, the equilibrium structure is unknown, making the equilibrium inference problem almost impossible via traditional theoretical methodology.

We say the beliefs converge when the new updating error is higher than a minimum error from the history,

$$\delta^{k_m^*} > min\{\delta^j, j = 1,2,\dots,k_m^* - 1\}.$$

This means that the beliefs converge when we cannot incorporate more precise information into the expectation and utility functions. Due to the randomness of model parameter, it is impossible to obtain a zero error. For the second criterion, after each updating, we calculate the correlation between the functions over two consequent updating, i.e., $cov(c^k, c^{k+1})$. In equilibrium, this correlation need to be higher than a threshold value (closer to 1).[91]

Secondly, we discuss the converge test for governing endogenous liquidity provision, i.e., determine the equilibrium cut-off strategy set Θ^*. After reaching the financial market equilibrium $F^{k_m^*}$ given the cut-off strategy set Θ^m, we can calculate the value Θ^{m+1} that makes the ELP indifferent between making and taking liquidity. For the benchmark model, we have

$$\hat{\theta}^{m+1}(\hat{\theta}^m) = \left\{\theta \,\middle|\, U_{\{E,T\}}^{k_m^*}(\theta, \hat{\theta}^m) = U_{\{E,M\}}^{k_m^*}(\theta, \hat{\theta}^m)\right\}$$

This value $\hat{\theta}^{m+1}(\hat{\theta}^m)$ is generated through comparing ELP's realized profit of taking and making liquidity, which depends on the given cut-off strategy $\hat{\theta}^m$ determining the financial market equilibrium. In the endogenous liquidity provision equilibrium, the financial market equilibrium and cut-off strategy are interconnected. Specifically, on the one hand, financial market equilibrium depends on the cut-off strategy; on the other hand, the ELP's liquidity provision decision is determined through comparing ELP's realized profit of taking and making liquidity, which are further influenced by financial market equilibrium. Considering computational limitation, we request that, for three continuous experiments, the values that make the ELP indifferent between making and taking liquidity locate in the same interval of the input parameter of utility functions, i.e., there exists an interval j,

[91] There are two ways to obtain a smaller error and a high correlation: 1) randomly draw a set of random outcomes which is highly consistent with the expectations; and 2) improve the accuracy of expectations. The probability of first way could be reduce significantly by demanding that the minimum errors is not be replaced continuously. In our benchmark experiment, we require that the minimum errors for each expectation and utility function in function set are not be replaced for 50 continuous runs and the correlations are higher than 99%. In the consideration of computational ability and time, in the extension of our model, the requested non-replaced time of minimum errors is reduced to 30 continuous runs.

$$-H + (j-2)L \leq \hat{\theta}^s < -H + (j-1)L, for\ s = m^*, m^* - 1, m^* - 2.$$

B2. Observation Generation and Scenario Comparison

We explore the nonlinear equilibrium in several scenarios (Sections 2.1 to 2.4). As described in Appendix B1, we need to continuously update the function set F^{k_m} with the observations of the economics outcomes from simulation until the F^{k_m} is self-consistent (i.e., $F^{k^*_{m^*}}$). During $k_m + 1$ round updating, we generate observations based on the function set F^{k_m} and cut-off strategy set Θ^m. Specifically, we illustrate how to generate the observations for updating expectation functions, i.e., generating $\{x_i^{k_m}, y_i^{k_m}\}\big|_{i=1,\ldots,T}$ used in Appendix C1. We drop the subscript k_m and m to simplify the expression. The procedure of generating observations for solving nonlinear equilibrium are similar among these scenarios except their differences in specific parameters choices such as noise trading volatility, limited market participation, and number of ELPs in the market as indicated in Table A1. We first describe how to generating observations for the most general case where the number of ELP equals to $N = 2$, limited participation parameter $\tau = 0$, and aggregate liquidity trading with general form σ_z^2. We then link the procedure to the specific scenarios.

To generate the observations for updating purpose, we first generate the realization of state variables and then calculate the demand functions and price functions to determine the equilibrium. Finally, based on the equilibrium results, we form the observations and calculate the different realized profits for ELP liquidity provision decision. Figure B2 summarizes the intuition for generating observations in general.

Simulating state variables: In order to generate the observations, we first simulate the state variables. Based on them, we calculate the inputs for observations and then determine the output for updating purpose. Based on the distributions of 1) the five state variables on the subinnovations about fundamentals: subinnovation observed by both ELPs and informed traders $\tilde{\delta}_0 \sim N(0, \sigma_0^2)$, by informed trader only $\tilde{\delta}_d \sim N(0, \sigma_d^2)$, by both ELPs only $\tilde{\delta}_f^s \sim N(0, \rho\sigma_f^2)$ and by each one of ELPs $\tilde{\delta}_f^1 \sim N\big(0, (1-\rho)\sigma_f^2\big)$ and $\tilde{\delta}_f^2 \sim N\big(0, (1-\rho)\sigma_f^2\big)$; 2) the noises for ELP's private information $\tilde{\varepsilon}_{\theta 1} \sim N(0, \sigma_{\varepsilon,\theta}^2)$ and $\tilde{\varepsilon}_{\theta 2} \sim N(0, \sigma_{\varepsilon,\theta}^2)$, and 3)

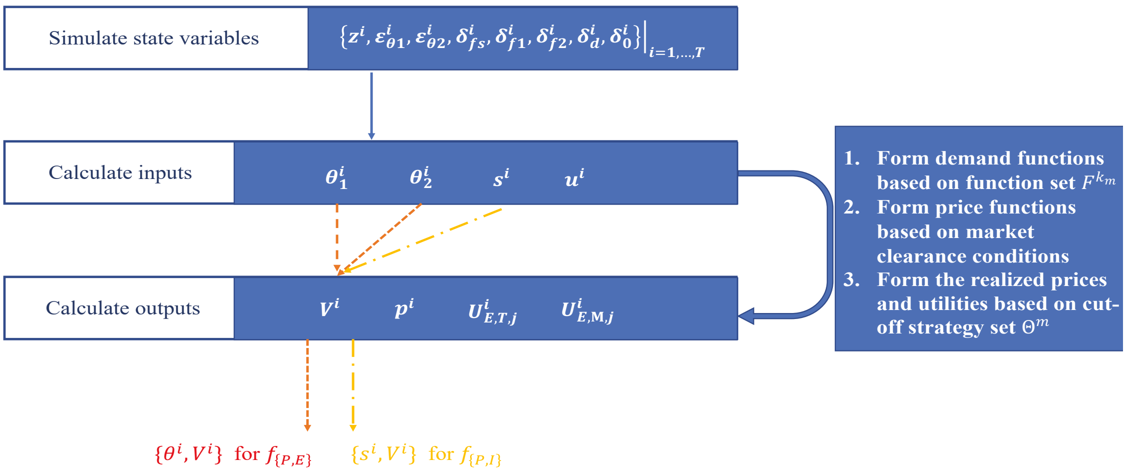


Figure B2: Intuitions for generating observations. This figure illustrates how to generate observations following the four steps in Appendix B1.

the liquidity demands from liquidity trading $\tilde{z} \sim N(0, \sigma_z^2)$, we draw T realizations about these state variables so that each draw forms an realized state variable input set, i.e., $\{z^i, \varepsilon_{\theta 1}^i, \varepsilon_{\theta 2}^i, \delta_{fs}^i, \delta_{f1}^i, \delta_{f2}^i, \delta_d^i, \delta_0^i\}\big|_{i=1,\dots,T}$. We can then calculate the realized fundamental value, $V^i = \delta_0^i + \delta_{fs}^i + \delta_{f1}^i + \delta_{f2}^i + \delta_d^i$, and the information signals for ELPs $\theta_1^i = \delta_0^i + \delta_{fs}^i + \delta_{f1}^i + \varepsilon_{\theta 1}^i$ and $\theta_2^i = \delta_0^i + \delta_{fs}^i + \delta_{f2}^i + \varepsilon_{\theta 2}^i$ for informed trader $s^i = \delta_0^i + \delta_d^i$, respectively, and apply them to calculate the demand functions.

<u>Demand functions</u>: Base on the realizations of state variables formed in the first step, we can calculate the demand function based on the function set. Specifically, market order of informed trader is described as:

$$x_I^i = \gamma\left(s^i - f_{\{P,I\}}(s^i)\right).$$

Market order $x_{E,T,j}^i$ of ELP j is described as

$$x_{E,T,j}^i = \gamma\left(\frac{\sigma_0^2 + \sigma_f^2}{\sigma_0^2 + \sigma_f^2 + \sigma_{\varepsilon,\theta}^2}\theta_j^i - f_{\{P,E\}}(\theta_j^i)\right).$$

Therefore, the aggregate order flow when both ELPs takes liquidity is given by $u_T^i = z^i + x_I^i + x_{E,T,1}^i + x_{E,T,2}^i$, while the limit order of DMM is given by

$$x_D^i = \gamma\left(f_{\{V,D\}}(u_T^i) - p_T^i\right).$$

We next calculate the demand functions under other scenarios. For the scenario with two ELPs, assume ELP j chooses to taking liquidity while another ELP -j chooses to providing liquidity, the aggregate order flow should be $u^i_{TM,j} = z^i + x^i_I + x^i_{E,T,j}$. Hence, the limit order of ELP is described as

$$x^i_{E,TM,-j} = \gamma\left(f_{\{V,1E\}}\left(u^i_{TMj}, \theta^i_{-j}\right) - p^i_{TMj}\right).$$

However, when both ELPs choose to providing liquidity, the aggregate order flow becomes $u^i_M = z^i + x^i_I$. Both ELPs can infer the other one's θ^i_{-j} from the equilibrium price. Therefore, two ELPs forms an identical expectation of fundamental, $f_{\{V,2E\}}\left(u^i_M, \theta^i_1 + \theta^i_2\right)$. Hence, the limit order from DMM can be described as

$$x^i_{E,M,1} = x^i_{E,M,2} = \gamma(f_{\{V,2E\}}\left(u^i_M, \theta^i_1 + \theta^i_2\right) - p^i_M)$$

Since there are only two ELPs existing in the market, the market making ELP j only faces two situations, i.e., either him as the only market maker (the equilibrium price will be $\tilde{p}^1_M\left(\tilde{\theta}_j, \tilde{u}\right)$ under this situation) or both ELPs make market (the equilibrium price will be $\tilde{p}^2_M\left(\tilde{\theta}_j, \tilde{\theta}_{-j}, \tilde{u}\right)$ under this situation). Hence, if

$$\varpi = \tilde{p} - \frac{\tilde{u}}{\gamma} - E\left(\tilde{V} \middle| \tilde{\theta}_j, \tilde{u}, \left|\tilde{\theta}_{-j}\right| > \theta^*\right) \neq 0,$$

both ELPs provide liquidity and learn the other ELP's information from equilibrium price. This means that the ELPs' demand functions is inconsistent with the equilibrium price (use both private information instead).[92]

Price function: Based on the demand functions and aggregate order flows, we can compute the equilibrium price according to market clearance condition. There are actually four different outcomes, depending on whether each ELP takes or makes liquidity. When both ELPs take liquidity, equilibrium price is calculated as

$$p^i_T = \frac{u^i_T}{\gamma} + f_{\{V,D\}}\left(u^i_T\right).$$

When one of ELPs is making liquidity, equilibrium price is calculated as

$$p^i_{TM,j} = \frac{u^i_{TM}}{\gamma} + f_{\{V,1E\}}\left(u^i_{TM,j}, \theta^i_{-j}\right),$$

where $-j$ indicates the ELP providing liquidity. When both ELPs are making liquidity, the equilibrium price is calculated as

[92] Otherwise, when $N \geq 3$, if $\varpi \neq 0$ ELP i would not know whether there are one or more ELPs who are making market except himself. Hence, ELP i could infer other ELPs' signal but only depends on price to form the expectation about fundamental value. This would lead to a no solution situation for reinforcement learning.

$$p_M^i = \frac{u_M^i}{2\gamma} + f_{\{V,2E\}}(u_M^i, \theta_1^i + \theta_2^i).$$

Notice that, in each simulation, there is only one realized price as introduced below, depending on ELP liquidity provision decision which is further decided by the cut-off strategy set Θ and realized signal of ELPs. To calculate ELP's utility functions, we need to calculate price under all situations.

Forming the observations: Finally, we form the observations based on the ELPs' liquidity provision decision depicted by the cut-off strategy set Θ. The equilibrium price actually is determined by

$$p^i = \mathbf{1}_{|\theta_1^i|\le\hat{\theta}_1}\mathbf{1}_{|\theta_2^i|\le\hat{\theta}_2}p_M^i + \mathbf{1}_{|\theta_1^i|\le\hat{\theta}_1}\mathbf{1}_{|\theta_2^i|>\hat{\theta}_2}p_{TM,2}^i + \mathbf{1}_{|\theta_1^i|>\hat{\theta}_1}\mathbf{1}_{|\theta_2^i|\le\hat{\theta}_2}p_{TM,1}^i + \mathbf{1}_{|\theta_1^i|>\hat{\theta}_1}\mathbf{1}_{|\theta_2^i|>\hat{\theta}_2}p_T^i.$$

The utility for taking or making liquidity for ELP j is given by

$$U_{E,T,j}^i = \mathbf{1}_{|\theta_{-j}^i|>\hat{\theta}_{-j}}\left[x_{E,T,j}^i(V^i - p_T^i) - \frac{1}{2\gamma}\left(x_{E,T,j}^i\right)^2\right] + \mathbf{1}_{|\theta_{-j}^i|\le\hat{\theta}_{-j}}\left[x_{E,T,j}^i(V^i - p_{TM,j}^i) - \frac{1}{2\gamma}\left(x_{E,T,j}^i\right)^2\right]$$

$$U_{E,M,j}^i = \mathbf{1}_{|\theta_{-j}^i|>\hat{\theta}_{-j}}\left[x_{E,TM,j}^i(V^i - p_{TM,-j}^i) - \frac{1}{2\gamma}\left(x_{E,TM,j}^i\right)^2\right] + \mathbf{1}_{|\theta_{-j}^i|\le\hat{\theta}_{-j}}\left[x_{E,M,j}^i(V^i - p_M^i) - \frac{1}{2\gamma}\left(x_{E,M,j}^i\right)^2\right]$$

The above definition means that the utility of ELP j who is taking or making liquidity depends on other ELP $-j$ actually liquidity provision decision. To calculate the utilities for ELP j under different liquidity provision decision, there are four possible outcomes for ELP utilities (i.e., each ELP can choose take or make liquidity). Even though there is only one realized state, we need to calculate all of them and generate the observations for the cut-off strategy updating purpose. Finally, after calculating all the outputs, we can form the observations. More specific, using expectation function $f_{\{P,I\}}$ as an example, we get a bunch of combinations about $\{s^i, p^i\}|_{i=1,\ldots,T}$.

We next link this general case to the specific scenarios in the chapter. For the benchmark model, there is only one ELP, the information structure is simplified as equation (1) and the case with $\tau = 1$ is to be introduced below. For circuit breaker scenario, the information structure is also simplified as equation (1) and the first period equilibrium can be delivered through inserting the different liquidity trading parameter; while the details about second period trading can be found in Appendix B3. For two ELPs case, there is no common information component, i.e., subinnovation $\tilde{\delta}_0$ equals to zero.

B3. Supplement Materials

In this part, we provide additional illustrations about both scenarios of $\tau = 1$ for the benchmark and the circuit breakers. As illustrated in Appendix A and Section 2.1, the DMM can infer the equilibrium state when there is only one ELP. Specifically, when $\tau = 1$, DMM can compete with ELP for providing liquidity. Although DMM cannot observe ELP's endogenous liquidity provision decision, DMM can infer the equilibrium regime from the equilibrium price and aggregate order flow, i.e., either him as the only market maker (the equilibrium price will be $\tilde{p}_T(\tilde{u})$ under this situation) or ELP also makes market (the equilibrium price will be $\tilde{p}_M(\tilde{\theta}, \tilde{u})$ under this situation). Hence, if

$$\psi = \tilde{p} - \frac{\tilde{u}}{\gamma} - E(\tilde{V}|\tilde{u}, |\tilde{\theta}| > \theta^*) \neq 0$$

ELP is surely providing liquidity and DMM would learn ELP's information from equilibrium price. This means that, comparing to the scenario of $\tau = 0$, when the ELP provides liquidity, the limit order from the DMM is modified as

$$x_D^i = x_{E,M}^i = \gamma(f_{\{V,E\}}(u_M^i, \theta^i) - p_M^i)$$

This further leads to the change of ELP's utility when providing liquidity as

$$U_{E,M}^i = x_{E,M}^i(V^i - p_M^i) - \frac{1}{2\gamma}(x_{E,M}^i)^2$$

Regarding the circuit breaker scenario, as shown in Appendix B2, the first period equilibrium is similar except different volatility of noise trading. Consider a single discretionary liquidity trader with an exogenous demand $2\tilde{l}$, who may choose to either split her trade equally across two periods or concentrate her trading in a single period. The random variable $\tilde{l}$ is normally distributed with mean zero and variance σ_l^2 and uncorrelated with other random variables. There is also a certain amount of nondiscretionary liquidity trading in each period ($t = 1, 1'$). The net nondiscretionary trade in period t is denoted by z_t, which is identically and independently normally distributed with mean zero and variance σ_z^2 and uncorrelated with other random variables. Therefore, when generating the realization value of state variables, we simulate demand z_l^i for discretionary liquidity trader in addition to the demand for nondiscretionary liquidity trader z_1^i in the first period and z_2^i in the second period. In the following we discuss two situations of the circuit breaker scenario: concentrate or split the order for discretionary liquidity trader.

When liquidity trader decides to concentrate orders in the first period, the trade in the second period is only between DMM and noise trader. The price in the first period neither provides any useful information for DMM nor impacts the cost of liquidity trader in the second period. Therefore, under this situation, we can ignore the computation in the second period. In the first period, the aggregate liquidity trading is describes as $z_1^i + 2z_l^i$. Under this scenario, the loss of discretionary trader who concentrates the trading in the first period can be described as

$$c_c^i = -2z_l^i p_1^i.$$

However, when the liquidity trader decides to split the orders into the first and second periods, the second period equilibrium matters due to its influence on the liquidity trader's loss. In the first period, the aggregate liquidity trading is describes as $z_1^i + z_l^i$; while the second period aggregate order flow (i.e., aggregate liquidity trading) is described as

$$u_{1'}^i = z_{1'}^i + z_l^i$$

Here, we focus on how to generate the second period equilibrium and calculate the liquidity trader's loss. As illustrated in Section 2.3, by observing the equilibrium order imbalance $\tilde{u}_1$ and price $\tilde{p}_1^k$ in period 1, DMM_l knows which equilibrium regime is in the first period (it must be in the taker regime when $\tilde{p}_1^k - \tilde{u}_1/\gamma - f_{V,D}^k(\tilde{u}_1) = 0$ or the maker regime otherwise) and bids based on the order flow $\tilde{u}_2$ in period 2 and equilibrium information in period 1. We highlight two points. First, if the ELP makes liquidity, observing equilibrium order imbalance $\tilde{u}_1$ and price $\tilde{p}_{1,M}^k$ means that $DMM_l 2$ can extract the value of $\tilde{\theta}$ according to Lemma 2.2. Second, the order flow $\tilde{u}_{1'}$ can only be useful when the discretionary liquidity trader chooses to split her trading across the two periods.[93] Specifically, under this situation, $\tilde{u}_{1'}$ can provide information to predict the demand of the discretionary liquidity trader in the first period.

In the second period, DMM_l can infer the equilibrium state in the first period, and extract information from first period equilibrium price. Based on the above analysis, when ELP takes liquidity in the first period, the DMM_l can infer non-liquidity market order in the first period as

$$s^i_{1',T} = u^i_T - \frac{\sigma_l^2}{\sigma_l^2 + \sigma_z^2} u^i_{1'}.$$

Notice that the detailed discussion about this signal has already been given in Section 4. Thus, when ELP takes liquidity, the limit order for DMM_l and the equilibrium price under this condition can be described as

$$x^i_{D1',T} = \gamma\left(f^i_{v,D1'T}\left(s^i_{1',T}\right) - p^i_{1',T}\right);\ p^i_{1',T} = \frac{u^i_{1'}}{\gamma} + f^i_{v,D1'T}(s^{k,i}_{1',T})$$

Similarly, when ELP makes liquidity, observing equilibrium order imbalance $\tilde{u}_1$ and price $\tilde{p}^k_{1,M}$ means that DMM_l can extract the value of $\tilde{\theta}$. Hence, the limit order from DMM 2 depends not only on the updated aggregate order flow but also on the ELP's signal; and the equilibrium price under this condition can be described as

$$x^i_{D1',M} = \gamma\left(f^i_{v,D1'M}\left(\theta^i, s^i_{1',M} = u^i_T - \frac{\sigma_l^2}{\sigma_l^2 + \sigma_z^2} u^i_{1'}\right) - p^i_{1',M}\right);\ p^i_{1',M} = \frac{u^i_{1'}}{\gamma} + f^i_{v,D1'M}(\theta^i, s^i_{1',M})$$

In general, the equilibrium price is determined by the ELP's liquidity provision decision, i.e.,

$$p^i_{1'} = \mathbf{1}_{|\theta^i| \le \hat{\theta}_s} p^i_{1',M} + \mathbf{1}_{|\theta^i| > \hat{\theta}_s} p^i_{1',T}$$

The trading in the first period will go through no matter whether the price in the first period moves outside the bounds or not; however, trading might be halted in the second period with probability $P\left(\left|\tilde{p}^s_1\right| > p^*\right)$. Therefore, the decision of the discretionary trader in the presence of a circuit breaker can be summarized as above. Based on this, the discretional liquidity trader's loss can be calculated as

$$c^i_s = -z^i_l p^i_1 - \mathbf{1}_{|p^i_1| \le p^*} z^i_l p^i_{1'} - \mathbf{1}_{|p^i_1| > p^*} K^*$$

We have now provided all the illustration about the reinforcement learning technique.

[93] If the discretionary liquidity trader chooses to concentrate her trading in periods 1, the order flow in the second period is purely a noise. Therefore, DMM_l would have exactly the same expectation (as either ELP or DMM_e, depending on the equilibrium regime) as in the first period.

Appendix C: Proofs for Chapter 3

C1. Proof for Proposition 3.1

We focus on the linear equilibrium. Based on market maker's pricing functions described in equation (3-4), we first solve the speculator's optimal investment problem. We then characterize the fixed-point problem of the strategic and optimal trading among the traders who are "*forecasting the forecasts of others*". Finally, with the speculator's strategy, we determine the pricing functions of the market makers.

Speculators' demands depend on the information they observe. In forming their demands, they take as given the number of speculators who are fast and slow, the speed that other fast traders choose, the trading strategies of these speculators, and the pricing strategy of the market maker.

