Essays on and Beyond Market Microstructure

by

Yenan Wang

Business Administration
Duke University

Date: _____

Approved:

_____
S. Viswanathan, Advisor

_____
Ming Yang, Advisor

_____
Adriano A. Rampini

_____
Luis Felipe Varas Greene

_____
Todd Sarver

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Business Administration
in the Graduate School of Duke University

2021

ABSTRACT

Essays on and Beyond Market Microstructure

by

Yenan Wang

Business Administration
Duke University

Date: _____

Approved:

_____
S. Viswanathan, Advisor

_____
Ming Yang, Advisor

_____
Adriano A. Rampini

_____
Luis Felipe Varas Greene

_____
Todd Sarver

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Business Administration in the Graduate School of Duke University

2021

Abstract

My dissertation consists of two parts. The first part examines standard microstructure topics from new perspectives. Chapter 2 investigates the market implications of high-frequency trading focusing on an ignored channel that existing market makers might reduce the capital exploited in market making facing the competition from high-frequency traders. I find that more high-frequency trading might not improve market quality exactly because of this channel. My model also generates several policy suggestions in improving the market quality with the presence of high-frequency trading. Chapter 3 studies the large traders' execution problem to understand whether electronic trading brings execution advantage. I find that although an increase in trading frequency improves their executions, losing the ability to adopt reputation-based trading strategies due to the prevalence of anonymous trading in this electronic trading era can be so costly that large traders become worse-off in execution. Chapter 4 considers an insider trading problem with uncertainty over the insider's existence. The competitive market making assumption is not consistent with this new element because the insider can achieve an unbounded payoff by taking advantage of the competition. Instead, I construct and analyze the equilibrium under the assumption of monopolistic market making. The model can explain the heterogeneous changes in liquidity of different stocks after a regulation.

The second part of my dissertation studies two elements outside the scope of current microstructure literature but (in my opinion) are important and should be incorporated to achieve a comprehensive understanding of financial markets. Chapter 5 studies an attention allocation problem under the context of contract and information design. The principal needs to jointly design compensation for the agent and his attention allocation in acquiring signals to motivate the agent to exert the unobservable effort. The flexibility to allocate attention can lead to perpetual cooperation between the principal and the agent. Chapter 6 studies belief updating when the state space contains elements that we are not aware of with an axiomatic approach. The agent can learn from what he is currently aware of and updates his belief when awareness is related to the state realization. In this case, besides leading to Bayesian updating, the arrival of a new signal can also expand the agent's awareness and trigger further updating. Its implication is discussed under the context of persuasion.

Dedication

To Mingjie Liu, for her continual support and company.

Acknowledgements

I always think the academic career as a solitary pilgrimage for enlightenment. The nature of this journey makes the help and kindness I received from many along the way immensely valuable. I want to take a moment to thank them.

Members of my dissertation committee are my academic role models. I would like to extend my deepest gratitude to my advisor S. Viswanathan for being an exemplary advisor granting me full freedom to explore my interest as well as offering insightful advice timely when I need guidance. I will always be grateful for your meetings with me on weekends. I am also deeply indebted to another advisor Ming Yang for his detailed academic guidance and his faith in me to work with me on multiple projects. I learned a lot from the collaboration process. I also want to thank committee members Adriano Rampini, Todd Sarver and Felipe Varas. They taught me general knowledge on various branches of economics and finance and provided many comments on my research that are helpful and inspiring.

Many others have helped and guided me academically. I would like to thank Simon Gervais and Pete Kyle for providing many insightful suggestions on an early version of my job market paper. I want to thank Bruno Biais for discussing my job market paper on the SaMMF job market candidate workshop. I wish to thank John Shim for discussing my job market paper on SFS Cavalcade and later co-authoring

a paper with me. Speaking of co-authoring, I also had great pleasure working with Liang Dai. I am grateful to other members at Fuqua Finance, especially Anna Cieslak, David Robinson and Manuel Adelino, for providing valuable suggestions on the job market as well as the academic career. Thanks should also go to Yao Zeng for giving advice on my job market spiel on a weekend.