Consider then the decision facing speculator *i* who is fast. Because that the payoff is to be announced at the end of trading period and that the speculator's order of x_F^i has probability of q_F to be executed at date 1 and $1-q_F$ at date 2, he expects to purchase at an average price of $q_F p_1 + (1-q_F)p_2$. Expected profits on each unit purchased are then $E\left[v - q_F p_1 - (1-q_F)p_2 \,\middle|\, v, \mathcal{E}_s, \{x_F^j\}\big|_{j\neq i} = x_F^*, \{x_S^j\} = x_S^*\right]$. The expected profit of fast speculator *i* conditional on the realization of *v* is:

$$U_F^i = \mathbb{E}\left\{x_F^i[l_F l_e(v-p_1) + (1-l_F l_e)(v-p_2)] \,\middle|\, v, \mathcal{E}_s, \{x_F^j\}\big|_{j\neq i} = x_F^*, \{x_S^j\} = x_S^*\right\}. \qquad \text{(C-1)}$$

Here, we plug the decomposition of q_F into the equation.

Similarly, consider then the decision facing speculator *j* who is slow. Because that the speculator's order of x_S^j has probability of q_s to be executed at date 1 and $1-q_s$ at date 2, she expects to purchase at an average price of $q_s p_1 + (1 - q_s)p_2$. Expected profits on each unit purchased are then $E\left[v - q_s p_1 - (1 - q_s)p_2 \,\middle|\, v, \mathcal{E}_s, \{x_S^j\}\big|_{j\neq i} = x_S^*, \{x_F^i\} = x_F^*\right]$. By plugging the decomposition of q_s, the expected net profit of slow speculator j conditional on the realization of v is given by

$$U_S^j = \mathbb{E}\left\{x_S^i[l_s l_e(v - p_1) + (1 - l_s l_e)(v - p_2)] \,\middle|\, v, \mathcal{E}_s, \{x_S^j\}\big|_{j\neq i} = x_S^*, \{x_F^i\} = x_F^*\right\}. \tag{C-2}$$

As assumed, both fast and slow speculators know the full information about the fundamental value v. The observed value of v also enables speculator to forecast prices at dates 1 and 2, since he or she knows the realization of v and the trading strategies of the other speculators who have observed v. However, the speculator knows nothing about the order flows generated by liquidity traders, though they have zero mean conditional on v. Thus, if a fast speculator's order is executed at time 1, his expectation of the price at that time becomes

$$\begin{aligned} E\left[p_1 \,\middle|\, v, \mathcal{E}_s, \{x_F^j\}\big|_{j\neq i} = x_F^*, \{x_S^j\} = x_S^*\right] &= \lambda_1 E\left[w_1 \,\middle|\, v, \mathcal{E}_s, \{x_F^j\}\big|_{j\neq i} = x_F^*, \{x_S^j\} = x_S^*\right] \\ &= \lambda_1\left((l_F l_e \mu N - 1)x_F^* + l_s l_e(1-\mu)N x_S^* + x_F^i\right). \end{aligned}$$

where x_F^* and x_S^* denote the conjectured demand of the other fast and slow speculators who have their orders executed at date 1. In contrast, the fast speculator's expectation of the date-2 price (which is independent of whether the order is executed at date 1 or date 2) is given by

$$\begin{aligned} &E\left[p_2 \,\middle|\, v, \mathcal{E}_s, \{x_F^j\}\big|_{j\neq i} = x_F^*, \{x_S^j\} = x_S^*\right] \\ &\quad = \rho_1 E\left[w_1 \,\middle|\, v, \mathcal{E}_s, \{x_F^j\}\big|_{j\neq i} = x_F^*, \{x_S^j\} = x_S^*\right] + \lambda_2 E\left[w_2 \,\middle|\, v, \mathcal{E}_s, \{x_F^j\}\big|_{j\neq i} = x_F^*, \{x_S^j\} = x_S^*\right] \\ &\quad = \rho_1(l_F l_e \mu N x_F^* + l_s l_e(1-\mu)N x_S^*) \\ &\quad + \lambda_2\left(((1 - l_F l_e)\mu N - 1)x_F^* + (1 - l_s l_e)(1-\mu)N x_S^* + x_F^i\right). \end{aligned}$$

Given these expectations, fast speculator i who has observed v chooses x^i to maximize

$$\begin{aligned} U_F^i = x_F^i \Big[& l_F l_e\left(v - \lambda_1\left((l_F l_e \mu N - 1)x_F^* + l_s l_e(1-\mu)N x_S^* + x_F^i\right)\right) \\ & + (1 - l_F l_e)\Big(v - \rho_1(l_F l_e \mu N x_F^* + l_s l_e(1-\mu)N x_S^*) \\ & + \lambda_2\left(((1 - l_F l_e)\mu N - 1)x_F^* + (1 - l_s l_e)(1-\mu)N x_S^* + x_F^i\right)\Big)\Big]. \end{aligned} \tag{C-3}$$

This expression shows clearly how trade by other speculators can have spillover effects on the utility of speculator i. More trade by other informed speculators (no matter fast or slow) lowers speculator i's expected utility. Negative spillovers like these are standard in most information-based asset pricing models. Each agent expects to gain only to the extent that he or she can trade on information that is not yet incorporated into price. If, however, other speculators who are informed trade aggressively, speculator *i* will earn profits if his or her order is executed at date 1. This occurs because additional information is impounded in the price at date-2. Taking the demands of other speculators and the market depth parameters as given, the first-order condition for x_F^* leads to

$$x_F^* = \frac{v - \big((l_F l_e \lambda_1 + (1 - l_F l_e)\rho_1) l_s l_e + (1 - l_F l_e)\lambda_2 (1 - l_s l_e)\big)\,(1-\mu) N x_S^*}{l_F l_e \lambda_1 (l_F l_e \mu N + 1) + (1 - l_F l_e)\rho_1 l_F l_e \mu N + (1 - l_F l_e)\lambda_2 \big((1 - l_F l_e)\mu N + 1\big)} \quad \text{(C-4)}$$

$$= \beta_F v.$$

Equation (C-4) shows that if speculators hold the asset, their demands are "strategic substitutes" with other speculators' demand. As other traders trade more aggressive, not only do speculator-*i*'s expected profits fall (due to the negative spillover), but speculator-*i* also trades less. This is because that more information about *v* is already in the price and so the marginal returns from trading on *v* are lower. We also have the similar results for slow traders, and their demand functions can be represented by

$$x_S^* = \frac{v - \big((l_s l_e \lambda_1 + (1 - l_s l_e)\rho_1) l_F l_e + (1 - l_s l_e)(1 - l_F l_e)\lambda_2\big)\mu N x_F^*}{l_s l_e \lambda_1 (l_s l_e (1-\mu) N + 1) + (1 - l_s l_e)\rho_1 l_s l_e (1-\mu) N + \lambda_2 (1 - l_s l_e)\big((1 - l_s l_e)(1-\mu) N + 1\big)} \quad \text{(C-5)}$$

$$= \beta_S v.$$

The more interesting case is about the influence of exchange speed. The increasing of the exchange speed decreases fast speculator's demand. This is because that the speed advantage of fast speculator decreases, leading to a lower profit. Furthermore, equation (C-4) and (C-5) determine the optimal demands for fast and slow speculators. The speculators and market maker are strategic in their trading and pricing, characterizing "*forecast the forecasts of others*". Indeed, in our model, the speculator's trading strategy depends on the trading strategies of the others and market maker's pricing rules, which in turn affect speculators' trading strategies. This endogenous feedback of "one agent's strategy affects other agents' strategies that affects himself own strategy" can be characterized by a fixed-point problem. We have the best response function for the fast and slow speculators to jointly determine the equilibrium. With the two equations and two parameters, the linear system uniquely determines the market equilibrium.

We then determine market maker's pricing rules based on the speculators' demand functions. At any time, the market makers absorb the net demand between speculators and liquidity traders. Based on the observed order flows, they conjecture the trading strategies of the speculators and form their beliefs about the asset value. Since x_F^* and x_S^* depend on the realized values of v, the order flows provide information about the fundamental value. Given that market makers' priors are normally distributed around a mean of zero, their posterior belief conditional on the order flow w_1 at date-1 is just w_1 multiplied by some constant, λ_1. This constant is equal to the coefficient in a regression of v on w_1. Thus, the price at date 1 (p_1) equals $\lambda_1 w_1$, where

$$\lambda_1 = \frac{Cov(V, w_1)}{Var(w_1)}. \tag{C-6}$$

Plug the aggregate order flow (3-2) in date 1 and speculators' demand function (3-7) into equation (C-6), we have,

$$\lambda_1 = \frac{(l_F l_e \mu N \beta_F + l_s l_e (1-\mu) N \beta_S)\sigma^2}{(l_F l_e \mu N \beta_F + l_s l_e (1-\mu) N \beta_S)^2 \sigma^2 + \sigma_z^2}. \tag{C-7}$$

Similarly, the order flow at date-2 provides information about the value of the asset. Given that the component of the order flow due to speculators' demands is different at dates 1 and 2, and that the variances of liquidity traders' order flow are the same for both periods, the two order flows give different information to market makers, who put unequal weight on these order flows. Thus, the price at date 2 (p_2) equals $\rho_1 p_1 + \lambda_2 w_2$, by applying the projection theorem, the parameter λ_2 follows,

$$\lambda_2 = \frac{Cov(V, w_2|w_1)}{Var(w_2|w_1)}.$$

Since the information about v in the combined order flow is more precise than the information in first period order flow alone, we have $\lambda_1 > \lambda_2$, which means more liquidity in date-2. Of course, the equilibrium value of λ_t depends on the trading strategies of the informed speculators, and these trading strategies, in turn, depend on the way in which the market maker sets prices. Besides, notice that term $(V - E[V|w_1])(w_2 - E[w_2|w_1])$ is independent of order flow w_1[94],

$$Cov(V, w_2|w_1) = E[(V - E[V|w_1])(w_2 - E[w_2|w_1])|w_1] = E[(V - E[V|w_1])(w_2 - E[w_2|w_1])].$$

On the other hand, by plugging the expression of conditional expected form into above covariance equation[95], we have

$$Cov(V, w_2|w_1) = Cov(V, w_2) - \frac{Cov(w_2, w_1)Cov(V, w_1)}{Var(w_1)}.$$

Following the same procedure, we can obtain the form of variance. Then, the pricing rule at date 2 can be represented by

$$\lambda_2 = \frac{Cov(V, w_2)Var(w_1) - Cov(w_2, w_1)Cov(V, w_1)}{Var(w_2)Var(w_1) - Cov(w_2, w_1)^2}. \quad \text{(C-8)}$$

Plugging the date 2 aggregate order flow (3-3) and speculators' demand function (3-7) into equation (C-8), we have,

$$\lambda_2 = \frac{((1 - l_F l_e)\mu N\beta_F + (1 - l_s l_e)(1 - \mu)N\beta_S)\sigma^2}{(l_F l_e \mu^* N\beta_F + l_s l_e(1 - \mu)N\beta_S)^2\sigma^2 + ((1 - l_F l_e)\mu N\beta_F + (1 - l_s l_e)(1 - \mu)N\beta_S)^2\sigma^2 + \sigma_Z^2}. \quad \text{(C-9)}$$

Similarly, we can determine ρ_1 by

$$\rho_1 = \frac{(l_F l_e \mu N\beta_F + l_s l_e(1 - \mu)N\beta_S)\sigma^2}{(l_F l_e \mu N\beta_F + l_s l_e(1 - \mu)N\beta_S)^2\sigma^2 + ((1 - l_F l_e)\mu N\beta_F + (1 - l_s l_e)(1 - \mu)N\beta_S)^2\sigma^2 + \sigma_Z^2} \quad \text{(C-10)}$$

---

[94] According to the definition of covariance,

$$Cov(V - E[V|w_1], w_1) = COV(V, w_1) - COV(E[V|w_1], w_1).$$

We also notice that,

$$COV(V, w_1) = E(Vw_1) - E(V)E(w_1);$$

And,

$$COV(E[V|w_1], w_1) = E(E[V|w_1]w_1) - E[V|w_1]E(w_1).$$

From the measurable and iteration expectation proposition,

$$E(E[V|w_1]w_1) = E(E[Vw_1|w_1]) = E(Vw_1);$$

And,

$$E[V|w_1] = E[V].$$

Thus, we have $Cov(V - E[V|w_1], w_1) = 0$. For the same procedure, we can prove that $Cov(w_2 - E[w_2|w_1], w_1) = 0$.

[95] Plug the conditional expectation's expression into the covariance form,

$$\begin{aligned}
&E[(V - E[V|w_1])(w_2 - E[w_2|w_1])] \\
&= E\left[\left(V - E(V) - \frac{Cov(V, w_1)}{Var(w_1)}(w_1 - E(w_1))\right)\left(w_2 - E(w_2) - \frac{Cov(w_2, w_1)}{Var(w_1)}(w_1 - E(w_1))\right)\right] \\
&= Cov(V, w_2) - \frac{Cov(V, w_1)}{Var(w_1)}Cov(w_1, w_2) - \frac{Cov(w_2, w_1)}{Var(w_1)}Cov(V, w_1) \\
&\quad + \frac{Cov(w_2, w_1)Cov(V, w_1)}{Var(w_1)^2}Var(w_1).
\end{aligned}$$

Solving the above equations (C-4), (C-5), (C-7), (C-9) and (C-10) we have the equilibrium conditions for the financial market equilibrium.

C2. Proof for Corollary 3.1

Based on Definition 3.2 and Proposition 3.1, the value of speed acquisition can be represented as

$$\begin{aligned}\phi(l_F,\mu) = \mathbb{E}\Big\{\mathbb{E}\Big\{&\beta_F v\Big[l_F l_e\big(v-\lambda_1(l_F l_e \mu N\beta_F v + l_s l_e(1-\mu)N\beta_S v)\big)\\
&+(1-l_F l_e)\Big(v-\rho_1(l_F l_e\mu N\beta_F v + l_s l_e(1-\mu)N\beta_S v)\\
&-\lambda_2\big((1-l_F l_e)\mu N\beta_F v + (1-l_s l_e)(1-\mu)N\beta_S v\big)\Big)\Big]\\
&-\beta_S v\Big[l_s l_e\big(v-\lambda_1(l_F l_e\mu N x_F^* + l_s l_e(1-\mu)N x_S^*)\big)\\
&+(1-l_s l_e)\Big(v-\rho_1(l_F l_e\mu N\beta_F v + l_s l_e(1-\mu)N\beta_S v)\\
&-\lambda_2\big((1-l_F l_e)\mu N\beta_F v + (1-l_s l_e)(1-\mu)N\beta_S v\big)\Big)\Big]\,|v,\mathcal{E}_s\Big\}\,|\mathcal{E}_m,\mu,l_F\Big\}.\end{aligned}$$

Simplifying the above expression, we have

$$\begin{aligned}\phi(l_F,\mu) = {}&(\beta_F l_F l_e - \beta_S l_s l_e)(1-\lambda_1\xi_1)\sigma^2\\
&+\big(\beta_F(1-l_F l_e)-\beta_S(1-l_s l_e)\big)(1-\rho_1\xi_1-\lambda_2\xi_2)\sigma^2,\end{aligned} \tag{C-11}$$

Where the trading intensities in early and late period are given as

$$\xi_1 = l_F l_e\mu N\beta_F + l_s l_e(1-\mu)N\beta_S;\ \xi_2 = (1-l_F l_e)\mu N\beta_F + (1-l_s l_e)(1-\mu)N\beta_S.$$

Holding the strategies for fast and slow speculators and market maker constant, a marginal increasing in the fraction of fast speculator influences the value of speed acquisition by

$$\begin{aligned}\frac{\partial\phi(l_F,\mu)}{\partial\mu} = {}&-(\beta_F l_F l_e-\beta_S l_s l_e)\rho_1\sigma^2(l_F l_e N\beta_F - l_s l_e N\beta_S)\\
&-\big(\beta_F(1-l_F l_e)-\beta_S(1-l_s l_e)\big)\lambda_1\sigma^2(l_F l_e N\beta_F - l_s l_e N\beta_S)\\
&-\big(\beta_F(1-l_F l_e)-\beta_S(1-l_s l_e)\big)\lambda_2\sigma^2\big((1-l_F l_e)N\beta_F-(1-l_s l_e)N\beta_S\big)\\
={}&-(\beta_F l_F l_e-\beta_S l_s l_e)^2\lambda_1\sigma^2 N\\
&-\big(\beta_F(1-l_F l_e)-\beta_S(1-l_s l_e)\big)\rho_1\sigma^2(l_F l_e N\beta_F - l_s l_e N\beta_S)\\
&-\big(\beta_F(1-l_F l_e)-\beta_S(1-l_s l_e)\big)^2\lambda_2\sigma^2 N<0.\end{aligned}$$

This partially illustrates that the value of speed acquisition is decreasing with respect to the fraction of fast speculators. Hence, given the speed of fast speculators, the maximum and minimum level of value of speed acquisition are

$$max\, \phi(l_F, \mu) = \phi(l_F, 0); \quad min\, \phi(l_F, \mu) = \phi(l_F, 1).$$

Hence, three types of speed-acquisition equilibrium arise. First, when $C_F + C_V(l_F)$, the cost in speed acquisition, is sufficiently small, all speculators become fast, due to

$$\phi(l_F, 1) > C_F + C_V(l_F).$$

The value from speed investment exceeds the cost being fast. Second, when $C_F + C_V(l_F)$ is sufficiently large, no trader chooses to become fast, due to

$$\phi(l_F, 0) < C_F + C_V(l_F).$$

The cost being fast exceeds the value from speed investment. Finally, when cost in speed acquisition takes an intermediate value, i.e., $\phi(l_F, 1) < C_F + C_V(l_F) < \phi(l_F, 0)$, let

$$g(\mu) = \phi(l_F, \mu) - C_F - C_V(l_F). \tag{C-12}$$

It is intuitive that $\partial g(\mu)/\partial\mu < 0$. Besides, considering that $g(0) > 0$ and $g(1) < 0$; there exists a unique solution for $g(\mu) = 0$, which determines the equilibrium fraction of fast speculators.

C3. Proof for Proposition 3.2

Equation (C-11) provides detailed description about the value of speed investment. When the cost in speed acquisition takes an intermediate value, i.e., $\phi(l_F, 1) < C_F + C_V(l_F) < \phi(l_F, 0)$, let the value of speed investment equals to the cost of speed investment; it implicitly determines the optimal response of μ^* with respect to a given value of l_F, as illustrated in Corollary 3.1. Following, we focus on deriving optimal response of l_F^* for given value of μ.

Considering that traders are strategic, given their trading strategy and take the equilibrium choice of the other fast speculators l_F^* as given, at time $t = 1$, fast speculator i investing in speed technology l_F^i has an expected gross profit (per capita)

$$\begin{aligned}\pi_F\left(l_F^i, l_F^*, \mu, v\right) &= E\left\{\mathbb{x}_F(l_F^*, \mu)\left[l_F^i l_e(v-p_1) + \left(1 - l_F^i l_e\right)(v-p_2)\right] | l_F^*, \mu, v\right\} \\ &= E\left\{\mathbb{x}_F(l_F^*, \mu)\left[l_F^i l_e\left(v\right.\right.\right. \\ &\left.- \lambda_1(l_F^* l_e(\mu N - 1)\mathbb{x}_F(l_F^*, \mu) + l_S l_e(1-\mu)N\mathbb{x}_S(l_F^*, \mu) + \mathbb{x}_F(l_F^*, \mu))\right) \\ &+ \left(1 - l_F^i l_e\right)\left(v - \rho_1(l_F^* l_e(\mu N - 1)\mathbb{x}_F(l_F^*, \mu) + l_S l_e(1-\mu)N\mathbb{x}_S(l_F^*, \mu))\right. \\ &- \lambda_2\left(\left(1 - l_F^i l_e\right)(\mu N - 1)\mathbb{x}_F(l_F^*, \mu) + (1 - l_S l_e)(1-\mu)N\mathbb{x}_S(l_F^*, \mu)\right. \\ &\left.\left.\left.\left.+ \mathbb{x}_F(l_F^*, \mu)\right)\right)\right] | l_F^*, \mu, v\right\},\end{aligned}$$

conditional on his signal. Thus, the *ex-ante* expected gross profit of fast speculators is given by $\bar{\pi}_F(l_F^i, l_F^*, \mu) \stackrel{\text{def}}{=} E[\pi_F(l_F^i, l_F^*, \mu, v)]$. We then compute the net profit (minus speed acquisition cost) for acquiring speed by $R_F(l_F^i, l_F^*, \mu) = \bar{\pi}_F(l_F^i, l_F^*, \mu) - C_F - C_V(l_F^i)$. Given (l_F^*, μ), fast speculator i optimally chooses her own speed acquisition by solving the following first-order condition of $R_F(l_F^i, l_F^*, \mu)$,

$$\frac{\partial R_F(l_F^i, l_F^*, \mu)}{\partial l_F^i} = 0 \Rightarrow \frac{\partial \bar{\pi}_F(l_F^i, l_F^*, \mu)}{\partial l_F^i} = C_V'(l_F^i).$$

Since the above equation holds for every fast speculator, we have $l_F^i = l_F^*$. Therefore, plugging this into the first order condition,

$$\underbrace{\left.\frac{\partial \bar{\pi}_F\left(l_F^i, l_F^*, \mu\right)}{\partial l_F^i}\right|_{l_F^i = l_F^*}}_{MR} = \underbrace{C_V'(l_F^i)\Big|_{l_F^i = l_F^*}}_{MC}. \tag{C-13}$$

Equation (C-13) simply means that the marginal revenue (MR) equals the marginal cost (MC). It implicitly determines the optimal response of l_F^* for given value of μ (and exchange latency l_e), as illustrated in equation (3-10).