With Fuqua's friendly Ph.D. student community, my experience as a Ph.D. student is very pleasant. Special thanks to my Ph.D. cohort Jing Huang, Jay Im and Youngjun Song for being wonderful classmates and great office neighbors. I very much appreciate that John Barry, Andrew Kane and James Pinnington offered extensive suggestions on my writing. I also want to thank Morad Elasify, who is also on the job market this year, for sharing job market information and discussing application strategies with me during this difficult job season.

On a more personal level, I am forever grateful to my parents, Feng Wang and Ying Sun, for offering me unconditional love, support and trust. As open-mind parents, they granted me complete freedom and shielded me from the pressure for good grades. This preserves my intellectual curiosity, which I considered as one of the most important traits for me to enjoy an academic career. I would also like to thank my wonderful friends including Rong Chen, Tiancheng Chen, Jiangyuan Li, Huidi Lin, Yingjie Liu, Joy Tong, Ge Wang, Yahui Wang, David Xiaoyu Xu, Qi Zhang, Xiaojue Zhang and many others. They make me feel cared, understood and supported. I have a lot of cherished memories with each one of them.

Finally, I want to thank Mingjie Liu. In all these years at Durham, I never felt envious when my friends were preaching their wonderful lives in large cities. Your company must be part of the reason for this preference.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

How financial markets are organized has significant implications for market participants' trading behaviors and prices of the assets. At a high level, this dissertation examines this general theme from two perspectives. The first part of this dissertation builds upon existing microstructure frameworks to address new topics emerging from the electronic trading era where tradings become faster and many orders are anonymous. The second part, using frameworks from game theory and decision theory, explores elements which have significant impacts on the organizations of financial markets but have not been incorporated into microstructure framework yet.

With the prevalence of trading technology, in the past decades, the trading form in financial markets gradually evolves from floor-based trading into electronic trading. This transformation calls for new frameworks to study its implications. The first part of my dissertation studies topics emerge in the electronic trading era: high-frequency trading and market quality, execution with electronic trading and uncertainty over the existence of an insider.

Chapter 2 considers the market quality implication of high-frequency trading focusing on the capital commitment channel. A huge body of empirical literature documents that increased activity of high-frequency traders is related to the decrease of bid-ask spread. Many considers this as the evidence that high-frequency trading improves market quality. However, there is also evidence pointing towards the other direction. For instance, it is also documented that traders have difficulties executing large trades and the average order size becomes smaller. To solve this discrepancy, I argue that existing models mainly focus on the price competition channel and ignore the capital commitment channel. Specifically, high-frequency trader's competition makes existing market makers become reluctant to commit capital in market making. On the other hand, limited by their business models and exogenous market conditions, high-frequency traders cannot indefinitely expand their operations to fill the gap. Despite the price competition brought by high-frequency traders reduces the spread, this effect deteriorates market quality by making the market shallower. Two effects need to be jointly considered to determine the net effect of high-frequency trading on market quality. Importantly, high-frequency trading is not always beneficial to the market and certain policies regarding high-frequency trading can be designed to improve market quality.

Chapter 3 examines large trader's execution in the electronic trading era. The transformation from floor-based trading to electronic trading brings two changes. First, the number of trades in a specific period of time increases. Second, many trades become anonymous such that large traders have hard time committing to future trades. Though higher trading speed improves the trader's execution, lack of

ability in implementing reputation-based trading strategy significantly deteriorates the trader's execution. Based on this analysis, large traders should try to regain the capacity to commit future trades (e.g., publicizing execution algorithms) to improve execution.

Chapter 4 analyzes an insider trading model with uncertainty over the existence of the insider. Competitive market making framework is not consistent with the uncertainty over insider's existence since the insider gains infinite payoff and the market breaks down. To overcome this problem, I construct and analyze the equilibrium under the assumption of monopolistic market making. In this case, the market maker's pricing strategy only depends on the ratio of two types of insider's existing probabilities. This explains the observation that different stocks' liquidity may change in opposite directions after a regulation.

The second part of the dissertation uses game theory and decision theory frameworks to explore two elements that have important implications for financial markets but are currently not incorporated into the microstructure literature: strategic allocation of limited attention and unexpected factors and their relation with belief updating. Although these elements are not investigated within the scope of microstructure in this dissertation, my exploration serves as a foundation for the future research of financial markets.

Chapter 5 considers a contracting problem where the principal needs to motivate the agent to exert effort. The principal cannot observe the agent's effort level but can allocate limited amount of attention in obtaining positive and negative signals with respect to the agent's effort. In other words, the principal needs to jointly

design the compensation scheme for the agent as while as the attention allocation scheme for himself. When the common interest of the principal and the agent is low, the principal allocation all attention on receiving positive signal of the agent's effort. With high common interest, the principal pay attention mostly to the negative signal of the agent's effort. Moreover, contrary to standard contracting literature without attention allocation, when the principal attention budget is high, it is possible for the principal to perpetuate the working relationship with the agent.

Chapter 6 studies belief formation and belief updating with presence of unexpected elements (unawareness) in the state space. In this case, the objective and subjective state space do not coincide and the agent may obtain information from the size of his state space. I adopt an axiomatic approach to analyze this situation. The agent can learn from what he is currently aware of and updates his belief when awareness is related to the state realization. Importantly, new signal expanding the agent's awareness may also lead to belief updating and the belief process under this situation is not a martingale. The implication is discussed under the context of persuasion.

Chapter 2

High-Frequency Trading, Endogenous Capital Commitment and Market Quality[1]

2.1 Introduction

Over the past decade, high-frequency trading has become increasingly prevalent worldwide. According to O'Hara (2015), high-frequency traders (henceforth HFTs) contribute more than half of market trading volume. This growing trend of high-frequency trading has led to a policy debate over proper regulatory measures to adapt to this change. Clearly, policy makers have yet to reach a consensus over this issue as different countries are implementing regulations with opposing intended effects.[2] Most European countries have carried out strict rules to reduce high-frequency trading and "level the playing field" while some Asian countries such as Japan and Singapore embrace high-frequency trading by providing systematic support including

[1]This chapter is based off of my paper (Wang (2021a)).

[2]For a comprehensive survey of the global high-frequency trading regulation environment, see Bell and Searles (2014)

introducing co-location service and rebating high-frequency trades.

Extant empirical research has documented that the presence of HFTs leads to lower spreads in the market. Some papers take this as direct evidence that high-frequency trading improves market quality.[3] There are essentially two rationales behind this claim. First, lower spreads indicate less information asymmetry. Second, lower spreads enhance market efficiency by facilitating assets moving to agents with higher valuations.

However, an implicit market clearing assumption lies behind the second claim. That is, at each instant, the asset price is determined by a centralized planner, who receives all market participants' supply and demand schedules, to clear the market. Although it is a reasonable assumption for analyzing the long-run behavior of the market, it is a strong assumption in modeling high-frequency trading for two reasons. First, since trading happens very fast, it is unlikely that each market participant has time to submit a sequence of limit orders to form a demand or supply schedule in each trade. Second, even if there is a planner with all the information, the price may not be adjusted quickly enough to clear the market at each instant. Without the market clearing assumption, the one to one link between price and quantity breaks; i.e., a lower spread level no longer indicates a larger trading volume. Specifically, facing competition from HFTs, a market maker might reduce his capacity in absorbing market imbalance as well as the spread since market making becomes less profitable. On the other hand, HFTs' abilities to provide liquidity are constrained by

[3]See Hendershott et al. (2011a), Boehmer et al. (2018), Brogaard et al. (2014), Hendershott et al. (2011a), Boehmer et al. (2018), Hendershott and Riordan (2013), Hasbrouck and Saar (2013), Brogaard et al. (2015), Conrad et al. (2015) and Conrad and Wahal (2018), among others.

market conditions and might be insufficient to fill the gap left by tte market maker. The decrease of market making capacity would lead to lower trading volume and deteriorate market quality. Indeed, Chordia et al. (2011), O'Hara et al. (2014) and Korajczyk and Murphy (2019a) show that the average order size becomes smaller and investors have difficulties executing large orders.