So far, we obtain two best response functions as in equations (3-10) and (3-11). Both response functions jointly determine the endogenous equilibrium of information production and speed acquisition, μ^* and l_F^*, by a fixed-point problem. Specifically, based on Definition 3.1 and Corollary 3.1, three types of speed-acquisition equilibrium could arise, i.e., all fast speculator, no speculator and part speculator regimes. Furthermore, as illustrated in Section 3.4, the speed of speculators becomes more concentrated, i.e., faster exchange speed shrinks the market fraction of fast speculators but stimulates up their optimal trading speed. In other words, the fraction of fast speculators decreases with exchange speed. Hence, we denote the thresholds $\hat{l}_e(C_F)$ and $\check{l}_e(C_F)(< \hat{l}_e)$ as

$$\phi\left(l_F(\mu=0,\hat{l}_e(C_F)),\mu=0\right)-C_F-C_V\left(l_F\left(\mu=0,\hat{l}_e(C_F)\right)\right)=0;$$

$$\phi\left(l_F\left(\mu=1,\check{l}_e(C_F)\right),\mu=1\right)-C_F-C_V\left(l_F\left(\mu=1,\check{l}_e(C_F)\right)\right)=0.$$

These thresholds make the fraction of fast speculators to be either zero or one. Therefore, given fixed speed acquisition cost C_F, there exists a unique equilibrium, depending on the exchange latency l_e relative to the thresholds $\hat{l}_e(C_F)$ and $\check{l}_e(C_F)(<\hat{l}_e)$, such that

1) (corner) all speculators stay slow *(with* $\mu^*=0$*)* when $l_e \geq min\{\hat{l}_e(C_F),1\}$;

2) (interior) a fraction μ^* of speculators acquire the speed with the optimal speed level of l_F^* when $max\{\check{l}_e(C_F),0\} \leq l_e < min\{\hat{l}_e(C_F),1\}$, while the equilibrium μ^* and l_F^* are determined by equations (3-10) and (3-11). Specifically, plugging one of the response functions into another one, we get that

$$l_F^* = l_F(\mu(l_F^*;\mathcal{E}_m);\mathcal{E}_m);$$
$$\mu^* = \mu(l_F(\mu^*;\mathcal{E}_m);\mathcal{E}_m).$$

3) (corner) all speculators acquire the speed *(with* $\mu^*=1$*)* when $l_e < max\{\check{l}_e(C_F),0\}$, while the optimal speed level is determined by

$$l_F^* = l_F(\mu=1;\mathcal{E}_m).$$

Besides, the expressions of the functions are

$$l_F^* = \mathbb{l}_F(\mathcal{E}_m) = \begin{cases} l_F(\mu=1;\mathcal{E}_m), \quad C_F \leq \phi(l_F,1)-C_V(l_F) \\ \{l_F^* | l_F^* = l_F(\mu(l_F^*;\mathcal{E}_m);\mathcal{E}_m)\}, \quad max\{\check{l}_e(C_F),0\} \leq l_e < min\{\hat{l}_e(C_F),1\}; \\ no\ definition, \quad l_e < max\{\check{l}_e(C_F),0\} \end{cases}$$

$$\mu^* = \mathbb{u}(\mathcal{E}_m) = \begin{cases} 1, \quad C_F \leq \phi(l_F,1)-C_V(l_F) \\ \{\mu^* | \mu^* = \mu(l_F(\mu^*;\mathcal{E}_m);\mathcal{E}_m)\}, \quad max\{\hat{l}_e(C_F),0\} \leq l_e < min\{\hat{l}_e(C_F),1\}. \\ 0, \quad l_e < max\{\check{l}_e(C_F),0\} \end{cases}$$

Summering all the discussion above, we obtain Proposition 3.2.

C4. Proof for Corollary 3.2

Based on Definition 1 and Corollary 3.1, three types of speed-acquisition equilibrium could arise, i.e., all fast speculator, no speculator and part speculator regimes. Furthermore, Proposition 3.2 divide the equilibrium into three different scenarios based on the value of exchange latency. Given exchange latency l_e, the trading platform's membership fee can be described as

1) (corner) when $l_e \geq min\{\hat{l}_e(C_F), 1\}$, the membership fee depends only on slow speculators' profit, that is $f^* = \psi\bar{\pi}_S(0, le)$. This is because there are only slow speculators existing in the market.

2) (interior) when $max\{\check{l}_e(C_F), 0\} \leq l_e < min\{\hat{l}_e(C_F), 1\}$, the membership fee can be described as $f^* = \psi[\bar{\pi}_F(\mathbb{u}(l_e; \cdot), \mathbb{l}_F(l_e; \cdot), l_e) - C_F - C_V(\mathbb{l}_F(l_e; \cdot))]$. Notice that the membership fee is the same for both fast and slow speculators even under interior solution scenario. This is because that under interior equilibrium, a speculator chooses to become fast only when his expected payoff from the trading is enough to cover the cost of acquiring the speed, compared to be a slow speculator. Indicated by equation (3-11), a speculator needs to be indifferent between being fast and slow.

3) (corner) when $l_e < max\{\check{l}_e(C_F), 0\}$, the membership fee depends only on fast speculators' profit, that is $f^* = \psi[\bar{\pi}_F(1, \mathbb{l}_F(l_e; \cdot), l_e) - C_F - C_V(\mathbb{l}_F(l_e; \cdot))]$. This is because there are only fast speculators existing in the market.

C5. Discussion on Proposition 3.3

As a supplement discussion for Proposition 3.3, we now illustrate how speculators' speed acquisition technology feeds back into exchange speed supply and financial market by conducting comparative statics with respect parameters governing the speculators' fixed speed acquisition cost C_F. The interested variables are a) speculators' speed demand, b) trading platforms' speed supply, c) trading platforms' membership fee, and d) market quality.

The very first results we want to address is the influence on the fraction of fast speculators. We denote $\check{C}_F(\alpha_e)$ and $\hat{C}_F(\alpha_e)$ by

$$\phi(l_F(\mu = 0, \hat{C}_F(\alpha_e)), \mu = 0) - \hat{C}_F(\alpha_e) - C_V(l_F(\mu = 0, \hat{C}_F(\alpha_e)))) = 0.$$

$$\phi(l_F(\mu = 1, \check{C}_F(\alpha_e)\ ), \mu = 1) - \check{C}_F(\alpha_e) - C_V\left(l_F\left(\mu = 1, \check{C}_F(\alpha_e)\right)\right) = 0.$$

These thresholds make the fraction of fast speculators to be either zero or one. The intuitive is straightforward; as the fixed cost of being fast increases, the fraction of fast speculators decreases.

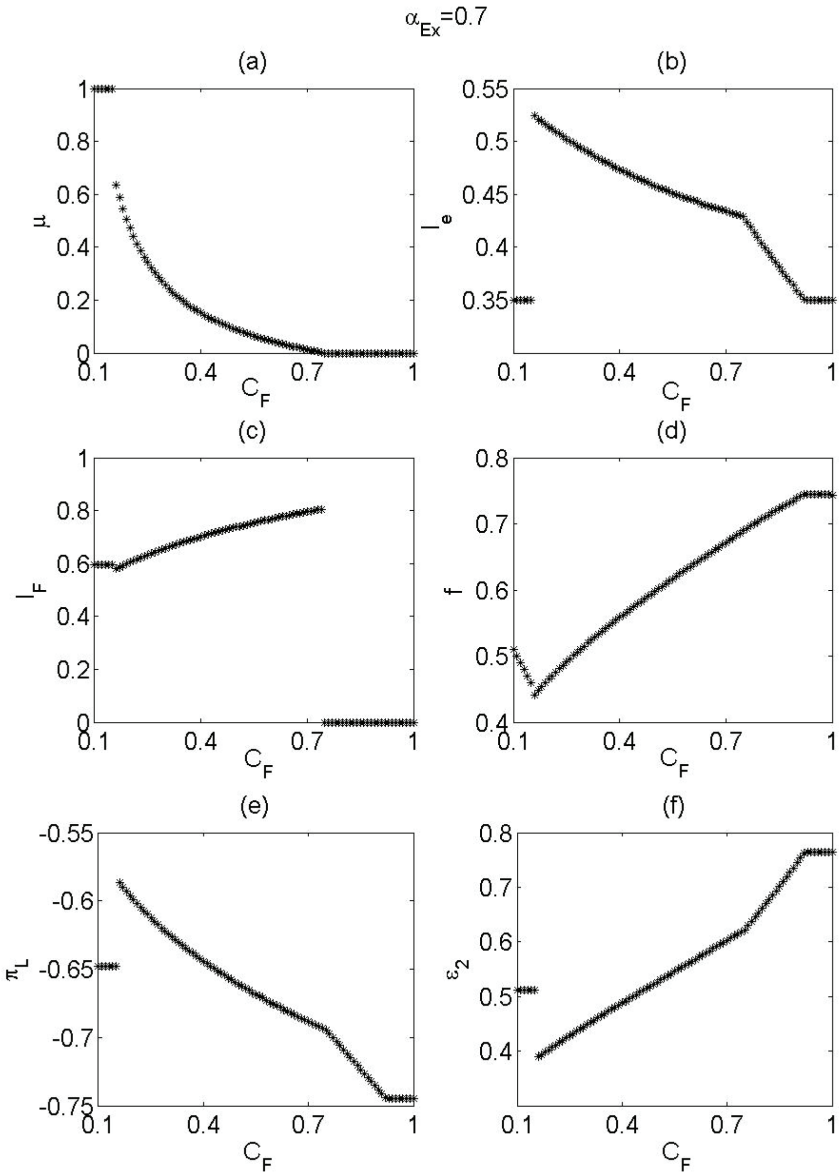


Figure C1. The influence of speculators' fixed speed acquisition cost. This figure plots how the speed demand and supply, membership fee speed, and market quality change with respect to fixed speed acquisition cost. The exchange variable cost function parameter $\alpha_e = 0.7$. The parameter values are the same as in Figure 3-4.

Figure C1 (a) shows that, with relatively low cost (i.e., $C_F < \check{C}_F(\alpha_e)$), all speculators rush to be fast; by contrast, all speculators would stay slow when cost is relatively high (i.e., $C_F \geq \hat{C}_F(\alpha_e)$). With moderate cost level (i.e., $\check{C}_F(\alpha_e) \leq C_F < \hat{C}_F(\alpha_e)$), part of speculators invest in speed technology and fast exchange speed crowds out fraction of fast speculators. In general, faster exchange speed crowds out fast speculators. Therefore, the influences of C_F on optimal exchange speed and optimal agent latency are different for three regimes, respectively.

We next turn to panel (b) and (c) in Figure C1. There are five points emphasized here. Firstly, with relatively low speed cost (i.e., $C_F < \breve{C}_F(\alpha_e)$), the exchange optimally chooses the default speed l_e^0 and fast speculators' optimal agent latency does not depend on the cost levels. From equation (C-13), the equilibrium optimal agent latency, given the optimal exchange latency l_e^0, is decided by

$$\left.\frac{\partial \bar{\pi}_F\left(l_F^i, l_F^*, \mu = 1, l_e = l_e^0\right)}{\partial l_F^i}\right|_{l_F^i = l_F^*} = C_V'(l_F^i)\big|_{l_F^i = l_F^*}. \tag{C-14}$$

Equation (C-14) is not affected by the fixed speed acquisition cost C_F. Therefore, when exchange always optimally chooses the default speed l_e^0 for $C_F < \breve{C}_F(\alpha_e)$, fast speculator sticks to the same agent latency. However, from Figure C1 (d), fast speculators' net profit, hence the trading platform's membership fee, indeed decreases with speed cost C_F.

We then turn to Figure C2 (a) to discuss the determination of exchange speed under this situation. From (3-16), trading platform makes speed investment based on the trade-off between speed acquisition cost and revenue through membership fee. Figure C2 (a) depicts this trade-off. Figure C2 (a) shows that, with relatively low exchange speed (i.e., $l_e < max\{\breve{l}_e(C_F), l_e^0\}$) all speculators rush to be fast; while part of speculators stay slow when $max\,\{\breve{l}_e(C_F), l_e^0\} \leq l_e < min\{\hat{l}_e(C_F), 1\}$. This divides the analysis into two sub-regimes: all-fast-speculator regimes and partial-fast-speculator regimes. When all speculators rush to be fast, the net profit from fast speculators, hence the revenue for exchange, decreases with fast exchange speed. Hence, it is intuitive that exchange chooses the default speed l_e^0 for all-fast-speculator regime. For partial-fast-speculator regime, there exists an optimal exchange speed level $l_e'(C_F)$ satisfying

$$\begin{gathered} l_e'(C_F) \in \underset{le}{argmax}\, N\psi\left[\bar{\pi}_F(\mathbb{u}(l_e;\cdot), \mathbb{l}_F(l_e;\cdot), l_e) - C_F - C_V\left(\mathbb{l}_F(l_e;\cdot)\right)\right] \\ - \alpha_e \times \left[l_e \times \ln\frac{l_e}{l_e^0} + (1 - l_e) \times \ln\frac{1 - l_e}{1 - l_e^0}\right], \\ s.t.\, max\{\breve{l}_e(C_F), l_e^0\} \leq l_e < min\{\hat{l}_e(C_F), 1\}. \end{gathered} \tag{C-15}$$

Therefore, the overall optimal exchange speed l_e^* is one of $\{l_e^0, l_e'(C_F)\}$ that maximise the profit, which can be described by

$$l_e^* = \underset{le\in\{l_e^0, l_e'(C_F)\}}{argmax}\ N\mathbb{f}(l_e;\cdot) - \alpha_e \times \left[l_e \times \ln\frac{l_e}{l_e^0} + (1-l_e)\times \ln\frac{1-l_e}{1-l_e^0}\right]. \tag{C-16}$$

When the cost is relatively low (i.e., $C_F < \breve{C}_F(\alpha_e)$), the optimal profit for exchange under all-fast-speculator regime is higher than the partial-fast-speculator regime; thus, the optimal exchange speed $l_e^* = l_e^0$.

Secondly, from Proposition 3.3, when fixed speed acquisition cost is above certain threshold level $\breve{C}_F(\alpha_e)$, some speculators stay slow. Here, Figure C1 (b) and (c) show the jumps in optimal exchange speed and optimal agent latency. The jump happens due to the optimal exchange speed level switches from the all-fast-speculator regime to the partial-fast-speculator regime. With the increasing of speed cost C_F, the difference between the optimal profit for exchange under the all-fast-speculator regime and the partial-fast-speculator regime $\Delta(C_F)$ decreases, where

$$\Delta(C_F) = N\mathbb{f}(l_e^0;C_F) - \left\{N\mathbb{f}(l_e';C_F) - \alpha_e \times \left[l_e' \times \ln\frac{l_e'}{l_e^0} + (1-l_e')\times \ln\frac{1-l_e'}{1-l_e^0}\right]\right\}. \tag{C-17}$$

The $\Delta(\breve{C}_F(\alpha_e))$ equals to zero at cost level $C_F = \breve{C}_F(\alpha_e)$. This means that the exchange optimally chooses l_e^0 with $C_F < \breve{C}_F(\alpha_e)$ while it optimally chooses $l_e'(C_F)$ with $\breve{C}_F(\alpha_e) \leq C_F < \hat{C}_F(\alpha_e)$. Considering that $l_e^0 < l_e^{'}(C_F)$, this triggers the discontinue change for optimal exchange speed with respect to speed acquisition cost. Besides, it is intuitive that higher exchange speed leads to an optimally low agent latency.

Furthermore, with moderate cost level (i.e., $\breve{C}_F(\alpha_e) \leq C_F < \hat{C}_F(\alpha_e)$), higher speed acquisition cost decreases exchange speed and increases the optimal agent latency. The result for agent latency is similar to the analysis in Section 3.3. However, the optimal exchange speed actually decreases, which shows the speculators' speed acquisition technology feeds back into exchange speed supply through the cross-complementary effect. This endogenous result is opposite to the one in the exogenous equilibrium. In exogenous equilibrium, faster exchange speed crowds out fast speculators, while both exchange speed and fraction of fast speculators decrease with speed cost C_F in endogenous equilibrium. Higher speed cost means that speculators' net profit decreases, which leads to lower exchange profit. Thus, exchange speed shrinks. Besides, the decreasing of both fraction of fast speculators and exchange speed ease the speed competition. Fast speculator, hence, has more incentive to trade faster.

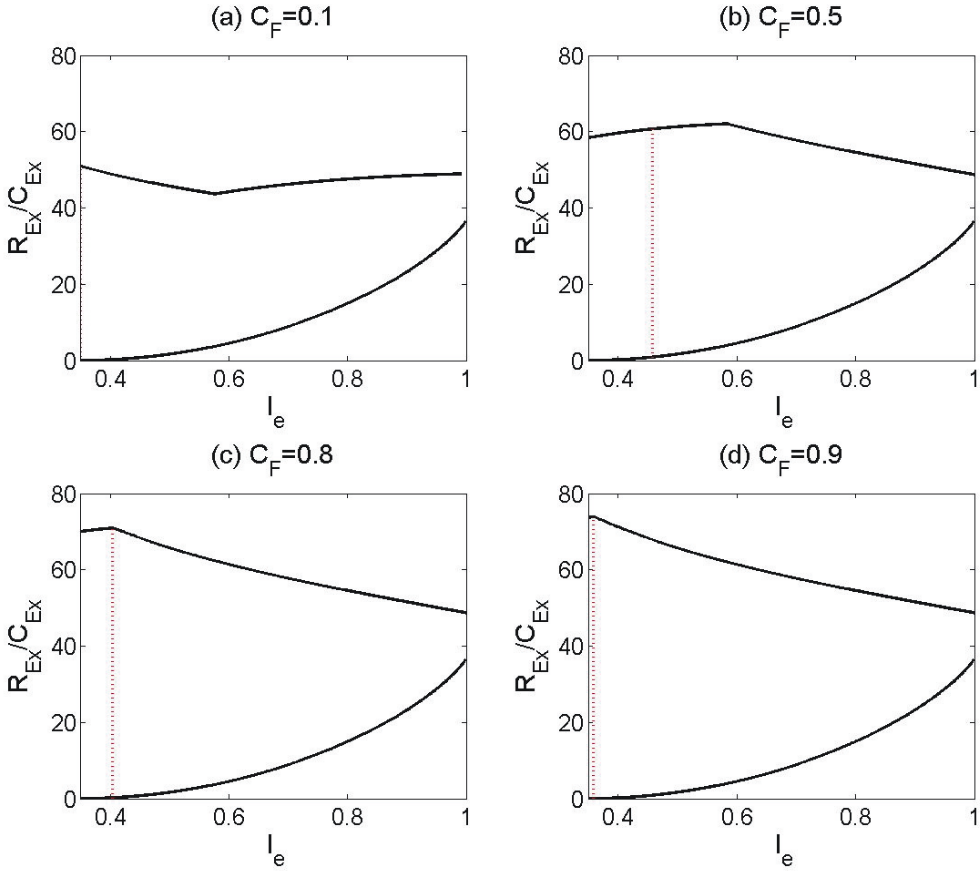


Figure C2. Determination of optimal exchange speed. This figure shows the determination of optimal exchange speed under four different fixed cost situations. The parameter values are the same as in Figure A1.

In addition, as cost C_F increases, the optimal exchange speed $l_e^*\left(\hat{C}_F(\alpha_e)\right)$ at cost level $\hat{C}_F(\alpha_e)$ does not decrease to the default speed l_e^0 eventually. We focus on this level due to all speculators stay slow when the cost is higher than $\hat{C}_F(\alpha_e)$. The details for this analysis are showed in Figure C2 (b) and (c). With moderate exchange speed (i.e., $max\{\breve{l}_e(C_F), l_e^0\} < l_e < min\{\hat{l}_e(C_F), 1\}$) part of speculators rush to be fast; while all speculators stay slow when $l_e \geq min\{\hat{l}_e(C_F), 1\}$. This means that there might be two sub-regimes for the overall optimal exchange speed: no-fast-speculator regimes and partial-fast-speculator regimes. When all speculators stay slow, the net profit from slow speculators, hence the revenue for exchange, decreases with fast exchange speed. Hence, it is intuitive that exchange chooses the threshold speed level $\hat{l}_e(C_F)$ for no-fast-speculator regime. Similarly, according to equation (C-15), for the partial-fast-speculator regime, there is an optimal exchange speed level $l_e^{'}(C_F)$. Therefore, the overall optimal exchange speed l_e^* is one of $\{\hat{l}_e(C_F), l_e^{'}(C_F)\}$ that maximise the profit, which can be described by

$$l_e^* = \underset{le\in\{\hat{l}_e(C_F), l_e^{'}(C_F)\}}{argmax} \; N\mathbb{f}(l_e;\cdot) - \alpha_e \times \left[l_e \times \ln\frac{l_e}{l_e^0} + (1-l_e)\times\ln\frac{1-l_e}{1-l_e^0}\right]. \tag{C-18}$$

When the cost is moderate (i.e., $\check{C}_F(\alpha_e) \le C_F < \hat{C}_F(\alpha_e)$), the optimal profit for exchange under the partial-fast-speculator regime is higher than in the no-fast-speculator regime; thus, the optimal exchange speed $l_e^* = l_e'(C_F)$. However, both optimal exchange speed level under the partial-fast-speculator regime $l_e^{'}(C_F)$ and threshold value $\hat{l}_e(C_F)$ decreases with speed acquisition cost C_F; while $\hat{l}_e(C_F)$ decreases with higher rate. At the cost level $C_F = \hat{C}_F(\alpha_e)$, these two values equal $\hat{l}_e(\hat{C}_F(\alpha_e)) = l_e^{'}(\hat{C}_F(\alpha_e)) > l_e^0$. Therefore, the optimal exchange speed $l_e^* = \hat{l}_e(\hat{C}_F(\alpha_e)) = l_e^{'}(\hat{C}_F(\alpha_e))$ is higher than the default speed l_e^0 at cost level $C_F = \hat{C}_F(\alpha_e)$ (when all speculators stay slow).

Finally, when the speed acquisition cost is too high (i.e., $C_F \ge \hat{C}_F(\alpha_e)$), all speculators stay slow. Under this situation, exchange speed first decreases (Figure C2 (c)) and then exchange makes no speed investment and stays in the default speed l_e^0 (Figure C2 (d)). Under this situation, there is no fast speculators and no investment in reduce agent latency. The decreasing rate is higher than the decreasing rate with moderate cost level (i.e., $\check{C}_F(\alpha_e) \le C_F < \hat{C}_F(\alpha_e)$). This is because that under this situation the optimal exchange speed $l_e^* = \hat{l}_e(C_F)$, while the threshold value $\hat{l}_e(C_F)$ decreases faster than the optimal exchange speed level under the partial-fast-speculator regime $l_e^{'}(C_F)$. When the cost is higher than certain level $C_F = C_F^{'}(\alpha_e)$ such that $\hat{l}_e\left(C_F^{'}(\alpha_e)\right) = l_e^0$, the threshold value $\hat{l}_e(C_F)$ is lower than the default speed l_e^0. Hence, exchange then stays at the default exchange speed l_e^0. We are especially interested in the situation where $\hat{C}_F(\alpha_e) < C_F < C_F'(\alpha_e)$; because exchange still invests in reducing exchange latency even though there is no fast speculators. Although this requires trading platforms' speed investment, it also increases membership fees. Hence, exchange endogenously speeds up and no speculator rushes to fast. Consistent with the findings about speed demand and speed supply, the results about membership fee and market quality are straightforward.

C6. Proof for Corollary 3.5

Taking a total differentiation of equations (3-10) and (3-11) with respect to l_e, we obtain

$$\frac{d\Phi_i}{dl_e} = \frac{\partial h_i}{\partial l_e} + \frac{\partial h_i}{\partial \Phi_j}\frac{\partial h_j}{\partial l_e}; \tag{C-19}$$

$$\frac{d\Phi_j}{dl_e} = \frac{\partial h_j}{\partial l_e} + \frac{\partial h_j}{\partial \Phi_i}\frac{\partial h_i}{\partial l_e}. \tag{C-20}$$

Substituting equation (C-20) into (C-19), we have

$$\frac{d\Phi_i}{dl_e} = \frac{\partial h_i}{\partial l_e} + \frac{\partial h_i}{\partial \Phi_j}\left(\frac{\partial h_j}{\partial l_e} + \frac{\partial h_j}{\partial \Phi_i}\frac{\partial h_i}{\partial l_e}\right).$$

Solving the above equation, we obtain Corollary 3.5.