I consider a model where the market maker and the HFT compete to sell shares to a potential buyer in each period.[4] For clarity, I use female pronouns for the HFT and male pronouns for the market maker and the buyer. The market maker contracts with the exchange to provide liquidity and is obliged to post quotes in the market.[5] As a firm, the market maker can either commit his capital in market making, i.e, buying shares from an inter-dealer market for sale, or paying out dividend to investors.[6] The amount of capital committed in market making is endogenously determined by equalizing the marginal value of market making and the marginal value of paying dividend. When no HFT exists, the market maker is modeled as a monopolist due to the market power he enjoys from advantageous terms provided by the exchange.[7] Under this circumstance, making the market is highly profitable and the market maker commits a large amount of capital in market making.

In my model, the HFT enters the market with exogenous probability and holding. This means to capture the reality that the HFT makes profit by anticipating

[4]This model bears similarities to Kreps and Scheinkman (1983).

[5]In practice, the market maker in my model can be considered as a designated market maker in NYSE or a specialist in NASDAQ.

[6]Alternatively, I can assume that there is a risk-free asset with unlimited supply the market maker can invest in.

[7]This differs from the competitive market making assumption in Kyle (1985a) and other models in market micro-structure.

the arrival of future orders.[8] If the HFT detects a buying order, she tries to quickly buy cheaper shares from other channels and sell to the buyer at a slightly higher price. This way of operation makes the HFT's presence and the amount of shares supplied highly depend on exogenous market conditions. The competition from the HFT affects the market maker's pricing and capital commitment decisions. The market maker may tighten the spread to compete with the HFT. This reduces buyers' transaction costs and improves market quality. On the other hand, market making becomes less attractive because of the competition and the market maker would reduce his capital commitment in market making. This weakens the market's capacity to satisfy large demands and effectively leads to a shallower market.[9]

I first consider the setting where the HFT possesses superior trading technology relative to the market maker. It enables the HFT to observe both the market maker's capital commitment and spread before making her pricing decision. In other words, the market maker and the HFT set spreads sequentially. The market maker faces a trade-off. If the market maker sets a high spread aiming to achieve a high expected payoff when the HFT does not enter, upon entering, the HFT would undercut and the market maker would only receive the residual demand. If the market maker sets a low spread, he sacrifices some profit when the HFT does not enter. Yet a low spread protects the market maker from the HFT's undercut. In the steady state, the market maker posts a high (low) spread if the HFT's entry probability is low (high).

[8]There are certainly other types of high-frequency traders. For instance, some market makers nowadays adopt advanced technology for trader. In my model, they fall into the market maker category.

[9]In my model, this corresponds to the buyer leaving the market with a smaller portion of his fulfilled. In practice, this may corresponds to the buyer purchasing a large portion of shares from other liquidity providers at a higher price.

In other words, competition from the HFT has a positive price effect on market quality but reduces the return of market making. Thus, the market maker's steady state capital commitment is (weakly) decreasing in the HFT's entry probability. This deteriorates market quality. I use liquidity, the expected shares sold to the buyer, as a proxy of market quality to measure the aggregate effect of high-frequency trading. Importantly, under mild assumptions, liquidity is not changing monotonically with respect to the HFT's entry probability. This lack of monotonicity has two implications. First, using linear regression to analyze high-frequency trading's market and welfare effects may lead to erroneous conclusions. Second, past observations on high-frequency trading's effects on financial markets may not be sufficient to guide policy making, which would change the market condition faced by HFTs dramatically.

I then analyze the setting where the market maker and the HFT's trading technologies are "head to head". The HFT and the market maker in this setting set spreads simultaneously. This corresponds to realistic situations with high-frequency market making or limitations on maximum trading speed. In the equilibrium, the market maker and the HFT both use mixed pricing strategies. The market maker's expected payoffs are the same setting spreads sequentially and simultaneously. However, the HFT's expected payoff is (weakly) lower when submitting spreads simultaneously.[10] Specifically, when the HFT's entry probability is low, the HFT has incentive to acquire superior trading technology at a low cost but the market maker would have no incentive to match the technology level. This is detrimental to market quality.

[10]This is in line with the evidence in Baron et al. (2018) that faster HFTs achieve higher payoffs.

In both settings, two regimes of equilibrium (the wide spread region and the tight spread region) exist depending on the HFT entry probability. In the wide spread region with low HFT entry probability, the market maker sets a high spread and his capital commitment is decreasing in HFT entry probability. An increase in HFT entry in this region has ambiguous effects on market quality since it increases liquidity supplied by the HFT but decreases liquidity supplied by the market maker. In the tight spread region where the HFT entry probability is high, the market maker sets a low spread and his capital commitment is not changing in the HFT's entry probability. Thus, more HFT entry leads to better market quality in this region. Moreover, in the wide spread region, equalizing trading technologies of the market maker and the HFT improves market quality. This is because by switching from sequential pricing to simultaneous pricing, the average spread becomes lower while the market maker's capital commitment remains the same.

My model differs from the existing theory in two ways.[11] First, I explicitly consider the market maker's capital commitment decision, which has critical implications for market quality. Second, liquidity suppliers in my model face asymmetric constraints. Specifically, the market maker has an affirmative obligation to provide liquidity and faces a trade off between committing capital in market making and paying dividend. On the contrary, the HFT's entry and the amount of liquidity supplied (extensive and intensive margin) depend on exogenous market conditions. Although market making is profitable for the HFT, these constraints limit the HFT's ability to fill the gap when the market maker commits less capital. Contrary to conven-

[11]For examples, see Goettler et al. (2009), Budish et al. (2015), Biais et al. (2015) and Foucault et al. (2016), etc.

tional wisdom, competition does not necessarily lead to better markets when there is asymmetry among liquidity suppliers.

I further consider three extensions. In the first extension, I endogenize the HFT's entry probability by imposing a fixed high-frequency trading participation cost. The HFT needs to pay the cost to enter the market with an exogenous probability.[12] If the cost is high, the exogenous entry probability is low or the market is competitive, the HFT would rationally not participate in high-frequency trading. The market maker in this extension enjoys an additional strategic advantage. He can make the market competitive by setting a low spread to deter the HFT from participating in trading. This deterring spread is increasing with the HFT's participation cost. The equilibrium outcome depends on the magnitude of the participation cost. When the participation cost is low, market quality is the same as in the baseline model since deterring the HFT's participation is too costly for the market maker. Conversely, with a high cost, the HFT may not participate in high-frequency trading and the market maker's spread and capital commitment increase with the participation cost and eventually converge to the monopolistic levels. The overall effect of the participation cost on market quality is ambiguous. Yet it is certain that high participation cost harms the market.

The second extension considers flipping. That is, the HFT can purchase shares from the market maker and re-supply them at a higher spread. With high HFT entry probability, the market maker sets a low spread to induce flipping. When the HFT flips shares, market quality appears to be good since the expected trading volume is

[12]For example, EU's trading tax on both executed and canceled orders can be considered as a cost of this type.

high and the average spread is low. However, these indicators are not characterizing market quality faithfully under this situation for two reasons. First, most of the cheaper shares are purchased by the HFT rather than the liquidity buyer. Second, the trading volume is "double-counted". The actual volume sold to the buyer is much lower. This extension demonstrates the importance of separating trades between liquidity suppliers and trades from liquidity suppliers to other investors to avoid over-estimating market quality.

In the third extension, the market maker can post a supply schedule to sell shares at different spreads.[13] With no HFT, the market maker chooses to sell all shares at the monopolistic spread. However, facing competition from the HFT, the market maker would sell shares at a continuum of spreads. I characterize conditions that determine the market maker's pricing strategy and capital commitment at the steady state and discuss implications for market quality. Furthermore, this extension illustrates how competition between the market maker and the HFT determines the shape of limit order book.