C7. Discussion on Market Quality

Liquidity is reflected by the price impacts as indicated by (C-7) and (C-9). We define the short and long run trading intensities ξ_1 and ξ_2, measuring how aggressively early and late traders, respectively, trade on their information about the fundamental, are defined by:

$$\xi_1 = q^*\mu N\beta_F + p(1-\mu)N\beta_S; \qquad \xi_2 = (1-q^*)\mu N\beta_F + (1-p)(1-\mu)N\beta_S. \tag{C-21}$$

Plugging equations (A-21) into (A-7) and (A-9), we can show that

$$\lambda_1 = \frac{\xi_1^2\sigma^2 + \sigma_z^2}{\xi_1\sigma^2}; \qquad \lambda_2 = \frac{(\xi_1^2 + \xi_2^2)\sigma^2 + \sigma_z^2}{\xi_2\sigma^2}.$$

As for price discovery, we use the reciprocal of conditional variances $\epsilon_1 = \left(VAR(V|w_1)\right)^{-1}$ and $\epsilon_2 = \left(VAR(V|w_1, w_2)\right)^{-1}$ to measure the price discovery about the fundamental payoff in the first and second periods, respectively. They indicate how much residual uncertainty the market maker faces after conditioning on the prices. When the market maker learns more about the fundamental value from the order flow information, the market becomes more informational efficient. Following the definition,

$$Var(V|w_1) = Var(V) - \frac{Cov(V, w_1)^2}{Var(w_1)} = \sigma^2 - \frac{\left((q^*\mu N\beta_F + p(1-\mu^*)N\beta_S)\sigma^2\right)^2}{(q^*\mu N\beta_F + p(1-\mu^*)N\beta_S)^2\sigma^2 + \sigma_z^2} \tag{C-22}$$

$$= \frac{\sigma^2\sigma_z^2}{(q^*\mu^* N\beta_F + p(1-\mu^*)N\beta_S)^2\sigma^2 + \sigma_z^2}.$$

For the long run price discovery,

$$Var(V|w_1, w_2) \tag{C-23}$$
$$= Var(V)$$
$$- \frac{Cov(V, w_1)^2 Var(w_2) + Cov(V, w_2)^2 Var(w_1) - 2Cov(V, w_1)Cov(w_2, w_1)Cov(V, w_2)}{Var(w_2)Var(w_1) - Cov(w_2, w_1)^2}.$$

Plugging the aggregate order flows in both periods and speculators' demand function into equation (C-23), we have,

$$Var(V|w_1, w_2) \tag{C-24}$$
$$= \frac{\sigma^2 \sigma_z^2}{(q^* \mu^* N\beta_F + p(1-\mu^*)N\beta_S)^2 \sigma^2 + ((1-q^*)\mu^* N\beta_F + (1-p)(1-\mu^*)N\beta_S)^2 \sigma^2 + \sigma_z^2}$$

Equations (C-22) and (C-24) illustrate the expression

$$\epsilon_1 = \frac{\xi_1^2 \sigma^2 + \sigma_z^2}{\sigma^2 \sigma_z^2}; \qquad \epsilon_2 = \frac{(\xi_1^2 + \xi_2^2)\sigma^2 + \sigma_z^2}{\sigma^2 \sigma_z^2}.$$

Equations (C-11) and (C-13) show that price discovery is positively affected by the trading intensities in both periods. Intuitively, when traders trade more aggressively, they reveal more information to the market maker who then impounds more information into the equilibrium prices. Clearly $Var(V|w_1, w_2) < Var(V|w_1)$, meaning that price discovery in the late period is always better than in the early period.

APPENDIX D: Supplement Materials and Analysis for Chapter 3

In this appendix, we provide additional details and analysis to the chapter. Specifically, we first introduce the institutional backgrounds about speed bumps. We then discuss some results not covered in the chapter, including how excess speed demand arise with advancing in speed technology, how lowering exchange latency feeds back to speculator's speed investment, and how technology influence the speed mismatch.

D1. Institutional Backgrounds about Speed Bumps

In United States and Canadian, equities trading is fragmented across multiple venues. Specifically, there are over 11 lit markets in U.S. and 6 lit markets in Canada. Despite different venues might have different trading rules, such as fee structure and priority rules, many exchanges have now applied so-called "speed bump" policy. A speed bump imposes an intentional delay on the arrival or execution of orders at a market, aiming to protect slow traders and liquidity providers from exposure to the risk mentioned above. For instance, Alpha Exchange was launched in 2008 and was acquired by the TMX Group in 2012. On the September 2015, the trading venue was relaunched as TSX Alpha with several changes, including: a) a randomized speed bump of 1-3 milliseconds for all orders except Post-Only limit orders; b) an inverted maker-taker pricing model; c) traders on Alpha are no longer subject to the Order Protection Rule. As documented in Chen et al. (2017a), institutional investors who require more liquidity than what is displayed on any

Table D1

Design of Speed Bumps in Financial Exchanges

Exchange	Date	Targets of delay	Length of delay
IEX	Oct-13	All	350 microseconds
NYSE American	Jul-17	All	350 microseconds
Thomson Reuters	Jun-16	All but cancel	0-3 milliseconds
Aequitas NEO	Mar-15	Liquidity takers	3-9 milliseconds
TSX Alpha	Sep-15	Liquidity takers	1-3 milliseconds
Eurex Exchange	Jun-19	Liquidity takers	1 or 3 milliseconds
Chicago Stock Exchange	Proposed	Liquidity takers	350 microseconds
NASDAQ OMX PHLX	Proposed	Liquidity takers	5 microseconds
ICE Futures	Proposed	Liquidity takers	3 microseconds
Interactive Brokers	Proposed	Liquidity takers	10-200 milliseconds

single trading venue typically utilize a SOR to spray marketable orders across multiple trading venues simultaneously, efficiently accessing consolidated quoted depth at the national best bid or offer price (O'Hara and Ye, 2011). Alpha's asymmetric, randomized speed bump for incoming marketable orders but not limit order entries or cancellations potentially enables its liquidity suppliers to observe the first legs of any large SOR spray being executed on other venues and cancel their limit orders to avoid adverse selection costs.

Some other exchange platforms also regulate traders' speed by introducing speed bumps. For example, Investors Exchange adopts a 350-microsecond (symmetric) speed bump on incoming orders and outgoing information from the exchange. Aequitas NEO and TSX Alpha, both Canadian exchanges, apply a few milliseconds of random delay to slow down only liquidity-taking orders (i.e., asymmetric speed bumps). As summarized by Table 1, many other exchanges have also adopted or proposed the adoption of a speed bump. In most cases, a delay is asymmetrically applied to liquidity taking orders, reflecting the primary purpose of the regulation. For research on speed bumps, one can find more details from Aoyagi (2020) (theory) and Chen et al. (2017a) (empirical).

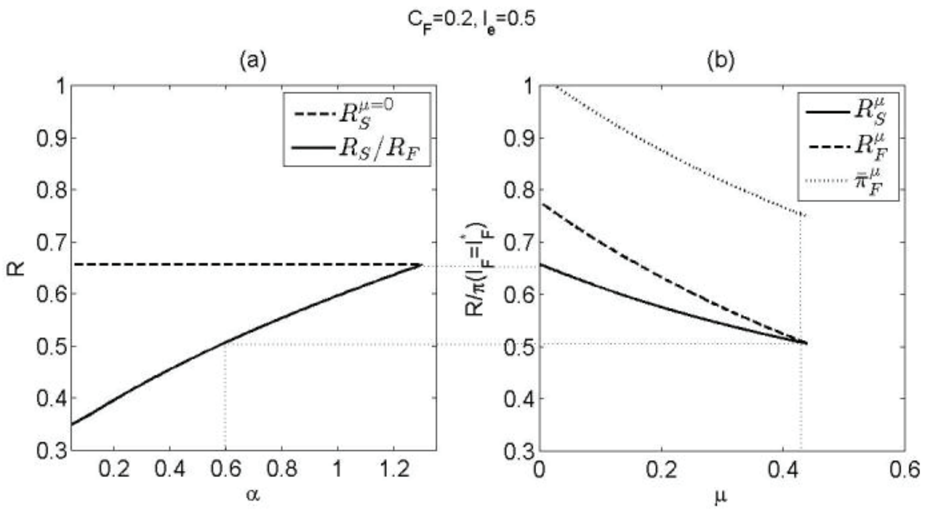


Figure D1: The influence of technology advancement on the gross, net profit and speed acquisition cost. The left panels plot the net profit under both equilibrium and no speed investment scenarios , while the right panels show the evolution of speed over-investment with respect to α. The parameter values are the same as Figure 3-4.

D2. Excessive Speed Demand

In this appendix, we provide further analysis for excessive speed demand. In Section 3.3, we have showed how excessive speed demand changes with respect to advancing in fixed speculator's speed acquisition cost. Here, Figure D1 provides an analysis on how excessive speed demand changes with respect to advancing in varied speculator's speed acquisition cost. In general, the mechanism is very similar to Section 3.3. The endogenous speed acquisition of speculators leads to an excessive speed investment of speculators due to that fast speculators having the first-mover advantages do not internalize the negative externality of their fast trading to other market participants. Essentially, the increasing trading from more fast speculators brings price closer to the fundamental value, which then shrinks the net profit of slow speculators, forcing them to become the fast. In equilibrium, netting the speed acquisition cost, the fast and slow speculators have the same net profit, which is however significantly lower comparing to the situation without fast speculators. Furthermore, the advancing in speculators' speed technology makes it cheaper acquire speed, while intensified competition between fast and slow speculators

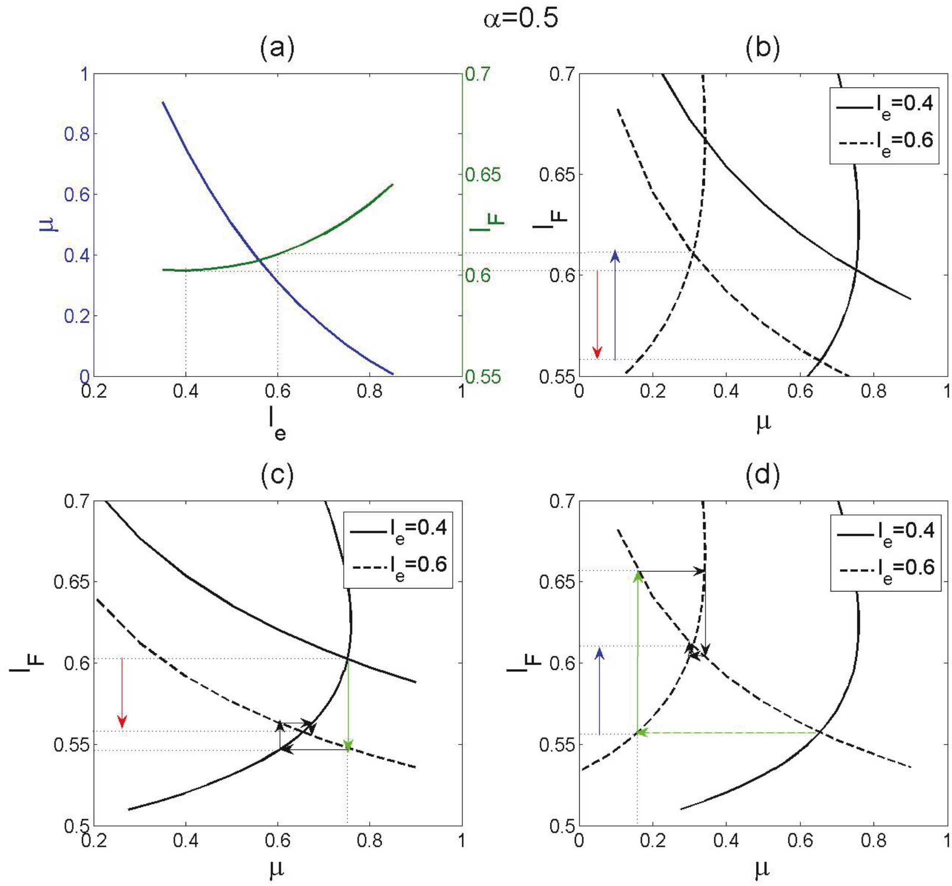


Figure D2: The influence of lowering exchange latency on speed investment. This figure illustrates how an increasing in exchange speed endogenously affects speed acquisition decision. Panel (a) depicts that lowering exchange latency can shrink the fraction of fast speculators but stimulate the optimal speed level. Panel (b) depicts the decomposition for influence of lowering exchange latency on fast speculators' agency latency under compliment scenario. Panels (c) and (d) show the decompositions for direct and indirect effect, respectively. Parameters are the same as in Figure 3-8.

reduce their profit. Therefore, the speed over-investment becomes worse with respect to technology advancing.

D3. Further Analysis for Spill-over Effect

Section 3.4 explains that the speed of speculators becomes more concentrated, i.e., faster exchange speed shrinks the market fraction of fast speculators but stimulates up their

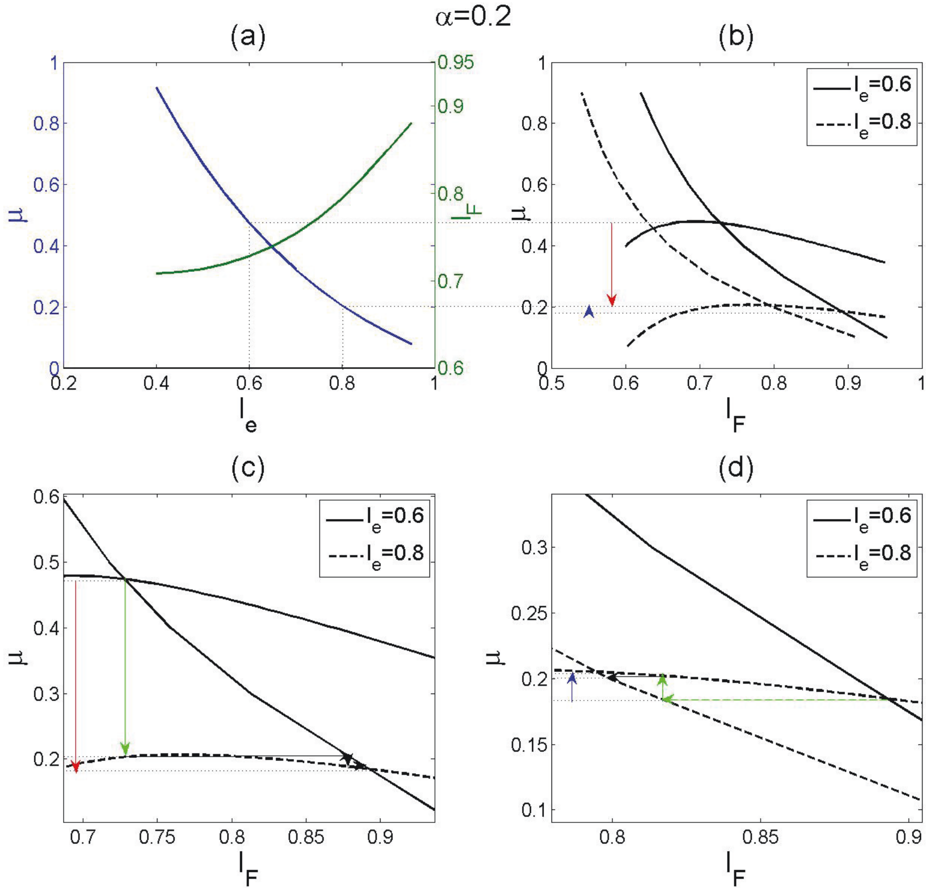


Figure D3: The influence of lowering exchange latency on speed investment. This figure illustrates how an increasing in exchange speed endogenously affects speed acquisition decision. Panel (a) depicts that lowering exchange latency can shrink the fraction of fast speculators but stimulate the optimal speed level. Panel (b) depicts the decomposition for influence of lowering exchange latency on fraction of fast speculators under substitution scenario. Panel (c) and panel (d) show the decompositions for direct and indirect effect, respectively. Parameters are the same as in Figure 3-8.

optimal trading speed. Figure 3-8 illustrates the decomposition of the spill-over effect of faster exchange speed on the fraction of fast speculators under compliment scenario based on Corollary 3.5. From Corollary 3.3, the term $\partial\mu(l_F;l_e)/\partial l_F$ can be either positive (compliment) or negative (substitute). We now discuss the analysis for other three cases (the fraction of fast speculator under substitute scenario, agency latency under both substitute and substitute scenarios) here. Intuitively, an advancing in exchange' speed acquisition technology stimulates trading platform's speed. Based on Corollary 3.4, faster exchange speed then demotes the fraction and speed of fast speculators. More explicitly,

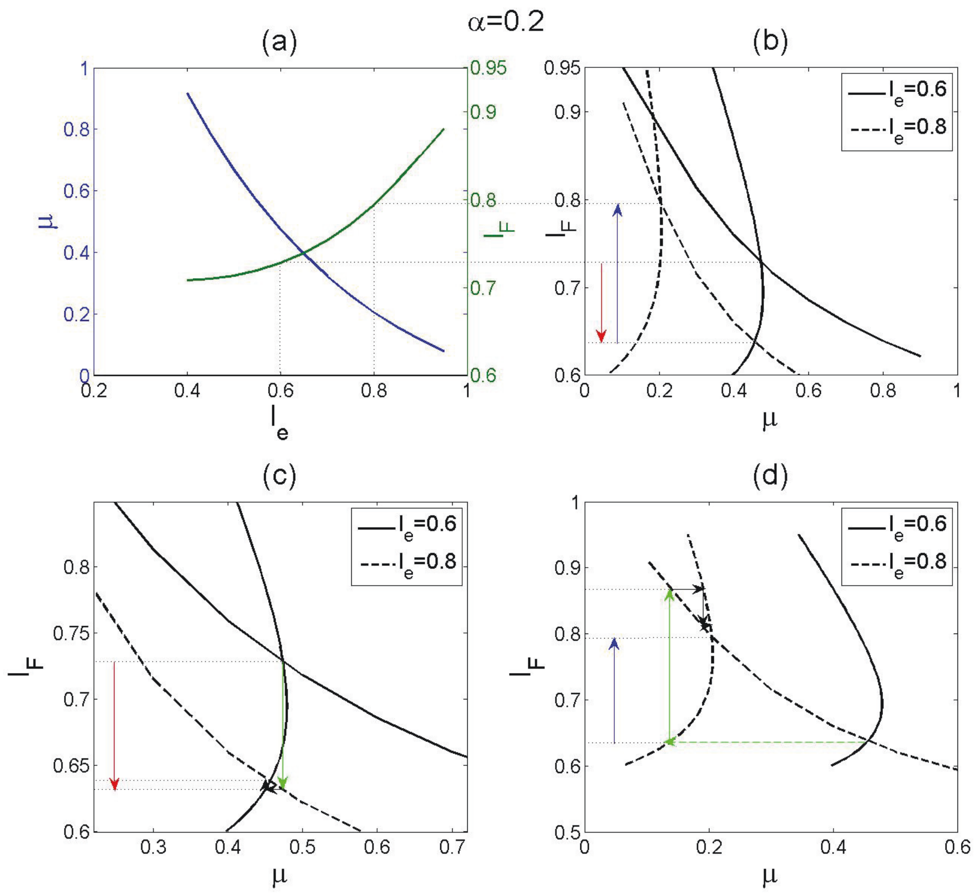


Figure D4: The influence of lowering exchange latency on speed investment. This figure illustrates how an increasing in exchange speed endogenously affects speed acquisition decision. Panel (a) depicts that lowering exchange latency can shrink the fraction of fast speculators but stimulate the optimal speed level. Panel (b) depicts the decomposition for influence of lowering exchange latency on fast speculators' agent latency under substitution scenario. Panel (c) and panel (d) show the decompositions for direct and indirect effect, respectively. Parameters are the same as in Figure 3-8.

Corollary 3.5 shows how the advancing in exchange speed technology spills over to speculator's investment in fast trading technology via the cross effects, supporting the above intuition. Following three figures graphically illustrate this decomposition.

Panels (b), (c) and (d) of Figure D2 illustrate the decomposition of the spill-over effect of faster exchange speed on the speed of fast speculators under compliment scenario based on Corollary 3.5, while Figure D2 (c) and (d) illustrate the direct and indirect effects, respectively, which can be divided into an initial effect and a multiplier effect in (3-19). Consequently, Figure D2 (b) shows the direct effect, indicated by the red arrow, and the indirect effect, indicated by the blue arrow, of lowering exchange latency (as in equation (3-18)). Therefore, the positive indirect effect dominates the negative direct effect, leading to an improvement in speed of fast speculators, as shown in Figure D2 (a).

Panels (b), (c) and (d) of Figure D3 illustrate the decomposition of the spill-over effect of faster exchange speed on the fraction of fast speculators under substitute scenario based on Corollary 3.5. Figure D3 (c) and (d) illustrate the direct and indirect effects, respectively, which can be divided into an initial effect and a multiplier effect in (3-19). Consequently, Figure D3 (b) shows the direct effect, indicated by the red arrow, and the indirect effect, indicated by the blue arrow, of lowering exchange latency (as in equation (3-18)). Therefore, the negative direct effect dominates the positive indirect effect, leading to a crowd effect in fraction of fast speculators, as shown in Figure D3 (a).

Panels (b), (c) and (d) of Figure D4 illustrate the decomposition of the spill-over effect of faster exchange speed on the speed of fast speculators under substitute scenario based on Corollary 3.5. Figure D4 (c) and (d) illustrate the direct and indirect effects, respectively, which can be divided into an initial effect and a multiplier effect in (3-19). Consequently, Figure D4 (b) shows the direct effect, indicated by the red arrow, and the indirect effect, indicated by the blue arrow, of lowering exchange latency (as in equation (3-18)). Therefore, the positive indirect effect dominates the negative direct effect, leading to an improvement in speed of fast speculators, as shown in Figure D4 (a).

In general, the direct effect of lowing exchange latency on the optimal speed level and the fraction of fast speculators is negative as indicated in Corollary 3.4; while the indirect effect, coming from the feedbacks through the interaction, can be either positive or negative for the fraction of fast trader but always positive for the optimal speed level. For the fraction of fast speculators, the negatively direct effect dominates the trade-off; while in contrast, the positively indirect effect dominates for the optimal speed level. Therefore, the speed of speculators becomes more concentrated, i.e., faster exchange speed shrinks the market fraction of fast speculators but stimulates up their optimal trading speed. The results show that speed that was once a ticket to fortunes in financial market is the bare minimum, only the largest trading companies remain in the race for speed.

D4. Comparative Analysis for Policy Maker's Goal

In Section 3.5, Figure 3-10 (b) focuses on the influence of advancing in exchange speed technology; while the influences of speculators' speed technology advancement is provided in Figure D5. Corollary 3.7 depicts that the socially optimal exchange speed weakly decreases with the policy weight ω. This result is also illustrated in Figure D5. Considering moderate weight $0 < \omega < 1$, it can be treated as weight average of the two extreme cases, $\omega = 0$ and $\omega = 1$, which is illustrated in the left panel of Figure D5 (a) and (b). Intuitively, with higher policy weight ω, the regulator cares more about market quality. Therefore, regulator desires a relatively lower speed to ensure high price efficiency. In general, the above results indicate that there is a speed mismatch in equilibrium. For instance, with $\omega = 1$, the regulator desires the default speed so that there is a speed overinvestment for exchange. By contrast, with $\omega = 0$, the regulator desires a moderate speed that is higher than the endogenous optimal exchange speed; so that there is a speed underinvestment for exchange. However, with $0 < \omega < 1$, socially optimal exchange speed decreases from $l_e^{SO}(\omega = 0) > l_e^*$ to default speed l_e^0. Hence, the desire speed from regulator's perspective can be either higher, lower, or equal to the optimal endogenous exchange speed. These results are illustrated in the middle panel of Figure D5 (a) and (b).

Finally, the speed mismatch does not necessarily remain unchanged. With the advancing in technology, both the optimal endogenous exchange speed and regulator's desired speed can be adjusted. The middle panels in Figure D5 (a) and (b) illustrate the results with respect to speculators' fixed and varied cost. Specifically, the influence on the optimal exchange speed has been analyzed in Section 3.4. Besides, for $\omega = 1$, the policy maker focuses only on market efficiency and would always desire the lowest exchange speed (the default exchange latency), i.e., $l_e^{SO} = l_e^0$. Hence, the exchange speed acquisition technology does not affect regulator's desired speed. However, the endogenous speed increases; hence, the speed mismatch (speed overinvestment) has been amplified under this situation.

More interestingly, for $\omega = 0$, the policy maker focuses only on deadweight loss and the desired speed also increases with the advancing in the advancement of speed acquisition technology, i.e., $\partial l_e^{SO}(\omega = 0)/\partial C_F < 0$ and $\partial l_e^{SO}(\omega = 0)/\partial \alpha < 0$. Different from

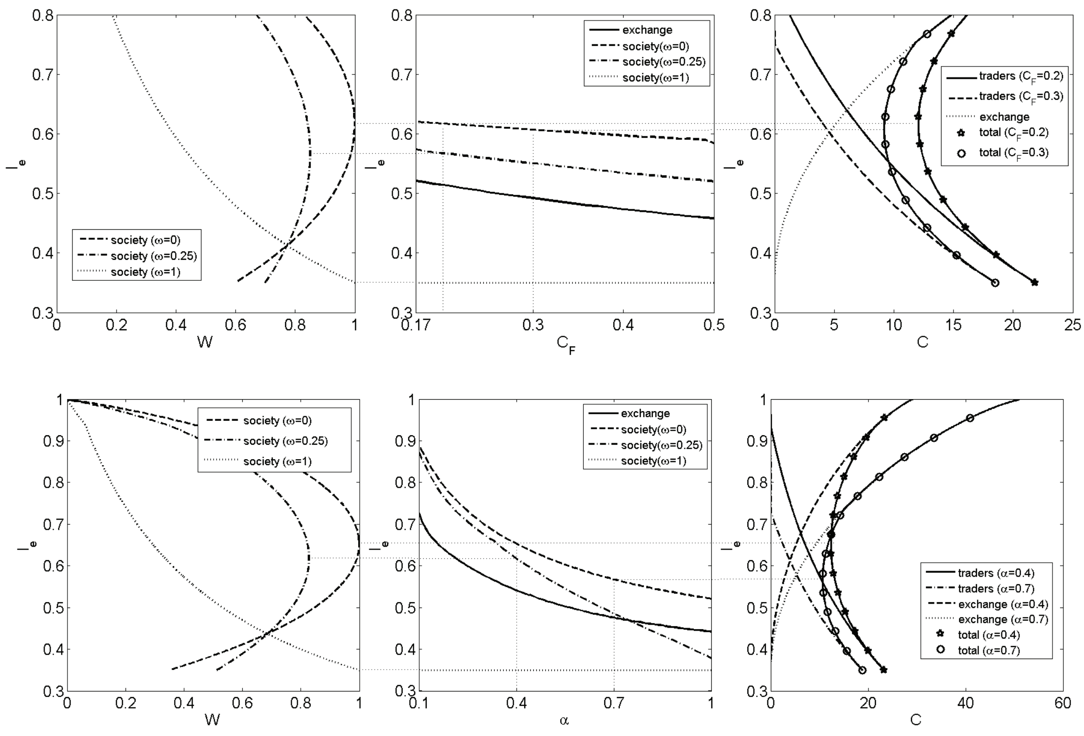


Figure D5: Speed mismatch between the desired speed of policy maker and the optimal speed supplied by the exchange. Panels (a) and (b) depicts how the speed mismatch between the desire exchange speed of policy maker and the optimal speed supplied by the exchange varies with respect to fast trading technology advancement measured by the fixed cost, c_F, the cost multiplier of faster speed of the speculators, α, under different policy weighting $\omega = 0, 0.3, 1$.

Figure 3-10 (b), with advancing in speculators' fixed and varied speed acquisition technology, the investment from speculators becomes relatively more efficient. This intuition has been illustrated in the right panels of Figure D5 (a) and (b). The deadweight loss is jointly determined by the cost of speculators and exchange. With advancing in speculators' fixed and varied speed acquisition technology, the exchange's cost is not influenced, which is indicated by the constant dotted line, while the speculators cost has been reduced. Aggregating the two effect, the deadweight loss that is depicted by the dashed line decreases with more advanced speed technology. Hence, regulator desires higher social optimal speed due to relatively more efficient speculator speed investment.

Appendix E: Proofs for Chapter 4

E1. Proof of Proposition 4.1

The Equilibrium in the Early Period: With the speed hierarchies and the first-mover advantage, fast trader's trading behaviour is not influenced by informed and slow traders. The equilibrium in the early period is thus similar to the standard one-period Kyle (1985) model. The only difference is that the homogenous information structure is replaced by the heterogeneous information structure. Standard procedure delivers price impact and fast traders' trading strategy as

$$\lambda_1 = \frac{\sqrt{I_F N n}}{I_F N + n}\frac{\sigma}{\sigma_z}; \qquad \beta_F = \frac{\sqrt{n}}{\sqrt{I_F N}}\frac{\sigma_z}{\sigma}. \tag{E-1}$$

The Equilibrium in the Late Period: With the partial information, the slow trader can infer the information from the early period equilibrium price. Based on the projection theorem, the slow traders have an expectation $E(V|v_i, p_1) = k_1 v_i + k_2 p_1$ about the fundamental value V with

$$k_1 = \frac{n^2}{(n-\delta)I_F N + n^2}; \qquad k_2 = \frac{(n-\delta)(I_F N + n)}{(n-\delta)I_F N + n^2}.$$

Based on market maker's linear pricing functions, we first solve the speculator's optimal investment problem. We then characterize the fixed-point problem of the strategic and optimal trading among the traders who are "*forecasting the forecasts of others*". Finally, with the speculator's strategy, we determine the pricing functions of the market makers.

Traders' Optimal Strategies: Instead of using (4-6) and (4-7) directly, we assume that the slow and informed trader's trading strategies follow $x_S(v_i, p_1) = l_1 v_i + l_2 p_1$

and $x_I(V,p_1) = h_1 V + h_2 p_1$. Under these assumptions, we have $h_1 = -h_2 = \beta_I$, $l_1 = \beta_S$ and $l_2 = \gamma_S - \beta_S$. Based on the model assumptions, there are $I_I N$ informed traders and $(1 - I_I - I_F)N$ slow informed traders ($(1 - I_I - I_F)N/n$ in each information group). Therefore the equilibrium price in the late period follows

$$p_2 = p_1 + \lambda_2 \left(I_I N x_I + \frac{(1 - I_I - I_F)N \sum_{i=1}^{n} x_S^i}{n} + z_2 \right).$$

Considering the trading strategy for the informed traders, they trade strategically (considering the price impact). Hence, they maximize their expected profits based on their information,

$$\max_x E\left[x_i \left(V - \lambda_2 \left(x_i + (I_I N - 1)x_I + \frac{(1 - I_I - I_F)N \sum_{i=1}^{n} x_S^i}{n} + z_2 \right) - p_1 \right) |V, w_1 \right],$$

conditional on slow informed traders' trading strategy $x_S(v_i, p_1) = l_1 v_i + l_2 p_1$. From the first-order condition, the informed trader's trading strategy satisfies

$$X^I = \frac{n - \delta l_1 \lambda_2 (1 - I_I - I_F)N}{\lambda_2 n (I_I N + 1)} V - \frac{1 + l_2 \lambda_2 (1 - I_I - I_F)N}{\lambda_2 (I_I N + 1)} p_1.$$

Therefore, the informed traders' trading intensity is the best response function of the slow traders' trading intensity,

$$h_1(l_1, l_2) = \frac{n - \delta l_1 \lambda_2 (1 - I_I - I_F)N}{\lambda_2 n (I_I N + 1)}; \qquad h_2(l_1, l_2) = -\frac{1 + l_2 \lambda_2 (1 - I_I - I_F)N}{\lambda_2 (I_I N + 1)}.$$

Each slow trader in group *i* maximizes the expected profits. Similarly, conditional on informed traders' trading strategy, we can calculate the first order condition on the expected profit of slow traders, substitute the informed traders' demand function, and obtain

$$l_1(h_1, h_2) = \frac{n k_1}{\lambda_2 \big((1 - I_I - I_F)N + n \big)} - \frac{n I_I N}{(1 - I_I - I_F)N + n} h_1 k_1 - \frac{(1 - I_I - I_F)N}{(1 - I_I - I_F)N + n} l_1 (\delta k_1 - 1);$$

$$l_2(h_1, h_2) = \frac{n(k_2 - 1)}{\lambda_2 \big((1 - I_I - I_F)N + n \big)} - \frac{n I N}{(1 - I_I - I_F)N + n} (h_1 k_2 + h_2) - \frac{(1 - I_I - I_F)N}{(1 - I_I - I_F)N + n} (\delta l_1 k_2 + (n - 1) l_2).$$

In general, informed traders' trading behaviour is the best response functions of the slow traders' behaviour, and visa verse, jointly determined the equilibrium.

Fixed-point Problem of the Strategic: The speculators and market maker are strategic in their trading and pricing, characterizing "*forecast the forecasts of others*". Indeed, in our model, the speculator's trading strategy depends on the trading strategies of the others and market maker's pricing rules, which in turn affect speculators' trading strategies. This endogenous feedback of "one agent's strategy affects other agents' strategies that affects himself own strategy" can be characterized by a fixed-point problem.

We have the best response function for the informed and slow traders to jointly determine the equilibrium. With the four equations and four parameters, the linear system uniquely determines the market equilibrium. We therefore have the informed and slow informed traders' trading strategies as the functions of fundamental parameters and market maker's pricing rule,

$$h_1 = \frac{n}{\lambda_2\big((I_I N + 1)n + \delta(1 - I_I - I_F)Nk_1\big)}; \qquad h_2 = -\frac{n(I_I N + 1) + nk_2(1 - I_I - I_F)N + \delta(1 - I_I - I_F)Nk_1}{\lambda_2\big((I_I N + 1)n + \delta(1 - I_I - I_F)Nk_1\big)(1 + (1 - I_F)N)};$$

$$l_1 = \frac{nk_1}{\lambda_2\big((I_I N + 1)n + \delta(1 - I_I - I_F)Nk_1\big)}; \qquad l_2 = \frac{n(k_2 - 1)(I_I N + 1) - \delta(1 - I_I - I_F)Nk_1}{\lambda_2\big((I_I N + 1)n + \delta(1 - I_I - I_F)Nk_1\big)(1 + (1 - I_F)N)}.$$

Market Maker's Pricing Rule: The market maker takes the informed and slow traders' trading strategies as given. The aggregate trading volume follows

$$w_2 = I_I N x_I + \frac{(1 - I_I - I_F)N\sum_{i=1}^{n} x_S^i}{n} + z_2.$$

Substituting the expression of the demand functions into the aggregate trading volume, we have

$$w_2 = \frac{I_I N n + \delta(1 - I_I - I_F)Nk_1}{\lambda_2((I_I N + 1)n + \delta(1 - I_I - I_F)Nk_1)}V + (I_I N h_2 + (1 - I_I - I_F)N l_2)p_1 + z_2.$$

In the late period, the market maker observes the aggregate trading volumes in both periods. Applying the projection theorem, we have

$$\begin{aligned}\lambda_2 &= \frac{Cov(V, w_1)Var(w_0) - Cov(w_1, w_0)Cov(V, w_0)}{Var(w_1)Var(w_0) - Cov(w_1, w_0)^2} \\ &= \left(\left(\frac{I_I N n + \delta(1 - I_I - I_F)Nk_1}{\lambda_2\big((I_I N + 1)n + \delta(1 - I_I - I_F)Nk_1\big)}\right)^2 \sigma^2 \right. \\ &\quad \left. + \left(\frac{I_F N}{n} + 1\right)\sigma_z^2\right)^{-1} \frac{I_I N n + \delta(1 - I_I - I_F)Nk_1}{\lambda_2((I_I N + 1)n + \delta(1 - I_I - I_F)Nk_1)}\sigma^2.\end{aligned}$$

Solving the above equations, we have the equilibrium conditions for the late period.

Equilibrium Conditions for the Late Period: In the equilibrium, the price impact and the informed and slow traders' trading strategy follow

$$\lambda_2 = \frac{\sqrt{I_I N n\big((n - \delta)I_F N + n^2\big)^2 + \delta(1 - I_I - I_F)Nn^2\big((n - \delta)I_F N + n^2\big)}}{\big((I_I N + 1)((n - \delta)I_F N + n^2) + \delta(1 - I_I - I_F)Nn\big)\sqrt{I_F N + n}}\frac{\sigma}{\sigma_z}; \tag{E-2}$$

$$\beta_I = \left(\frac{I_I N n}{I_F N + n} + \frac{\delta(1 - I_I - I_F) N n^2}{(I_F N + n)\big((n-\delta) I_F N + n^2\big)}\right)^{-1/2} \frac{\sigma_z}{\sigma}; \tag{E-3}$$

$$\beta_S = \frac{n^2}{(n-\delta) I_F N + n^2} \beta_I; \qquad \gamma_S = \frac{n^2 - \delta n}{(n-\delta) I_F N + n^2} \beta_I. \tag{E-4}$$

Equations (E-1) to (E-4) describe the equilibrium.

E2. Proof of Proposition 4.2

Informed Trading Intensity: To derive the informed traders' best response function, conditional on trading intensities of the fast and slow traders, informed trader maximizes his expected profit,

$$\max_x E\big[x_k\big(V - p_1 - \lambda_2(x_k + (I_I N - 1) X_I + \Phi_S V + \Omega_S p_1 + z_2)\big) | V, w_1\big].$$

Following the first-order condition and Definition 2, we obtain

$$\Phi_I = \underbrace{\frac{I_I N\big((\Phi_F^2 + (\Phi_I + \Phi_S)^2)\sigma^2 + \sigma_z^2\big)}{(\Phi_I + \Phi_S)(I_I N + 1)\sigma^2}}_{\text{direct effect from observing signal}} - \underbrace{\frac{I_I N}{I_I N + 1} \Phi_S}_{\text{indirect effect from inferring slow traders}}. \tag{E-5}$$

$$= \Phi_{I,Direct} - \Phi_{I,Indirect}.$$

Rearranging (D-5) and calculating the quadratic function, we have informed trading intensity

$$\Phi_I = h_I(\Phi_S; \Phi_F, N, n, I_I, I_F, \sigma, \sigma_z) = -\frac{\Phi_S}{2} + \sqrt{\frac{\Phi_S^2}{4} + I_I N \Phi_F^2 + I_I N \frac{\sigma_z^2}{\sigma^2}}. \tag{E-6}$$

To better understand the strategic trading between the informed and slow traders and the trade-off between the direct and indirect effects in (E-5), we substitute (E-6) into (E-5) and obtain (4-10) (in the main text). From (4-10), the direct effect of the informed trading intensity is given by

$$\Phi_{I,Direct}(\Phi_F, \Phi_S) \tag{E-7}$$
$$= \frac{I_I N}{2(I_I N + 1)} \frac{4\Phi_F^2 \sigma^2 + 4\sigma_z^2 + \left(\Phi_S \sigma + \sqrt{(\Phi_S^2 + 4 I_I N \Phi_F^2)\sigma^2 + 4 I_I N \sigma_z^2}\right)^2}{\Phi_S \sigma^2 + \sqrt{(\Phi_S^2 + 4 I_I N \Phi_F^2)\sigma^4 + 4 I_I N \sigma^2 \sigma_z^2}}.$$

Calculating the derivatives of (E-7) with respect to slow traders' trading intensity Φ_S,

$$\frac{\partial \Phi_{I,Direct}(\Phi_F,\Phi_S)}{\partial \Phi_S} = \left[\left(\Phi_S\sigma + \sqrt{\Phi_S^2\sigma^2 + 4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)}\right)^2 - 4(\Phi_F^2\sigma^2 + \sigma_Z^2)\right] G$$

$$= \left[2\Phi_S^2\sigma^2 + 4(I_I N - 1)(\Phi_F^2\sigma^2 + \sigma_Z^2) + 2\Phi_S\sigma\sqrt{\Phi_S^2\sigma^2 + 4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)}\right] G > 0.$$

Here the term G is given by

$$G = \frac{I_I N}{I_I N + 1}\frac{1}{\Phi_S\sigma + \sqrt{\Phi_S^2\sigma^2 + 4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)}}\frac{1}{2\sqrt{\Phi_S^2\sigma^2 + 4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)}} > 0. \quad \text{(E-8)}$$

Therefore, an increase in Φ_S strengthens the direct effect positively. Also, it is easy to show that an increase in Φ_S also strengthens the indirect effect negatively. This implies that the slope of the best response function is jointly determined by these two effects. Overall, we calculate the derivative of (4-10) with respect to slow traders' trading intensity Φ_S,

$$\frac{\partial \Phi_I}{\partial \Phi_S} = G\left[2\Phi_S^2\sigma^4 + 4(I_I N - 1)(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2 + 2\Phi_S\sigma^2\sqrt{\Phi_S^2\sigma^4 + 4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2}\right]$$

$$- G\left(\Phi_S\sigma^2 + \sqrt{\Phi_S^2\sigma^4 + 4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2}\right)2\sqrt{\Phi_S^2\sigma^4 + 4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2}$$

$$= -4G(I_I N + 1)(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2.$$

By substituting (D-8) into the above expression,

$$\frac{\partial \Phi_I}{\partial \Phi_S} = -\frac{2I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2}{4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2}\frac{\sqrt{\Phi_S^2\sigma^4 + 4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2} - \Phi_S\sigma^2}{\sqrt{\Phi_S^2\sigma^4 + 4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2}} = -\frac{\sqrt{\Phi_S^2\sigma^4 + 4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2} - \Phi_S\sigma^2}{2\sqrt{\Phi_S^2\sigma^4 + 4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2}}$$

$$< 0.$$

This shows that the negative and indirect effect dominates the trade-off. In general, the increase in slow traders' trading intensity Φ_S weakens informed traders' trading intensity. By directly calculating the derivative of (E-6), we can obtain the same results about how the slow traders' trading intensity Φ_S affects the informed traders' trading intensity Φ_I.

Similarly, we examine how the fast trading intensity Φ_F affects the informed trading behaviour. Calculating the derivatives about equation (E-7) and by using (E-8), we have

$$\frac{\partial \Phi_{I,Direct}(\Phi_F,\Phi_S)}{\partial \Phi_F} = 4G\Phi_F\left[2\left(\Phi_S\sigma^2 + \sqrt{\Phi_S^2\sigma^4 + 4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2}\right)\sqrt{\Phi_S^2\sigma^4 + 4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2}\right.$$

$$\left. + I_I N\left(\left(\Phi_S\sigma^2 + \sqrt{\Phi_S^2\sigma^4 + 4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2}\right)^2 - 4(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2\right)\right] =$$

$$= 4G\Phi_F\sigma^2(I_I N + 1)\left(\Phi_S\sigma^2 + \sqrt{\Phi_S^2\sigma^4 + 4I_I N(\Phi_F^2\sigma^2 + \sigma_Z^2)\sigma^2}\right)^2 > 0.$$

This leads to the asymmetry reduction effect for the fast trading. Besides, fast traders' trading intensity has no influence on the indirect effect due to no competition. Therefore, informed traders' trading intensity is a complement to fast traders' trading intensity. Similarly, we can directly calculate the derivative of (E-6) to analyze how the fast traders' trading intensity Φ_F can affect the informed traders' trading intensity Φ_I.

Slow Trading Intensity: Given the fast and informed traders' trading intensities, slow trader k in group i maximizes his expected profit,

$$\max_x E\left[x_k\left(V - p_1 - \lambda_2\left(x_k + \left(\frac{(1-I_I-I_F)N}{n} - 1\right)x_S^i + \frac{(1-I_I-I_F)N}{n}\sum_{j\neq i} x_S^j + \Phi_I V + \Omega_I p_1 + z_2\right)\right)\middle| v_i, w_1\right].$$

From the first-order condition, we obtain that

$$\Phi_S = \underbrace{\frac{\delta(1-I_I-I_F)N\left((\Phi_F^2 + (\Phi_I+\Phi_S)^2)\sigma^2 + \sigma_z^2\right)\sigma_z^2}{(\Phi_I+\Phi_S)\sigma^2\left((\delta(1-I_I-I_F)N+n)\sigma_z^2 + (n-\delta)\Phi_F^2\sigma^2\right)}}_{\text{direct effect from observing signal}} \qquad \text{(E-9)}$$
$$- \underbrace{\frac{\delta(1-I_I-I_F)N\sigma_z^2}{(\delta(1-I_I-I_F)N+n)\sigma_z^2 + (n-\delta)\Phi_F^2\sigma^2}\Phi_I}_{\text{indirect effect from inferring informed traders}}$$
$$= \Phi_{S,Direct} - \Phi_{S,Indirect}.$$

Rearranging the best response function (E-9) and calculating the quadratic function, we have

$$\Phi_S = h_S(\Phi_I; \Phi_F, N, n, I_I, I_F, \sigma, \sigma_z) \qquad \text{(E-10)}$$
$$= -\frac{\Phi_I}{2} + \sqrt{\frac{\Phi_I^2}{4} + \delta(1-I_I-I_F)N\frac{\Phi_F^2\sigma^2 + \sigma_z^2}{(n-B)\Phi_F^2\sigma^2 + n\sigma_z^2}\frac{\sigma_z^2}{\sigma^2}}.$$

To examine slow informed trader's trading intensity, we can calculate the derivative of (D-10) with respect to the trading intensity of informed and fast trader respectively and show that slow traders' trading intensity is a substitute to informed traders' trading intensity; but a complement to fast traders' trading intensity.