My model contributes to the theoretical literature on high-frequency trading by exploring how high-frequency trading affects market quality via the capital commitment channel. Competition from the HFT leads the market maker to commit less capital in market making. This effect dampens the benefit brought by pricing competition, and, if large enough, the presence of a potential HFT might even deteriorate market quality. Ait-Sahalia and Saglam (2017a) and Han et al. (2014) also consider market quality implications with competition between the HFT and the

[13]In the baseline model, I assume the market maker has to sell all shares at one spread.

market maker. However, in these papers, the size of orders is fixed. This assumption constrains these models' abilities to capture how capital commitment of the market maker affects market quality.[14] In my model, it is possible that a market with wide spread has better quality than a market with tight spread. The reason is that in the latter market, the market maker commits much less capital in market making.

The implications of my model are consistent with the following empirical findings in the literature: (1) High-frequency trading leads to lower average spreads in the market; (2) the average trade size becomes smaller; (3) market makers commit less capital in market making; (4) Large orders might face higher trading costs with the presence of HFTs; (5) market quality improves when all market participants have similar trading speeds. My model also provides several insights for future studies. First, the price information alone does not provide a complete characterization of market quality. The volume information is equally important. Second, market quality may not change monotonically with increasing HFT presence. In this sense, we cannot only rely on linear regression for accurate welfare implications of high-frequency trading. Third, when the HFT can flip orders, it is important to differentiate trades between liquidity providers and trades from liquidity providers to other investors. Otherwise, the data cannot faithfully reflect market quality since HFTs would exploit most of the cheaper orders with superior trading technology.

This paper also generates important insights for HFT regulations. The model suggests that if high-frequency trading is prevalent in the market, encouraging high-

[14]In Ait-Sahalia and Saglam (2017a), the HFT, as a long run market maker, also holds inventory. However, since the supply is fixed to one, the inventory does not have a quantity effect. Instead, it has a price effect due to the inventory aversion assumption.

frequency trading benefits liquidity. On the other hand, when high-frequency trading is less prevalent, more HFT's presence drives out the market maker's capital and has ambiguous effects on market quality. Second, when the HFT's entry probability is low, equalizing the trading speeds of the HFT and the market maker improves market quality. When the HFT's entry probability is high, it benefits mid-valuation buyers yet hurts low-valuation buyers. I also analyze implications of high-frequency trading participation cost. A low participation cost does not affect the market quality while a high participation cost increases market maker's capital commitment but also drives up the spread. The aggregate effect is ambiguous.

Finally, my paper complements the literature of limit order book formation by illustrating the effect of asymmetric competition between the market maker and the HFT over limit order book shape.[15] Specifically, with no HFT, the market maker would sell all shares at the monopolistic spread. Facing the competition from the HFT, the market maker sets a non-degenerate limit order book to avoid the HFT's undercutting. My model predicts a downward sloping limit order book at lower spreads and a large volume supplied at the monopolistic spread.

The rest of the paper is organized as follows. Section 2.2 reviews related literature. Section 2.3 presents baseline models. Section 2.4 analyzes baseline models. Section 2.5 considers the costly participation extension. Section 2.6 uses results developed in Sections 2.3, 2.4 and 2.5 to discuss market quality implications of various high-frequency trading regulations. Section 2.7 considers the flipping extension. Section

[15]For other papers on limit order book formation, see Glosten (1994), Chakravarty and Holden (1995), Seppi (1997), Biais et al. (2000), Viswanathan and Wang (2002) Parlour and Seppi (2003), Foucault et al. (2005), Roşu (2009), Back and Baruch (2013), Baruch and Glosten (2019), etc.

2.8 considers the extension in which the market maker can submit supply schedules. Section 2.9 concludes.