E3. Lemma E1

We use $\Phi_i = h_i(\Phi_j; \Phi_F, N, n, I_I, I_F, \sigma, \sigma_z)$ to represent the best response functions of the informed (with i=I, j=S) and slow (with i=S, j=I) traders. Let Q be one of the seven exogenous parameters $(N, n, \delta, I_I, I_F, \sigma, \sigma_z)$ that determine Φ_I and Φ_S. Then the effect of Q on the trading intensity is given by

$$\underbrace{\frac{d\Phi_i}{dQ}}_{total\ effect} = \underbrace{M\left(\frac{\partial h_i}{\partial Q} + \frac{\partial h_i}{\partial \Phi_j}\frac{\partial h_j}{\partial Q} + \left(\frac{\partial h_i}{\partial \Phi_F} + \frac{\partial h_i}{\partial \Phi_j}\frac{\partial h_j}{\partial \Phi_F}\right)\frac{d\Phi_F}{dQ}\right)}_{direct\ effect},$$

where the three terms in the bracket capture the direct effects of changing in Q that come from type-i, type-j speculators and the fast traders respectively, while the coefficient M is a multiplier given by

$$M = \left(1 - \frac{\partial h_i}{\partial \Phi_j}\frac{\partial h_j}{\partial \Phi_i}\right)^{-1} \geq 1.$$

Furthermore, M=1 with either no informed traders ($I_I = 0$) *or no slow traders* ($I_I + I_F = 1$)*; otherwise, M>1 and therefore the effect of Q is amplified in the equilibrium.*

Proof: Taking a total differentiation of equations (E-6) and (E-10) with respect to Q, we obtain

$$\frac{d\Phi_i}{dQ} = \frac{\partial h_i}{\partial Q} + \frac{\partial h_i}{\partial \Phi_j}\frac{\partial h_j}{\partial Q} + \frac{\partial h_i}{\partial \Phi_F}\frac{d\Phi_F}{dQ}; \tag{E-11}$$

$$\frac{d\Phi_j}{dQ} = \frac{\partial h_j}{\partial Q} + \frac{\partial h_j}{\partial \Phi_i}\frac{\partial h_i}{\partial Q} + \frac{\partial h_j}{\partial \Phi_F}\frac{d\Phi_F}{dQ}. \tag{E-12}$$

Substituting equation (E-12) into (E-11), we have

$$\frac{d\Phi_i}{dQ} = \frac{\partial h_i}{\partial Q} + \frac{\partial h_i}{\partial \Phi_F}\frac{d\Phi_F}{dQ} + \frac{\partial h_i}{\partial \Phi_j}\left(\frac{\partial h_j}{\partial Q} + \frac{\partial h_j}{\partial \Phi_i}\frac{\partial h_i}{\partial Q} + \frac{\partial h_j}{\partial \Phi_F}\frac{d\Phi_F}{dQ}\right).$$

Solving the above equation, we obtain Lemma E1.

Next, we examine the sign and magnitude of parameter M. When $I_I = 0$ *or* $I_I + I_F = 1$, there is no informed or slow informed trader. Thus, the multiplier of the cross-derivatives $\partial h_i/\partial \Phi_j$ and $\partial h_j/\partial \Phi_i$ equals to zero. When both the informed and slow traders are active in the market, this multiplier depends on whether trading on the two types of speculators is a complement or substitute (the signs of the cross-derivatives $\partial h_i/\partial \Phi_j$ and $\partial h_j/\partial \Phi_i$). When both the informed and slow traders are active in the market, we have from (E-6) and (E-10) that

$$-\frac{1}{2} < \frac{\partial h_I}{\partial \Phi_S} = -\frac{1}{2} + \frac{\sigma \Phi_S}{2\sqrt{\Phi_S^2 \sigma^2 + 4(\Phi_F^2 \sigma^2 + {\sigma_z}^2) I_I N}} < 0;$$

$$-\frac{1}{2} < \frac{\partial h_S}{\partial \Phi_I} = -\frac{1}{2} + \frac{\sigma \Phi_I}{2\sqrt{\Phi_I^2 \sigma^2 + 4\delta(1 - I_I - I_F) N \sigma_z^2 \dfrac{n\Phi_F^2 \sigma^2 + n\sigma_z^2}{(n-B)\Phi_F^2 \sigma^2 + n\sigma_z^2}}} < 0.$$

This means that the two response functions are decreasing with higher trading intensities. Note that they both are greater than -1/2. Therefore $M > 1$. This represents an amplified trading intensity multiplier, meaning that the interaction between the two trading intensity measures tends to amplify the initial effect.

E4. Proof of Corollary 4.2

Informed Trading Intensity: Following Lemma E1, we consider the influence of the fraction of different traders on trading intensity. An increase in I_F (I_I) indicates an increase in the fast (informed) traders and a decrease in the slow traders. On the effect of changing in the fraction of informed traders on the informed traders' trading intensity, we have from Lemma E1 that

$$\frac{d\Phi_I}{dI_I} = M\left(\frac{\partial h_I}{\partial I_I} + \frac{\partial h_I}{\partial \Phi_S}\frac{\partial h_S}{\partial I_I}\right). \tag{E-13}$$

From Proposition 4.2, $\frac{\partial h_I}{\partial \Phi_S} < 0$. In addition, $\frac{\partial h_I}{\partial I_I} > 0$ by the expressions of h_i in equation (E-6), while $\frac{\partial h_S}{\partial I_I} < 0$ by (E-10). These observations lead to the following results on the effect of increasing in I_I. First, conditional on fast trader's trading sensitivity, when the fraction of the informed increases, the informed trader, as a group, trade more aggressively. Second, more informed trading reduces slow trading intensity, which, in turn, increases informed trading intensity due to the strategic substitution. In addition, the effect becomes more significant due to the amplified trading intensity multiplier through the interconnection between the informed and slow trading intensities. Therefore, with more informed traders, they trade more aggressively.

We now examine the effect of changing in the fraction of fast traders on informed traders' trading intensity. Applying Lemma E1, we have

$$\frac{d\Phi_I}{dI_F} = M\left(\underbrace{\left(\frac{\partial h_I}{\partial \Phi_F} + \frac{\partial h_I}{\partial \Phi_S}\frac{\partial h_S}{\partial \Phi_F}\right)\frac{d\Phi_F}{dI_F}}_{\text{direct effect through fast trader}} + \underbrace{\frac{\partial h_I}{\partial \Phi_S}\frac{\partial h_S}{\partial I_F}}_{\text{direct effect through slow informed}}\right); \tag{E-14}$$

In (E-14), the direct effect can be decomposed, respectively, through fast and slow traders. From (E-6) and (E-10), we have $\frac{\partial h_S}{\partial I_F} < 0$ and $\frac{\partial h_I}{\partial \Phi_S} < 0$ (Proposition 4.2). Thus, the direct effect through the slow traders is also positive. Note that $d\Phi_F/dI_F$ is positive. Therefore the sign of (E-14) is determined by the sign of

$$\frac{\partial h_I}{\partial \Phi_F}+\frac{\partial h_I}{\partial \Phi_S}\frac{\partial h_S}{\partial \Phi_F}. \tag{E-15}$$

According to (E-6) and (E-10), we have,

$$\frac{\partial h_I}{\partial \Phi_F}=\frac{2\sigma^2 I_I N\Phi_F}{\sqrt{\Phi_S^2\sigma^4+4I_I N\left(\Phi_F^2\sigma^2+\sigma_z{}^2\right)\sigma^2}}>0;$$

$$\frac{\partial h_S}{\partial \Phi_F}=\frac{2\Phi_F\delta^2\left(1-I_I-I_F\right)N\sigma_z{}^4\sigma^2}{\left((n-\delta)\Phi_F^2\sigma^2+n\sigma_z{}^2\right)^2\sqrt{\Phi_I^2\sigma^4+4\sigma^2\dfrac{(\Phi_F^2\sigma^2+\sigma_z{}^2)\delta\left(1-I_I-I_F\right)N\sigma_z{}^2}{(n-\delta)\Phi_F^2\sigma^2+n\sigma_z{}^2}}}>0.$$

Therefore, term (E-15) can be written as

$$\frac{\partial h_I}{\partial \Phi_F}+\frac{\partial h_I}{\partial \Phi_S}\frac{\partial h_S}{\partial \Phi_F}=\frac{2\sigma^2 I_I NTI_F}{\sqrt{\Phi_S^2\sigma^4+4I_I N(\Phi_F^2\sigma^2+\sigma_z{}^2)\sigma^2}}(1+H), \tag{E-16}$$

where

$$H=\frac{\sigma^2\Phi_S-\sqrt{\Phi_S^2\sigma^2+4(\Phi_F^2\sigma^2+\sigma_z{}^2)I_I N}}{4\sigma^2 I_I N\Phi_F}\frac{2\Phi_F\delta^2(1-I_I-I_F)N\sigma_z{}^4\sigma^2}{\left((n-\delta)\Phi_F^2\sigma^2+n\sigma_z{}^2\right)^2\sqrt{\Phi_I^2\sigma^4+4\sigma^2\dfrac{(\Phi_F^2\sigma^2+\sigma_z{}^2)\delta(1-I_I-I_F)N\sigma_z{}^2}{(n-\delta)\Phi_F^2\sigma^2+n\sigma_z{}^2}}}$$

$$>-\sqrt{4\sigma^2\sigma_z{}^2\frac{I_F N+n}{n}I_I N}\frac{\delta^2(1-I_I-I_F)N}{2I_I N\left((n-\delta)\frac{I_F N}{n}+n\right)^2\sqrt{4\sigma^2\dfrac{\frac{I_F N+n}{n}\delta(1-I_I-I_F)N\sigma_z{}^2}{(n-\delta)\frac{I_F N}{n}+n}}}$$

$$=-\frac{\delta^2\sqrt{1-I_I-I_F}}{2\left((n-\delta)\frac{I_F N}{n}+n\right)\sqrt{(n-\delta)\frac{I_F N}{n}+n}}>-1.$$

Hence, the expression of (E-16) is positive. Therefore, with an increase in the fast traders' fraction, informed traders trade more aggressively.

Slow Trading Intensity: We can express the effect of I_I to slow trading intensity as follows,

$$\frac{d\Phi_S}{dI_I}=M\left(\frac{\partial\Phi_S}{\partial I_I}+\frac{\partial\Phi_S}{\partial\Phi_F}\frac{\partial\Phi_F}{\partial I_I}+\frac{\partial\Phi_S}{\partial\Phi_I}\left(\frac{\partial\Phi_I}{\partial I_I}+\frac{\partial\Phi_I}{\partial\Phi_F}\frac{\partial\Phi_F}{\partial I_I}\right)\right)=M\left(\frac{\partial\Phi_S}{\partial I_I}+\frac{\partial\Phi_S}{\partial\Phi_I}\frac{\partial\Phi_I}{\partial I_I}\right)<0.$$

From Proposition 4.2, the term $\frac{\partial h_S}{\partial \Phi_I} < 0$. Furthermore, by (E-6) and (E-10), we can easily see that $\frac{\partial h_I}{\partial I_I} > 0$ and $\frac{\partial h_S}{\partial I_I} < 0$. Thus more informed traders decrease slow informed trading intensity.

On the fraction of the fast traders on slow informed traders' trading intensity, by directly calculating the derivative based on the closed-form solution, we can show that the increase in I_F reduces slow informed traders' trading intensity Φ_S.

Aggregate Late Period Trading Intensity: From Proposition 4.1, we have

$$\Phi_L = \Phi_I + \Phi_S = \sqrt{\frac{I_F N + n}{n}\left(I_I N + \frac{\delta(1 - I_I - I_F)Nn}{(n-\delta)I_F N + n^2}\right)}\,\frac{\sigma_z}{\sigma}. \tag{E-17}$$

From (E-17), it is easy to show that more informed traders increase the aggregate trading intensity and hence price discovery in late period.

On the impact of the fraction of fast traders on the trading intensity in the late period, we have from (E-17) that

$$\frac{d\Phi_L}{dI_F} = \frac{1}{2n\Phi_L}\left[N\left(I_I N + \frac{\delta(1 - I_I - I_F)Nn}{(n-\delta)I_F N + n^2}\right) + (I_F N + n)\frac{-\delta Nn\big((n-\delta)I_F N + n^2\big) - \delta(1 - I_I - I_F)Nn(n-\delta)N}{\big((n-\delta)I_F N + n^2\big)^2}\right]\frac{\sigma_z^2}{\sigma^2}.$$

It is easy to show that the above equation monotonically increases in the fraction of informed traders. Therefore, we can define the threshold value I_I^a as,

$$I_I^a = \left[N\left(N - \frac{\delta Nn}{(n-\delta)I_F N + n^2}\right) + (I_F N + n)\frac{\delta Nn(n-\delta)N}{\big((n-\delta)I_F N + n^2\big)^2}\right]^{-1}\left[(I_F N + n)\frac{\delta Nn\big((n-\delta)I_F N + n^2\big) + \delta(1 - I_F)Nn(n-\delta)N}{\big((n-\delta)I_F N + n^2\big)^2} - N\frac{\delta(1 - I_F)Nn}{(n-\delta)I_F N + n^2}\right]. \tag{E-18}$$

The aggregate trading intensity in the late period decreases in the fast traders for $0 < I_I < I_I^a$, but increases in fast trader for $I_I^a < I_I < 1 - I_F$.

E5. Proof of Corollary 4.3

To examine the effect of information diffusion, we substitute Proposition 4.1 into Definition 4.2 and obtain a closed-form solution of the informed traders' trading intensity,

$$\Phi_I = \frac{I_I N}{\sqrt{\frac{I_I N n}{I_F N + n} + \frac{\delta(1 - I_I - I_F) N n^2}{(I_F N + n)\left((n-\delta) I_F N + n^2\right)}}} \frac{\sigma_z}{\sigma}. \tag{E-19}$$

Calculating the derivative with respect to information disclosure level,

$$\frac{d\Phi_I}{dn} = -\frac{\Phi_I}{2\left(\frac{I_I N n}{I_F N + n} + \frac{\delta(1 - I_I - I_F) N n^2}{(I_F N + n)\left((n-\delta) I_F N + n^2\right)}\right)} \left[\frac{I_I N I_F N}{(I_F N + n)^2} + \frac{\delta(1 - I_I - I_F) N n[(n-\delta) I_F N I_F N + n(-\delta I_F N - n^2)]}{(I_F N + n)^2 \left((n-\delta) I_F N + n^2\right)^2}\right]. \tag{E-20}$$

The sigh of (E-20) is pinned down by the sign of

$$I_I N I_F N\left((n-\delta) I_F N + n^2\right)^2 + \delta((1 - I_F) N - I_I N) n[(n-\delta) N^2 I_F^2 + n(-\delta I_F N - n^2)]. \tag{E-21}$$

The first term of (E-21) is positive. The second term depends on the level of fast trader, which determines the value of quadratic function $[(n-\delta) N^2 I_F^2 + n(-\delta I_F N - n^2)]$. When the fraction of the fast traders satisfies

$$I_F > I_F^b = \frac{\delta n + \sqrt{\delta^2 n^2 + 4n^3(n-\delta)}}{2(n-\delta) N}, \tag{E-22a}$$

the second term of (E-21) is also positive. Instead, when $I_F < I_F^b$, the second term of (E-21) is negative. Note that in (E-20),

$$\frac{d}{dI_I N}\left[\frac{I_I N I_F N}{(I_F N + n)^2} + \frac{\delta(1 - I_I - I_F) N n[(n-\delta) I_F N I_F N + n(-\delta I_F N - n^2)]}{(I_F N + n)^2 \left((n-\delta) I_F N + n^2\right)^2}\right] > 0.$$

Therefore, we can define a threshold value on the number of informed traders to make (E-21) equal to zero,

$$I_I^b = \frac{-\delta(1 - I_F) N n[(n-\delta)(I_F N)^2 + n(-\delta I_F N - n^2)]}{I_F N\left((n-\delta) I_F N + n^2\right)^2 - \delta n[(n-\delta)(I_F N)^2 + n(-\delta I_F N - n^2)]}. \tag{E-22b}$$

Thus, only when $I_F < I_F^b$ *and* $I_I < I_I^b$, (E-21) is negative, which means that higher disclosure level demotes informed trading intensity.

On slow trading intensity Φ_S,

$$\Phi_S = \frac{\delta(1-I_I-I_F)Nn}{\big((n-\delta)I_FN+n^2\big)\sqrt{\frac{I_INn}{I_FN+n}+\frac{\delta(1-I_I-I_F)Nn^2}{(I_FN+n)\big((n-\delta)I_FN+n^2\big)}}}\frac{\sigma_z}{\sigma}. \quad \text{(E-23)}$$

It is easy to show that higher disclosure level always promotes slow trading. From (E-17), it can be shown that high disclosure level always promotes late trading intensity.

Calculating the derivative of (E-19) and (E-23) with respect to δ, we can show the result on information dissemination.

E6. Proof of Proposition 4.3

We first define the expression in Lemma 4.1.

$$\Psi(I_F,I_I,\theta,\delta) = \frac{n\sqrt{I_FN+n}}{\Big((I_IN+1)\big((n-\delta)I_FN+n^2\big)+\delta(1-I_I-I_F)Nn\Big)\sqrt{I_INn+\frac{\delta(1-I_I-I_F)Nn^2}{(n-\delta)I_FN+n^2}}}\sigma\sigma_z. \quad \text{(E-24)}$$

Substituting $\Psi(I_F,I_I,\theta,\delta)$ into the value of information, we have

$$\Gamma_I(I_F,I_I,\theta,\delta) = \frac{(n-\delta)\sqrt{(I_FN+n)n}}{\Big((I_IN+1)\big((n-\delta)I_FN+n^2\big)+\delta(1-I_I-I_F)Nn\Big)\sqrt{I_INn+\frac{\delta(1-I_I-I_F)Nn^2}{(n-\delta)I_FN+n^2}}}\sigma\sigma_z.$$

Therefore, we have

$$\frac{\partial\Gamma_I(I_F,I_I,n)}{\partial I_F} = \frac{\big(\Gamma_I(I_F,I_I,n)\big)^2}{(n-1)\sigma\sigma_z}H.$$

The sign of $\partial\Gamma_I(I_F,I_I,n)/I_F$ is pinned down by the sign of term H, which follows

$$H = \frac{\partial}{\partial I_F}\left[-\Big((I_IN+1)\big((n-\delta)I_FN+n^2\big)+\delta(1-I_I-I_F)Nn\Big)\sqrt{\frac{I_INn}{I_FN+n}+\frac{\delta(1-I_I-I_F)Nn^2}{\big((n-\delta)I_FN+n^2\big)(I_FN+n)}}\right]. \quad \text{(E-25)}$$

Simplifying (E-25), we have $H = KG(I_I)$, in which

$$K = -\frac{1}{2(I_FN+n)^2\big((n-\delta)I_FN+n^2\big)^2\sqrt{\frac{I_INn}{I_FN+n}+\frac{\delta(1-I_I-I_F)Nn^2}{\big((n-\delta)I_FN+n^2\big)(I_FN+n)}}}$$

is always negative and function G has following expression,

$$\begin{aligned}G(I_I) = {} & 2\big((I_IN+1)(n^2-\delta n)N-\delta Nn^2\big)\Big[I_INn\big((n-\delta)I_FN+n^2\big)^2(I_FN+n) \\ & +\delta(1-I_I-I_F)Nn^2\big((n-\delta)I_FN+n^2\big)(I_FN+n)\Big] \\ & +\Big((I_IN+1)n\big((n-\delta)I_FN+n^2\big) \\ & +\delta(1-I_I-I_F)Nn^2\Big)\Big\{-I_INnN\big((n-\delta)I_FN+n^2\big)^2 \\ & -\delta Nn^2\Big[(I_FN+n)\big((n-\delta)I_FN+n^2\big) \\ & +(1-I_I-I_F)[(n-\delta)I_FN^2+n^2N+(I_FN+n)(n-\delta)N]\Big]\Big\}.\end{aligned} \tag{E-26}$$

We can show that $G''(I_I)>0$. Besides, substituting $I_I=0$ into $G(I_I)$ in (E-26), we have $G(0)<0$. Combining these observations, we see that quadratic equation function G of I_I has a unique and positive solution

$$I_I^c = \{I_I | G(I_I)=0, I_I>0\}. \tag{E-27}$$

Hence, when $0<I_I<I_I^c$, function G is negative; term H in (E-25) is positive. This means that, with more fraction of fast traders, the value of information increases.

E7. Proof of Corollary 4.5

Information dissemination: It is easy to show that faster dissemination rate decreases the value of information. We only examine how faster dissemination rate influences value of speed. Since information dissemination only affects the second period equilibrium, we focus on the slow traders' unconditional expected profit here,

$$\bar{\pi}_S = \frac{n^2\sqrt{I_FN+n}}{(I_FN+n)\left((I_IN+1)\left(\left(\frac{n}{\delta}-1\right)I_FN+\frac{n^2}{\delta}\right)+(1-I_I-I_F)Nn\right)\sqrt{I_INn+\dfrac{(1-I_I-I_F)Nn^2}{\left(\frac{n}{\delta}-1\right)I_FN+\frac{n^2}{\delta}}}}.$$

Then

$$\frac{\partial\bar{\pi}_S}{\partial\delta} = -\frac{\sqrt{I_FN+n}\bar{\pi}_S^2}{n^2}G,$$

where

$$G=\frac{\partial}{\partial I_F}\left[\left((I_IN+1)\left(\left(\frac{n}{\delta}-1\right)I_FN+\frac{n^2}{\delta}\right)+(1-I_I-I_F)Nn\right)\sqrt{I_INn+\frac{(1-I_I-I_F)Nn^2}{\left(\frac{n}{\delta}-1\right)I_FN+\frac{n^2}{\delta}}}\right]. \tag{E-28}$$

Therefore, the sign of $\partial\Gamma_I(I_F,I_I,n)/I_F$ is pinned down by the sign of G. Let

$$\Delta=\left(\frac{n}{\delta}-1\right)I_FN+\frac{n^2}{\delta}$$

Simplifying (E-28), we have

$$G=(I_IN+1)\sqrt{I_INn+\frac{(1-I_I-I_F)Nn^2}{\Delta}}\frac{\partial\Delta}{\partial\delta}-\frac{1}{2}\left((I_IN+1)\Delta+(1-I_I-I_F)Nn\right)\frac{1}{\sqrt{I_INn+\frac{(1-I_I-I_F)Nn^2}{\Delta}}}\frac{(1-I_I-I_F)Nn^2}{\Delta^2}\frac{\partial\Delta}{\partial\delta}=KH(I_I).$$

The parameter K is always negative and function H has following expression,

$$H(I_I)=2(I_IN+1)\left(I_INn+\frac{(1-I_I-I_F)Nn^2}{\Delta}\right)\Delta^2-\left((I_IN+1)\Delta+(1-I_I-I_F)Nn\right)(1-I_I-I_F)Nn^2. \tag{E-29}$$

Calculating the derivation for H about I_I, we can show that $H'(I_I)>0$, so H is increasing in I_I. Besides, substituting $(1-I_F)$ into (E-29), it is easy to show that $H(1-I_F)>0$. Substituting $I_I=0$ into $H(I_I)$ in (E-29), we have

$$H(0)=2(1-I_F)Nn^2\Delta-(\Delta+(1-I_F)Nn)(1-I_F)Nn^2=\left[\left(\frac{n}{\delta}-1\right)I_FN+\frac{n^2}{\delta}-(1-I_F)Nn\right](1-I_F)Nn^2.$$

Therefore, when $I_F<I_F^d$,

$$I_F^d=\frac{\delta nN-n^2}{(n+(n-1)\delta)N}. \tag{E-30a}$$

Under this situation, we can find the unique

$$I_I^d=min\{1-I_F,\{I_I|H(I_I)=0,I_I>0\}\} \tag{E-30b}$$

meaning $H\left(I_I^d\right)=0$. Therefore, only when $I_F<I_F^d$ and $I_I<I_I^d$, G is positive. This means that the unconditional slow trader's expected revenue decreases in faster dissemination rate. This leads to an increasing in value of speed.

Information disclosure: Rearranging the expression of value of information, we have

$$\Gamma_I = \frac{n-\delta}{(n-\delta)I_F N + n^2} \frac{n^2}{\left((I_I N+1)n + \frac{\delta(1-I_I-I_F)Nn^2}{(n-\delta)I_F N + n^2}\right)\sqrt{\frac{I_I Nn + \frac{\delta(1-I_I-I_F)Nn^2}{(n-\delta)I_F N + n^2}}{I_F N + n}}} \sigma\sigma_z$$

$$= \frac{1}{\sqrt{n} + \frac{I_F N}{\sqrt{n}} + 1 + \frac{1}{n-\delta}} \frac{1}{\left(I_I N + 1 + \frac{\delta(1-I_I-I_F)N}{\left(1-\frac{\delta}{n}\right)I_F N + n}\right)\sqrt{\frac{I_I N + \frac{\delta(1-I_I-I_F)N}{\left(1-\frac{\delta}{n}\right)I_F N + n}}{I_F N + n}}} \sigma\sigma_z.$$

We also know that

$$\frac{d}{dn}\left[\frac{1}{\sqrt{n} + \frac{I_F N}{\sqrt{n}} + 1 + \frac{1}{n-\delta}}\right] > 0; \qquad \frac{d}{dn}\left[I_I N + 1 + \frac{\delta(1-I_I-I_F)N}{\left(1-\frac{\delta}{n}\right)I_F N + n}\right] < 0.$$

Therefore, higher disclosure level decreases value of information. The proof for the value of speed is similar; thus we omit it here.

E8. Proof of Corollary 6

Equilibrium Fraction of Fast Traders: On the cost of fast trading, according to the equilibrium conditions ($\Gamma_F(I_F, I_I, \theta, \delta) = C_F$ and $\Gamma_I(I_F, I_I, \theta, \delta) = C_I$), we have from (4-11) and (4-12) the condition for equilibrium speed acquisition,

$$C_F = \frac{n}{(I_F N + n)\sqrt{I_F Nn}} \sigma\sigma_z - \frac{\delta n}{I_F N + n} \Psi(I_F, I_I, \theta, \delta).$$

Similarly, the equilibrium information condition becomes

$$C_I = (n-\delta)\Psi(I_F, I_I, \theta, \delta).$$

Combining the above two equations, we have

$$C_F = \frac{n}{I_F N + n}\left(\frac{\sigma\sigma_z}{\sqrt{I_F Nn}} - \frac{1}{n-\delta} C_I\right). \tag{E-31}$$

Applying the implicit function theorem to (E-31), we obtain

$$\frac{dI_F}{dC_F} = -\left(\frac{nN}{(I_F N + n)^2}\left(\frac{\sigma\sigma_z}{\sqrt{I_F Nn}} - \frac{1}{n-\delta} C_I\right) + \frac{n}{I_F N + n} \frac{\sigma\sigma_z}{2I_F\sqrt{I_F Nn}}\right)^{-1} < 0. \tag{E-32}$$

This implies that a decrease in the cost C_F of acquiring speed always increases informed trading in the equilibrium. This is intuitive since a lower C_F implies a higher net benefit of acquiring speed. More interestingly, acquiring speed exhibits a strategic complementarity to producing information. Thus, the complementarity effect characterized in Proposition 4.3 implies that a decrease in C_I, the cost of producing information, may surprisingly increase fast trading in equilibrium. This is demonstrated by the following result (after applying the implicit function theorem to (E-31)),

$$\frac{dI_F}{dC_I} = -\frac{n}{(n-\delta)(I_F N + n)}\left(\frac{nN}{(I_F N + n)^2}\left(\frac{\sigma\sigma_z}{\sqrt{I_F Nn}} - \frac{1}{n-\delta}C_I\right) + \frac{n}{I_F N + n}\frac{\sigma\sigma_z}{2I_F\sqrt{I_F Nn}}\right)^{-1} < 0.$$

Equilibrium Fraction of Informed Traders: We apply the implicit function theorem to the equilibrium conditions $\Gamma_I(I_F, I_I, \theta, \delta) = C_I$ and $\Gamma_F(I_F, I_I, \theta, \delta) = C_F$, respectively, and obtain

$$\frac{dI_I}{dC_I} = -\left(\frac{\partial\Gamma_F}{\partial I_I}\right)^{-1}\frac{\partial\Gamma_F}{\partial I_F}\frac{dI_F}{dC_I}; \qquad \frac{dI_I}{dC_F} = -\left(\frac{\partial\Gamma_I}{\partial I_I}\right)^{-1}\frac{\partial\Gamma_I}{\partial I_F}\frac{dI_F}{dC_F}. \tag{E-33}$$

The term $\frac{dI_I}{dC_I} < 0$ since a lower information cost C_I implies a higher net benefit of acquiring the information. All these results are consistent with Proposition 4.3 that both producing information and acquiring speed are strategic substitute. More interestingly, Proposition 4.3 shows that producing information exhibits either a strategic substitute or complement to acquiring speed. From (E-33), dI_I/dC_F depends on $\partial\Gamma_I/\partial I_I$ (which is negative based on Proposition 4.3), dI_F/dC_F (which is negative as in (E-32)) and $\partial\Gamma_I/\partial I_F$(which is determined by whether producing information exhibits substitutability or complementarity). Thus, these complementarity and substitutability effects imply that a decrease in the cost of acquiring speed may increase (decrease) the equilibrium market fraction of the informed traders (to producing information). Specially, Proposition 4.3 shows that, when $I_I > I_I^c$, a decrease in C_F decreases the fraction of the informed traders. Therefore an increase in the fast trading reduces trader's incentive to produce information. This means that a decrease in C_F, which directly increases the fast trading, indirectly reduces the informed trading. Similarly, when $I_I < I_I^c$, a decrease in C_F increases the fraction of the informed traders.

E9. Equilibrium Corner Solutions

In this subsection, we consider two corner solutions, without either informed traders or fast traders. Under these situations, the cost to be informed or fast is too high so that the expected payoff from the trading is not sufficient enough to cover the cost of acquiring the speed or information, comparing to the slow trader.

Without Informed Trader: This is a special case of Theorem 4.1 when the fraction of informed traders equals to zero. Plugging $I_I = 0$ into (E-2) to (E-4), we have the financial market equilibrium as:

$$\beta_F^{WI} = \sqrt{\frac{n}{I_F N}}\frac{\sigma_z}{\sigma}; \qquad \beta_S^{WI} = \sqrt{\frac{n^2(I_F N + n)}{\left((n-\delta)I_F N + n^2\right)\delta(1-I)N}}\frac{\sigma_z}{\sigma}; \qquad \lambda_1^{WI} = \frac{\sqrt{I_F N n}}{I_F N + n}\frac{\sigma}{\sigma_z};$$

$$\lambda_2^{WI} = \sqrt{\frac{\left((n-\delta)I_F N + n^2\right)\delta(1-I_F)Nn^2}{I_F N + n}}\frac{1}{(n-\delta)I_F N + n^2 + \delta(1-I_F)Nn}\frac{\sigma}{\sigma_z}. \tag{E-34}$$

By calculating the fast and slow informed traders' unconditional expected revenues, we have

$$\bar{\pi}_F = \frac{n}{(I_F N + n)\sqrt{nI_F N}}\sigma\sigma_z; \qquad \bar{\pi}_S = \sqrt{\frac{\delta\left((n-\delta)I_F N + n^2\right)}{(I_F N + n)(1-I_F)N}}\frac{n}{(n-\delta)I_F N + n^2 + \delta(1-I_F)Nn}\sigma\sigma_z.$$

In the equilibrium, a trader chooses to become fast only when his expected payoff from the trading is sufficient to cover the cost of the speed, comparing to a slow trader. Thus, when a trader chooses to acquire speed, the value of speed must be equal to its cost; that is, the trader is indifferent between acquiring speed and being a slow trader. Formally, we have

$$\frac{n}{(I_F N + n)\sqrt{nI_F N}}\sigma\sigma_z - \sqrt{\frac{\delta\left((n-\delta)I_F N + n^2\right)}{(I_F N + n)(1-I_F)N}}\frac{n}{(n-\delta)I_F N + n^2 + \delta(1-I_F)Nn}\sigma\sigma_z = C_F. \tag{E-35}$$

This determines the equilibrium fraction of fast traders.

On the one hand, according to Corollary 4.5, higher disclosure level decreases the fraction of fast traders. On the other hand, faster dissemination rate increases the fraction of faster traders due to $I_F < I_F^d$ and $I_I < I_I^d$. It is easy to show that the left hand side of (E-35) is decreasing in the fraction of fast traders. At the same time, substituting I_F^d into the left hand side of (E-35) leads to a negative value. However, the value of speed must be positive so that both fast and slow traders participate in the market. This means $I_F < I_F^d$.

Coupled with the fact that $I_I = 0$, fast dissemination rate increases the fraction of faster traders.

Without Fast Traders: Consider the trading strategy for the informed traders who trade strategically (considering the price impact). Hence, they maximize their expected profits based on their information,

$$\max_x E\left[x_i\left(V-\lambda\left(x_i+(I_IN-1)x_I+\frac{(1-I_I)N\sum_{i=1}^n x_S^i}{n}+z\right)\right)|V\right],$$

given the slow informed traders' trading strategy $x_S(v_i)=\beta_S^{WF}v_i$. From the first-order condition, the informed trader's trading intensity follows

$$\beta_I^{WF}=\frac{1}{\lambda^{WF}(I_IN+1)}-\frac{(1-I_I)N}{n(I_IN+1)}\delta\beta_S^{WF}.$$

Similarly, we can calculate the first order condition on the expected profit of a slow trader in group i. Then, by substitute the informed traders' demand function, we obtain

$$\beta_S^{WF}=\frac{n}{\lambda^{WF}\big((1-I_I)N+n\big)}-\frac{nI_IN}{(1-I_I)N+n}\beta_I^{WF}-\frac{(1-I_I)N}{(1-I_I)N+n}(\delta\beta_S^{WF}-\beta_S^{WF})$$

These two equations for two parameters jointly determine the unique equilibrium.

On pricing rule, the market maker takes the informed and slow traders' trading strategies as given. In aggregate, trading volume follows

$$w^{WF}=I_INx_I+\frac{(1-I_I)N\sum_{i=1}^n x_S^i}{n}+z=\frac{(1-I_I)N+nI_IN}{\lambda^{WF}\big((1-I_I)N+n+nI_IN\big)}V+z.$$

Applying the projection theorem, we have

$$\lambda^{WF}=\frac{Cov(V,w^{WF})}{Var(w^{WF})}$$

$$=\left(\left(\frac{(1-I_I)N\delta+nI_IN}{\lambda^{WF}\big((1-I_I)N\delta+n+nI_IN\big)}\right)^2\sigma^2+\sigma_z^2\right)^{-1}\frac{(1-I_I)N\delta+nI_IN}{\lambda^{WF}\big((1-I_I)N\delta+n+nI_IN\big)}\sigma^2.$$

Solving the above equations, we have the following equilibrium conditions,

$$\beta_I^{WF}=\beta_S=\sqrt{\frac{n}{(1-I_I)N\delta+nI_IN}}\frac{\sigma_z}{\sigma};\qquad \lambda^{WF}=\frac{\sqrt{n\big((1-I_I)N\delta+nI_IN\big)}}{(1-I_I)N\delta+n+nI_IN}\frac{\sigma}{\sigma_z}.\tag{E-36}$$

By calculating the informed and slow informed traders' unconditional expected revenues, we have

$$\bar{\pi}_I=\frac{n^2}{\lambda_1\big((1-I_I)N\delta+n+nIN\big)^2}\sigma^2;\qquad \bar{\pi}_S=\frac{n\delta}{\lambda_1\big((1-I_I)N\delta+n+nI_IN\big)^2}\sigma^2.$$

In the equilibrium, a trader chooses to become the informed only when his expected payoff from the trading is sufficient to cover the cost of the information, comparing to a slow trader. When a trader chooses to produce information, then the value of information must be equal to its cost; that is, the trader is indifferent between producing information and being a slow trader. Formally, we have

$$\left((1-I_I)N\delta+n+nI_IN\right)^2\left((1-I_I)N\delta+nI_IN\right)=\frac{n(n-\delta)^2}{C_I^2}\sigma^2{\sigma_z}^2. \tag{E-37}$$

This determines the equilibrium fraction of informed traders. On the other hand, the value of information can be represented by

$$\Gamma_I=\frac{1}{\left(\frac{(1-I_I)N}{\frac{n}{\delta}-1}+\frac{1+I_IN}{1-\frac{\delta}{n}}\right)\sqrt{\frac{(1-I_I)N\delta}{n}+IN}}\sigma_v\sigma_z.$$

It is easy to show that both faster information dissemination and more information disclosure decrease the value of information.

E10. Proof of Proposition 4.4

Faster dissemination rate: Applying implicit function theorem to (E-31) delivers,

$$-\frac{nN}{(I_FN+n)^2}\left(\frac{\sigma\sigma_z}{\sqrt{I_FNn}}-\frac{\delta}{n-\delta}C_I\right)\frac{dI_F}{d\delta}+\frac{n}{I_FN+n}\left(-\frac{1}{2}\frac{\sigma\sigma_z}{I_F\sqrt{I_FNn}}\frac{dI_F}{d\delta}-\frac{1}{(n-\delta)^2}C_I\right)=0.$$

Rearranging the above equation, we have

$$\frac{dI_F}{d\delta}=-\frac{\frac{n}{I_FN+n}\frac{1}{(n-\delta)^2}C_I}{\frac{nN}{(I_FN+n)^2}\left(\frac{\sigma\sigma_z}{\sqrt{I_FNn}}-\frac{\delta}{n-\delta}C_I\right)+\frac{n}{I_FN+n}\frac{1}{2}\frac{\sigma\sigma_z}{I_F\sqrt{I_FNn}}}<0. \tag{E-38}$$

Therefore, with faster dissemination rate, the fraction of fast traders decreases.

As for equilibrium fraction of informed traders, in Figure E1, we choose two different values of information cost ($C_I=0.1,0.2$) and keep the speed cost ($C_F=2$) fixed. We plot the equilibrium fraction of the informed traders against information dissemination, together with the threshold value for the information production to be a substitute to information acquisition based on Proposition 4.3. The results confirm the dominance of the crowding-out effect of faster information dissemination on the trade-off. Therefore

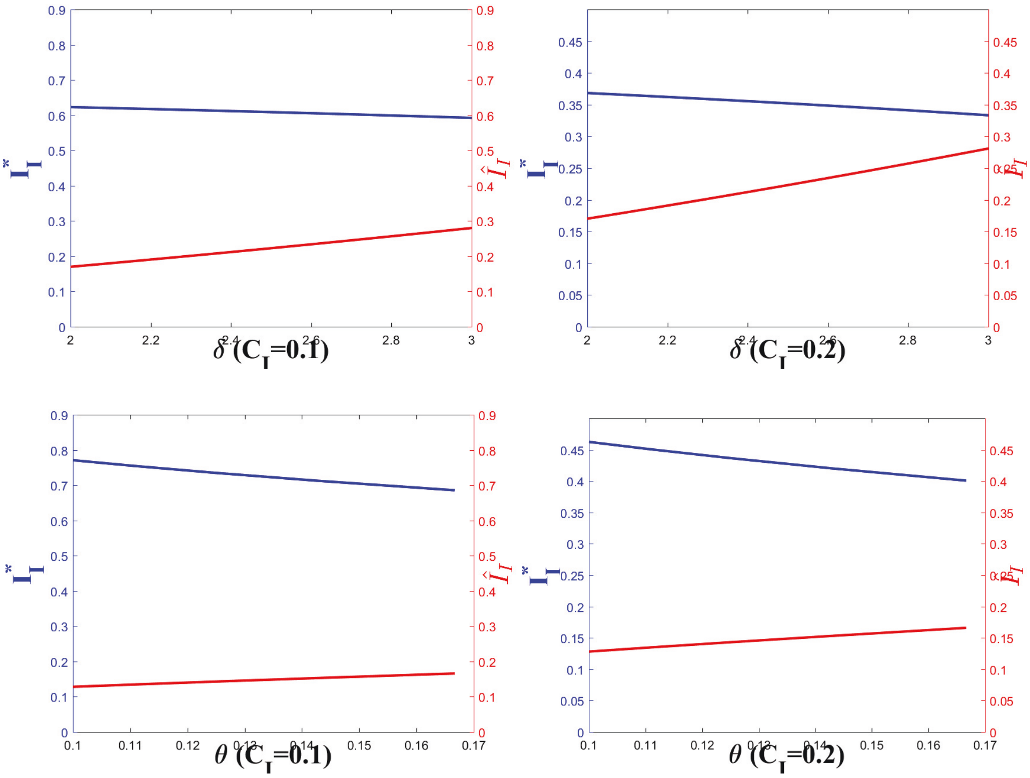


Figure E1: The effect of more information diffusion on information production. The equilibrium fraction of the informed traders (blue lines), together with the threshold value for the information production to be a substitute to information acquisition (red lines), with respect to either the dissemination rate (upper panels; the speed acquisition cost $C_F = 2$, the information disclosure level $n = 10$) or disclosure level (bottom panels; the speed acquisition cost $C_F = 6$, the information dissemination rate $\delta = 1$); other parameters are the same as in Figure 4.3.

with faster dissemination, the equilibrium fraction of informed traders decreases. In general, the significant crowding-out effect dominates the trade-off, leading to a decrease in informed traders with faster dissemination.

Higher Disclosure Level: Applying the implicit function theorem to (D-31) delivers,

$$\frac{dI_F}{dn} = \frac{\frac{I_F N}{I_F N + n}\left(\frac{\sigma\sigma_z}{\sqrt{I_F N n}} - \frac{1}{n-1}C_I\right) + n\left(-\frac{\sigma\sigma_z}{2n\sqrt{I_F N n}} + \frac{C_I}{(n-1)^2}\right)}{\frac{Nn}{I_F N + n}\left(\frac{\sigma\sigma_z}{\sqrt{I_F N n}} - \frac{1}{n-1}C_I\right) + n\frac{\sigma\sigma_z}{2I_F\sqrt{I_F N n}}} > 0. \tag{D-39}$$

This shows that higher disclosure level crowds out fast traders

As for equilibrium fraction of informed traders, in Figure D1, we choose two different values of information production cost ($C_I = 0.1, 0.2$) and keep the speed acquisition cost ($C_F = 6$) fixed for all the scenarios; together with an information diffusion speed $\delta = 1$.

We plot the equilibrium fraction of the informed traders against the disclosure level for the combinations of information production and speed acquisition costs. In addition, by substituting the equilibrium outcomes into $\hat{I}_I$, we show that all equilibria satisfy the condition for the information production to be substitute to information acquisition. In both panels of Figure D1, the numerical results confirm our analyzis that the trade-off of the higher disclosure level is dominated by the crowding-out effect, meaning that, with higher disclosure level, the equilibrium fraction of informed traders decreases. In general, although the crowding-out effect and the fast trading effect can either strengthen or weaken each other, depending on the learning complementarity and substitutability, the significant crowding-out effect dominates the trade-off in both scenarios, leading to a decrease in informed traders with higher disclosure level.

E11. Disclosure's Impact on Fast Trading

In this Appendix, we mainly show that higher disclosure promotes fast trading. Applying the implicit function theorem to Equation (4-17), we have

$$\frac{d\Phi_F}{dn} = \underbrace{\frac{1}{2}\sqrt{\frac{N}{nI_F}}\frac{\sigma_z}{\sigma}\frac{dI_F}{dn}}_{Crowding-out\ effet} - \underbrace{\frac{1}{2n}\sqrt{\frac{I_F N}{n}}\frac{\sigma_z}{\sigma}}_{Weakening\ effect} . \tag{E-40}$$

In the early trading period, higher disclosure not only brings more information to fast traders, weakening information asymmetry, but also reduces trader's incentive to invest in the speed technology, crowding-out fast trading. In (E-40), the first term captures the crowding-out effect on fast traders; while the second term captures the effect of more information for fast traders.

With the crowding-out effect discussed in Subsection 4.3.2, we now characterize the second term. The information diffusion determines the unique information structure and intra-temporal competition among the fast traders. Corollary 4.3 shows that, with higher disclosure, individual fast trader reduces his trading sensitivity β_F to the signal, but also receives more precise information. In equilibrium, this trade-off for faster traders is dominated by more transparent information, which increases the trading intensity of the fast traders Φ_F (given their fraction). This is essentially reflected by the second term of equation (E-40). When considering the general effect, the sigh of equation (E-40) is then pinned down by

$$-\frac{\sigma\sigma_z}{\sqrt{I_F N n}}+\frac{n}{(n-1)^2}C_I. \tag{E-41}$$

In the equilibrium,

$$C_F=\frac{n}{I_F N+n}\left(\frac{\sigma\sigma_z}{\sqrt{I_F N n}}-\frac{1}{n-1}C_I\right).$$

Thus, equation (E-41) can be represented as,

$$-\frac{\sigma\sigma_z}{\sqrt{I_F N n}}+\frac{n}{(n-1)^2}C_I=-\frac{I_F N+n}{n}C_F+\frac{1}{(n-1)^2}C_I.$$

When both the informed and fast traders are active in the market, we show based on the numerical analysis that the speed cost should be larger than the information cost. Thus, the expression in (E-41) is positive. This means that, with higher disclosure and the speed technology, the first-mover advantage makes the fast traders trade more aggressively. Therefore, although higher disclosure and the interaction between the fast and informed traders amplify the crowding-out effect to the fast traders (as discussed in subsection 4.3.2), the positive information effect dominates the negative crowding-out effect, leading the fast traders to trade more aggressively.