2.2 Related Literature

2.2.1 HFT Behavior

An existing theory literature analyzes how high-frequency trading effects market quality from the information perspective.[16] Han et al. (2014) demonstrate how adverse selection problem arising from fast order cancellation leads to wide spreads when the HFT enters the market with probability between 0 and 1. Budish et al. (2015) show how mechanical arbitrage in high-frequency time horizon hurts liquidity and propose frequent batch auctions mechanism as a solution. Biais et al. (2015) endogenize investment decisions on fast trading technology and show that equilibrium investment level on fast trading is higher than the social optimal level because high-frequency trading has a negative externality. Van Kervel (2015) analyzes the link between high-frequency trading and order cancellations across trading venues. Foucault et al. (2016) analyzes news trading by fast speculators and its implications for trading volume and asset price. Ait-Sahalia and Saglam (2017a) and Ait-Sahalia and Saglam (2017b) analyze high-frequency market making and show that the faster market maker provides more liquidity. Baldauf and Mollner (2019) consider the informational implication of high-frequency trading and conclude that the bid-ask

[16]For a comprehensive survey, see Menkveld (2016).

spread narrows yet the information production also diminishes. Budish et al. (2019) consider a model where exchanges capture economic rents by selling speed technologies to discuss exchanges' incentives to adopt new market designs. Li et al. (2020) model competition between slower execution algorithms and high-frequency traders featuring the implication of the tick size. My model differs from the existing literature by explicitly considering the market maker's capital commitment decision facing competition from HFT and its implications for market quality.

Many empirical papers test high-frequency trading's impact on liquidity. Research generally documents an increase in liquidity with high-frequency trading. For instance, Hendershott et al. (2011a), Hendershott and Riordan (2013), Hasbrouck and Saar (2013), Conrad et al. (2015) and Conrad and Wahal (2018),[17] using spread as a proxy for liquidity, conclude that liquidity is improved by high-frequency trading. Brogaard et al. (2014), using order flow data, conclude that HFT is liquidity improving around macroeconomic news since liquidity supply is greater than liquidity demand. Boehmer et al. (2018) using execution shortfalls as a proxy, reach the similar conclusion. My model does not contradict these evidences. However, it does suggest that some important quantity aspects of market quality cannot be captured by these proxies. Specifically, spread measures might not capture the quantity information related to the market maker's capital commitment. The execution shortfall can better capture the price change facing large demand. Yet even the execution shortfall does not incorporate information about unexecuted and canceled orders.

[17]Hasbrouck and Saar (2013) also examines number of shares displayed on the order book as a proxy for depth. One concern is that since HFTs can cancel orders with fast speed, this "NearDepth" might not able to capture real market depth.

Moreover, order flow as a proxy of liquidity often includes trades between HFTs. This might lead to an over estimate of market quality. The extension on flipping directly addresses this concern. Recently, Korajczyk and Murphy (2019b) and Korajczyk and Murphy (2019a) document that less high-frequency trading is associated with higher transaction costs for small trades and lower transaction costs for large trades. Hu (2019) shows that market quality improves when IEX, an institute implementing a trading speed bumps to all participant, became a national securities exchange. These findings are in line with predictions in my model.

Some empirical papers focus on characteristics of traditional market makers and HFTs. Kirilenko et al. (2017) document that, different from traditional market makers, HFTs behaviors during the flash crash are more consistent with the latency arbitrage theory. Hirschey (2018) shows that HFTs can anticipate and trade ahead of other investors' order flow. Baron et al. (2018) find that faster HFTs gain higher payoffs. This is in line with the prediction of my model that small HFTs has incentive to upgrade trading technology to be able to undercut the market maker. Van Kervel and Menkveld (2019) document that HFTs initially lean against institutional orders but eventually trade along long-lasting orders since they are likely to be information-motivated.[18] Yao and Ye (2018) document that HFTs provide more liquidity for stocks with higher relative tick size. Clark-Joseph et al. (2017) use data of two trading halts to show that designated market makers' participation has important liquidity implications. This clearly shows that designated market makers and HFTs operate on

[18]This finding is consistent with my assumption that the HFT acts as a liquidity provider. However, my model is silent on the HFT trading alone the information-motivated orders since my model does not consider informed trading.

different business models. Bessembinder et al. (2019) also highlight the importance of designated market makers by showing that an improving of contract terms for designate market makers in NYSE improves market quality. This is consistent to the prediction of my model. If the market maker receives extra rebate on each share, he will commit more capital in market making and posts a lower spread.[19]