E12. Scenario without Learning

The early period equilibrium remains the same as in Appendix E1. For late period equilibrium, due to both informed and slow traders do not observe the early period equilibrium price, they make decision based only on signals. Considering the trading strategy for the informed traders, they trade strategically (considering the price impact). Hence, they maximize their expected profits based on their information,

$$\max_x E\left[x_i\left(V-\lambda_2\left(x_i+(I_I N-1)x_I+\frac{(1-I_I-I_F)N\sum_{i=1}^n x_S^i}{n}+z_2\right)-p_1\right)|V\right],$$

given the slow traders' trading strategy $x_S(s_i^S)=\beta_S s_i^S$. From the first-order condition we obtain the informed trader's trading strategy

$$X^I = \frac{n^2 - \lambda_2(I_F N + n)(1 - I_I - I_F)N\delta\beta_S}{\lambda_2 n(I_I N + 1)(I_F N + n)} V.$$

Therefore, the informed traders' trading intensity is the best response function of the slow informed traders' trading intensity,

$$\beta_I(\beta_S) = \frac{n^2 - \lambda_2(I_F N + n)(1 - I_I - I_F)N\delta\beta_S}{\lambda_2 n(I_I N + 1)(I_F N + n)}.$$

Each of the slow traders in group *i* maximizes the expected profits. Similarly, given informed traders' trading strategy, we can calculate the first order condition on the expected utility of slow traders, substitute the informed traders' demand function, and obtain

$$\beta_S(\beta_I, \beta_S) = \frac{n^2}{\lambda_2\big((1 - I_I - I_F)N + n\big)(I_F N + n)} - \frac{n I_I N}{(1 - I_I - I_F)N + n}\beta_I - \frac{(1 - I_I - I_F)N}{(1 - I_I - I_F)N + n}(\delta - 1)\beta_S.$$

Therefore the informed traders' trading behaviour is the best response functions of the slow traders' behaviour, and visa verse, jointly determining the equilibrium. With the two equations and two parameters, the linear system uniquely determines the market equilibrium. We therefore have the informed and slow traders' trading strategies as the functions of fundamental parameters and market maker's pricing rule,

$$\beta_I = \beta_S = \frac{n^2}{\lambda_2(I_F N + n)\big((1 - I_I - I_F)N\delta + I_I N n + n\big)}.$$

On the pricing rule, the market maker takes the informed and slow traders' trading strategies as given. The aggregate trading volume follows

$$w_2 = I_I N x_I + \frac{(1 - I_I - I_F)N \sum_{i=1}^{n} x_S^i}{n} + z_2.$$

Substituting the demand functions into the aggregate trading volume, we have

$$w_2 = \frac{n(n I_I N + (1 - I_I - I_F)N\delta)}{\lambda_2(I_F N + n)\big((1 - I_I - I_F)N\delta + I_I N n + n\big)} V + z_2.$$

In the late period, the market maker sees to aggregate trading volumes in both early and late periods. Applying the projection theorem, we have

$$\lambda_2 = \frac{Cov(V, w_1)Var(w_0) - Cov(w_1, w_0)Cov(V, w_0)}{Var(w_1)Var(w_0) - Cov(w_1, w_0)^2}$$

$$= \left(\left(\frac{n(n I_I N + (1 - I_I - I_F)N\delta)}{\lambda_2(I_F N + n)\big((1 - I_I - I_F)N\delta + I_I N n + n\big)}\right)^2 \sigma^2 + \left(\frac{I_F N}{n} + 1\right)\sigma_z^2\right)^{-1} \frac{n(n I_I N + (1 - I_I - I_F)N\delta)}{\lambda_2(I_F N + n)\big((1 - I_I - I_F)N\delta + I_I N n + n\big)}\sigma^2.$$

Solving the above equations, we have the equilibrium conditions for the late period,

$$\lambda_2 = \frac{n\sqrt{(nI_I N + (1 - I_I - I_F)N\delta)\big(I_F N\big((1 - I_I - I_F)N\delta + I_I Nn + n\big) + n^2\big)}}{(I_F N + n)\big((1 - I_I - I_F)N\delta + I_I Nn + n\big)\sqrt{I_F N + n}}\frac{\sigma}{\sigma_z}; \tag{E-42}$$

$$\beta_I = \beta_S = \frac{n\sqrt{I_F N + n}}{\sqrt{(nI_I N + (1 - I_I - I_F)N\delta)\big(I_F N\big((1 - I_I - I_F)N\delta + I_I Nn + n\big) + n^2\big)}}\frac{\sigma_z}{\sigma}; \tag{E-43}$$

Equations (E-1), (E-42) (E-43) describe the equilibrium.

By calculating the informed and slow traders' unconditional expected revenues, we have

$$\bar{\pi}_I = \frac{n^4}{\lambda_2 (I_F N + n)^2 \big((1 - I_I - I_F)N\delta + I_I Nn + n\big)^2}\sigma^2;$$

$$\bar{\pi}_S = \frac{n^3\delta}{\lambda_2 (I_F N + n)^2 \big((1 - I_I - I_F)N\delta + I_I Nn + n\big)^2}\sigma^2.$$

In the equilibrium, the value of information and speed must be equal to their costs. Formally, we have the following conditions that determine the endogenous fraction of fast and informed traders.

$$C_I = \frac{n^3(n - \delta)}{\lambda_2 (I_F N + n)^2 \big((1 - I_I - I_F)N\delta + I_I Nn + n\big)^2}\sigma^2; \tag{E-44}$$

$$C_F = \frac{\sqrt{n}}{\sqrt{I_F N}(I_F N + n)}\sigma\sigma_z - \frac{n^3\delta}{\lambda_2 (I_F N + n)^2 \big((1 - I_I - I_F)N\delta + I_I Nn + n\big)^2}\sigma^2. \tag{E-45}$$

Here, we finished the proof of scenario without learning.

Bibliography

Abadie, A. & Kasy, M. 2017. The risk of machine learning. *arXiv preprint arXiv:1703.10935*.

Aït-Sahalia, Y. & Saglam, M. 2013. High frequency traders: Taking advantage of speed. National Bureau of Economic Research.

Anand, A. & Venkataraman, K. 2016. Market conditions, fragility, and the economics of market making. *Journal of Financial Economics,* 121**,** 327-349.

Anand, B. N. & Galetovic, A. 2000. Information, nonexcludability, and financial market structure. *The Journal of Business,* 73**,** 357-402.

Anderson, S. P., De Palma, A. & Thisse, J.-F. 1992. *Discrete choice theory of product differentiation*, MIT press.

Aoyagi, J. 2020. The Dark Side of Regulating Fast Informed Trading.

Avdis, E. 2016. Information tradeoffs in dynamic financial markets. *Journal of Financial Economics,* 122**,** 568-584.

Avery, C. & Zemsky, P. 1998. Multidimensional uncertainty and herd behavior in financial markets. *American economic review***,** 724-748.

Back, K., Crotty, K. & Li, T. 2018. Identifying information asymmetry in securities markets. *The Review of Financial Studies,* 31**,** 2277-2325.

Back, K. & Pedersen, H. 1998. Long-lived information and intraday patterns. *Journal of financial markets,* 1**,** 385-402.

Bai, J., Philippon, T. & Savov, A. 2016. Have financial markets become more informative? *Journal of Financial Economics,* 122**,** 625-654.

Banerjee, S., Davis, J. & Gondhi, N. 2018. When transparency improves, must prices reflect fundamentals better? *The Review of Financial Studies,* 31**,** 2377-2414.

Banerjee, S. & Green, B. 2015. Signal or noise? Uncertainty and learning about whether other traders are informed. *Journal of Financial Economics,* 117**,** 398-423.

Barlevy, G. & Veronesi, P. 2003. Rational panics and stock market crashes. *Journal of Economic Theory,* 110**,** 234-263.

Baron, M., Brogaard, J., Hagströmer, B. & Kirilenko, A. 2019. Risk and return in high-frequency trading. *Journal of Financial and Quantitative Analysis,* 54**,** 993-1024.

Bernardo, A. E. & Judd, K. L. 2000. Asset market equilibrium with general tastes, returns, and informational asymmetries. *Journal of Financial Markets,* 3**,** 17-43.

Bessembinder, H., Carrion, A., Tuttle, L. & Venkataraman, K. 2016. Liquidity, resiliency and market quality around predictable trades: Theory and evidence. *Journal of Financial economics,* 121, 142-166.

Biais, B. & Foucault, T. 2014. HFT and market quality. *Bankers, Markets & Investors*, 5-19.

Biais, B., Foucault, T. & Moinas, S. 2015. Equilibrium fast trading. *Journal of Financial economics,* 116, 292-313.

Boehmer, E., Li, D. & Saar, G. 2018. The competitive landscape of high-frequency trading firms. *The Review of Financial Studies,* 31, 2227-2276.

Bond, P., Edmans, A. & Goldstein, I. 2012. The real effects of financial markets. *Annu. Rev. Financ. Econ.,* 4, 339-360.

Bongaerts, D. & Van Achter, M. 2020. Competition among Liquidity Providers with Access to High-Frequency Trading Technology. *Available at SSRN 2698702*.

Breugem, M. & Buss, A. 2019. Institutional investors and information acquisition: Implications for asset prices and informational efficiency. *The Review of Financial Studies,* 32, 2260-2301.

Brogaard, J., Carrion, A., Moyaert, T., Riordan, R., Shkilko, A. & Sokolov, K. 2018. High frequency trading and extreme price movements. *Journal of Financial Economics,* 128, 253-265.

Brogaard, J., Hendershott, T. & Riordan, R. 2014. High-frequency trading and price discovery. *The Review of Financial Studies,* 27, 2267-2306.

Budish, E., Cramton, P. & Shim, J. 2015. The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics,* 130, 1547-1621.

Cespa, G. & Foucault, T. 2014. Illiquidity contagion and liquidity crashes. *The Review of Financial Studies,* 27, 1615-1660.

Chen, H., Foley, S., Goldstein, M. A. & Ruf, T. 2017a. The value of a millisecond: Harnessing information in fast, fragmented markets. *Fragmented Markets (November 18, 2017)*.

Chen, H., Petukhov, A. & Wang, J. 2017b. The dark side of circuit breakers. *Work. Pap., Mass. Inst. Technol., Cambridge, MA*.

Chinco, A., Clark-Joseph, A. D. & Ye, M. 2019. Sparse signals in the cross-section of returns. *The Journal of Finance,* 74, 449-492.

Chowdhry, B. & Nanda, V. 1991. Multimarket trading and market liquidity. *The Review of Financial Studies,* 4, 483-511.

Christensen, K., Oomen, R. C. & Podolskij, M. 2014. Fact or friction: Jumps at ultra high frequency. *Journal of Financial Economics,* 114, 576-599.

De Prado, M. L. 2018. *Advances in financial machine learning*, John Wiley & Sons.

Duffie, D., Malamud, S. & Manso, G. 2009. Information percolation with equilibrium search dynamics. *Econometrica,* 77, 1513-1574.

Duffie, D. & Manso, G. 2007. Information percolation in large markets. *American Economic Review,* 97, 203-209.

Dugast, J. & Foucault, T. 2018. Data abundance and asset price informativeness. *Journal of Financial Economics,* 130, 367-391.

Fishe, R. P., Haynes, R. & Onur, E. 2019. Anticipatory traders and trading speed. *Journal of Financial and Quantitative Analysis,* 54, 729-758.

Fishman, M. J. & Hagerty, K. M. 1992. Insider trading and the efficiency of stock prices. *The RAND Journal of Economics*, 106-122.

Foster, F. D. & Viswanathan, S. 1993. The effect of public information and competition on trading volume and price volatility. *The Review of Financial Studies,* 6**,** 23-56.

Foster, F. D. & Viswanathan, S. 1996. Strategic trading when agents forecast the forecasts of others. *The Journal of Finance,* 51**,** 1437-1478.

Foucault, T., Hombert, J. & Roşu, I. 2016. News trading and speed. *The Journal of Finance,* 71**,** 335-382.

Foucault, T. & Moinas, S. 2018. Is Trading Fast Dangerous? *Global Algorithmic Capital Markets: High Frequency Trading, Dark Pools, and Regulatory Challenges*, 9.

Froot, K. A., Scharfstein, D. S. & Stein, J. C. 1992. Herd on the street: Informational inefficiencies in a market with short‐term speculation. *The Journal of Finance,* 47**,** 1461-1484.

Gao, F., Song, F. & Wang, J. 2013. Rational expectations equilibrium with uncertain proportion of informed traders. *Journal of Financial Markets,* 16**,** 387-413.

Gârleanu, N. & Pedersen, L. H. 2018. Efficiently inefficient markets for assets and asset management. *The Journal of Finance,* 73**,** 1663-1712.

Gennotte, G. & Leland, H. 1990. Market liquidity, hedging, and crashes. *The American Economic Review*, 999-1021.

Gervais, S. & Hall, S. H. D. 1997. Market microstructure with uncertain information precision: A new framework. Tech. rep., Working Paper, University of Pennsylvania.

Goldstein, I., Li, Y. & Yang, L. 2014. Speculation and hedging in segmented markets. *The Review of Financial Studies,* 27**,** 881-922.

Goldstein, I. & Yang, L. 2017. Information disclosure in financial markets. *Annual Review of Financial Economics,* 9**,** 101-125.

Goldstein, I. & Yang, L. 2019. Good disclosure, bad disclosure. *Journal of Financial Economics,* 131**,** 118-138.

Greenwald, B. C. & Stein, J. C. 1991. Transactional risk, market crashes, and the role of circuit breakers. *Journal of Business*, 443-462.

Grossman, S. J. & Stiglitz, J. E. 1980. On the impossibility of informationally efficient markets. *The American economic review,* 70**,** 393-408.

Gu, S., Kelly, B. & Xiu, D. 2018. Empirical asset pricing via machine learning. National Bureau of Economic Research.

Haas, M., Khapko, M. & Zoican, M. Speed and learning in high-frequency auctions. Paris December 2016 Finance Meeting EUROFIDAI-AFFI, 2019.

Han, B., Tang, Y. & Yang, L. 2016. Public information and uninformed trading: Implications for market liquidity and price efficiency. *Journal of Economic Theory,* 163**,** 604-643.

Han, B. & Yang, L. 2013. Social networks, information acquisition, and asset prices. *Management Science,* 59**,** 1444-1457.

Hasbrouck, J. 1991. Measuring the information content of stock trades. *The Journal of Finance,* 46**,** 179-207.

Hausman, D. M. & A Simon, H. 1992. *The inexact and separate science of economics*, Cambridge University Press.

Heimer, R. Z. & Simon, D. Facebook finance: How social interaction propagates active investing. AFA 2013 San Diego Meetings, 2012.

Hirshleifer, D., Subrahmanyam, A. & Titman, S. 1994. Security analysis and trading patterns when some investors receive information before others. *The Journal of Finance,* 49**,** 1665-1698.

Hoffmann, P. 2014. A dynamic limit order market with fast and slow traders. *Journal of Financial Economics,* 113**,** 156-169.

Holden, C. W. & Subrahmanyam, A. 1996. Risk aversion, liquidity, and endogenous short horizons. *The Review of Financial Studies,* 9**,** 691-722.

Holden, C. W. & Subrahmanyam, A. 2002. News events, information acquisition, and serial correlation. *The Journal of Business,* 75**,** 1-32.

Hong, D., Hong, H. G. & Ungureanu, A. An epidemiological approach to opinion and price-volume dynamics. AFA 2012 Chicago Meetings Paper, 2011.

Hong, H., Kubik, J. D. & Stein, J. C. 2004. Social interaction and stock‐market participation. *The journal of finance,* 59**,** 137-163.

Hong, H. & Stein, J. C. 1999. A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of finance,* 54**,** 2143-2184.

Hong, H., Stein, J. C. & Yu, J. 2007. Simple forecasts and paradigm shifts. *The Journal of Finance,* 62**,** 1207-1242.

Hu, G. X., Pan, J. & Wang, J. 2016. Early peek advantage? efficient price discovery with tiered information disclosure. *Second Annual Volatility Institute at NYU Shanghai (VINS)*.

Huang, S. & Yueshen, B. Z. 2018. Speed acquisition. *Available at SSRN 2845864*.

Hughes, J. S., Liu, J. & Liu, J. 2007. Information asymmetry, diversification, and cost of capital. *The accounting review,* 82**,** 705-729.

Ivković, Z. & Weisbenner, S. 2007. Information diffusion effects in individual investors' common stock purchases: Covet thy neighbors' investment choices. *The Review of Financial Studies,* 20**,** 1327-1357.

Jackson, M. O. 2005. The economics of social networks.

Jensen, M. C. & Heckling, W. H. 1995. Specific and general knowledge, and organizational structure. *Journal of applied corporate finance,* 8**,** 4-18.

Jovanovic, B. & Menkveld, A. J. 2016. Middlemen in limit order markets. *Available at SSRN 1624329*.

Keim, D. B. & Madhavan, A. 1996. The upstairs market for large-block transactions: Analysis and measurement of price effects. *The Review of Financial Studies,* 9**,** 1-36.

Kendall, C. 2018. The time cost of information in financial markets. *Journal of Economic Theory,* 176**,** 118-157.

Kim, O. & Verrecchia, R. E. 1994. Market liquidity and volume around earnings announcements. *Journal of accounting and economics,* 17**,** 41-67.

Kirilenko, A., Kyle, A. S., Samadi, M. & Tuzun, T. 2017. The flash crash: High‐frequency trading in an electronic market. *The Journal of Finance,* 72**,** 967-998.

Korajczyk, R. A. & Murphy, D. 2019. High-frequency market making to large institutional trades. *The Review of Financial Studies,* 32**,** 1034-1067.

Krauss, C., Do, X. A. & Huck, N. 2017. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research,* 259, 689-702.

Kyle, A. S. 1985. Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, 1315-1335.

Lee, S. S. & Mykland, P. A. 2012. Jumps in equilibrium prices and market microstructure noise. *Journal of Econometrics,* 168, 396-406.

Lee, T. 2019. Latency in fragmented markets. *Review of Economic Dynamics,* 33, 128-153.

Li, S., Wang, X. & Ye, M. 2019. Who Provides Liquidity, and When? : National Bureau of Economic Research.

Li, T. 2013. Insider trading with uncertain informed trading. *Available at SSRN 946324*.

Lillo, F., Farmer, J. D. & Mantegna, R. N. 2003. Master curve for price-impact function. *Nature,* 421, 129-130.

Malinova, K., Park, A. & Riordan, R. 2018. Do retail investors suffer from high frequency traders? *Available at SSRN 2183806*.

Manela, A. 2014. The value of diffusing information. *Journal of Financial Economics,* 111, 181-199.

Mcfadden, D. 1974. The measurement of urban travel demand. *Journal of public economics,* 3, 303-328.

Mendelson, H. & Tunca, T. I. 2004. Strategic trading, liquidity, and information acquisition. *The Review of Financial Studies,* 17, 295-337.

Menkveld, A. J. 2016. The economics of high-frequency trading: Taking stock. *Annual Review of Financial Economics,* 8, 1-24.

Menkveld, A. J. & Zoican, M. A. 2017. Need for speed? Exchange latency and liquidity. *The Review of Financial Studies,* 30, 1188-1228.

Morris, S. & Shin, H. S. 2002. Social value of public information. *american economic review,* 92, 1521-1534.

Mullainathan, S. & Spiess, J. 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives,* 31, 87-106.

O'hara, M., Yao, C. & Ye, M. 2014. What's not there: Odd lots and market data. *The Journal of Finance,* 69, 2199-2236.

O'hara, M. & Ye, M. 2011. Is market fragmentation harming market quality? *Journal of Financial Economics,* 100, 459-474.

Ozsoylev, H. N. & Walden, J. 2011. Asset pricing in large information networks. *Journal of Economic Theory,* 146, 2252-2280.

Pagnotta, E. S. & Philippon, T. 2018. Competing on speed. *Econometrica,* 86, 1067-1115.

Popper, N. 2012. High-speed trading no longer hurtling forward. *The New York Times,* 14.

Raman, V., Robe, M. A. & Yadav, P. K. 2014. Electronic market makers, trader anonymity and market fragility. *Trader Anonymity and Market Fragility (May 29, 2014)*.

Riordan, R. & Storkenmaier, A. 2012. Latency, liquidity and price discovery. *Journal of Financial Markets,* 15, 416-437.

Romer, D. 1992. Rational asset price movements without news. National bureau of economic research.

Rossi, A. G. 2018. Predicting stock market returns with machine learning. Working paper.

Roşu, I. 2019. Fast and slow informed trading. *Journal of Financial Markets,* 43**,** 1-30.

Simon, H. A. 1955. A behavioral model of rational choice. *The quarterly journal of economics,* 69**,** 99-118.

Subrahmanyam, A. 1994. Circuit breakers and market volatility: A theoretical perspective. *The Journal of Finance,* 49**,** 237-254.

Van Kervel, V. & Menkveld, A. J. 2019. High - frequency trading around large institutional orders. *The Journal of Finance,* 74**,** 1091-1137.

Varian, H. 1989. Survey evidence on the diffusion of interest and information among investors. Kluwer Academic Publishers, Boston, MA.

Varian, H. R. 2014. Big data: New tricks for econometrics. *Journal of Economic Perspectives,* 28**,** 3-28.

Verrecchia, R. E. 1982. Information acquisition in a noisy rational expectations economy. *Econometrica: Journal of the Econometric Society***,** 1415-1430.

Verrecchia, R. E. 2001. Essays on disclosure. *Journal of accounting and economics,* 32**,** 97-180.

Walden, J. 2019. Trading, profits, and volatility in a dynamic information network model. *The Review of Economic Studies,* 86**,** 2248-2283.

Weller, B. 2016. Efficient prices at any cost: does algorithmic trading deter information acquisition? *SSRN Scholarly Paper ID,* 2662254**,** 43.

Yang, L. & Zhu, H. 2020. Back-running: Seeking and hiding fundamental information in order flows. *The Review of Financial Studies,* 33**,** 1484-1533.

Ye, M., Yao, C. & Gai, J. 2013. The externalities of high frequency trading. *Available at SSRN 2066839*.

Yuan, K. 2005. Asymmetric price movements and borrowing constraints: A rational expectations equilibrium model of crises, contagion, and confusion. *The Journal of Finance,* 60**,** 379-411.