2.2.2 Capital Constraint and Capital Commitment

Many models explore the link between capital constraints of intermediaries and liquidity provision. Kyle and Xiong (2001) describe the situation that when convergence traders lose capital, their liquidation leads to excess volatility and more correlation among different markets. Gromb and Vayanos (2002) show that constrained arbitrageurs might provide too much or too little liquidity compare to the social optimal level, depending on their initial investment positions. Weill (2007) and Brunnermeier and Pedersen (2008) both demonstrate that insufficient capital of the market maker would lead to lower liquidity provision then the optimal level. In Weill (2007), lack of capital prevents the market maker to absorb enough order imbalance when the economy is recovering from a negative shock. In Brunnermeier and Pedersen (2008), traders' lack of funding and market liquidity deterioration reinforce each other and let to "liquidity spiral". My paper contributes to this strand of literature by showing that, even when the market maker is not constrained, his capital commitment decision plays an important role to market quality when facing competition from

[19]Bessembinder et al. (2019) also document the spillover effect in market quality improvement because of the strategic complementary effect in market making. My model is silent on this aspect because I assume a deep inter-dealer market.

high-frequency trading.

A relatively small empirical literature examines the capital commitment of market makers. Hameed et al. (2010) show that negative market return decreases liquidity asymmetrically. The authors attribute the decrease to the market maker's capital constraint. Comerton-Forde et al. (2010) find a similar result using data on NYSE specialist positions and revenues. Bessembinder et al. (2018) document that capital commitments of corporate bond dealers are decreasing overtime, specifically in markets with more electronically facilitated trades. The authors interpret this as a result of electronic trading reducing search cost and required capital. This model suggests an alternative explanation. The decrease of capital commitment might due to the growing entry of HFTs facilitated by electronic trading. Brogaard and Garriott (2019) document similar capital commitment decreases of market makers in the stock market.

2.3 Model Setting

2.3.1 The Setup

Consider a game with infinite many periods and three (kinds of) players: a long-run market maker, a short-run HFT and a short-run buyer. The market maker's discount rate is δ and has net worth w_0 in period 0. In each period, the market maker can either pay dividend d or acquire shares from a inter-dealer market at the fair price

1 for market making.[20] The market maker maximizes $E_0(\sum_{t=0}^{\infty} \delta^t d_t)$, the expected dividend payout. In each period, a short-run HFT enters the market with probability π. Upon entering, the HFT holds q_h shares and aims at maximizing her expected profit. The market maker and the HFT are both sellers and compete to provide liquidity for the short-run buyer. Due to liquidity or hedging needs, the buyer is willing to pay $v > 1$ for each share and demands q_b shares; i.e., he is willing to pay a premium $v - 1$ for each share within his demand q_b.

The sequence of events in a single period, illustrated in Figure 2.1, can be specified as follows: Let w_t be the market maker's net worth at the beginning of period t. The market maker first chooses a non-negative dividend level d_t. He then commits the remaining capital, $w_t - d_t$, to purchase $q_{m,t} = w_t - d_t$ shares from the inter-dealer market at the fair price 1.[21] The market maker then posts a spread $x_{m,t}$, committing to sell all shares at the ask price $1 + x_{m,t}$. After the market maker sets his spread, a short-run HFT holding q_h shares enters the market with probability π. If the HFT's trading technology is superior to the market maker, she observes the market maker's capital commitment $q_{m,t}$ and spread $x_{m,t}$ before setting her spread x_h (the sequential pricing game). Otherwise, the HFT only observes the market maker's capital commitment $q_{m,t}$ (the simultaneous pricing game). After the market maker and the HFT determine their spreads, the short-run buyer arrives with random demand q_b and random buying threshold $v > 1$. After the buyer finishes buying, the

[20] Another interpretation can be that the market maker invests some capital into a safe asset and deposits the rest of capital into a margin account to cover the cost of potential short selling.

[21] It is without loss of generality to assume that the market maker commits all remaining net worth in market making. If he chooses to commit less, he may raise his dividend payout to achieve a higher payoff.

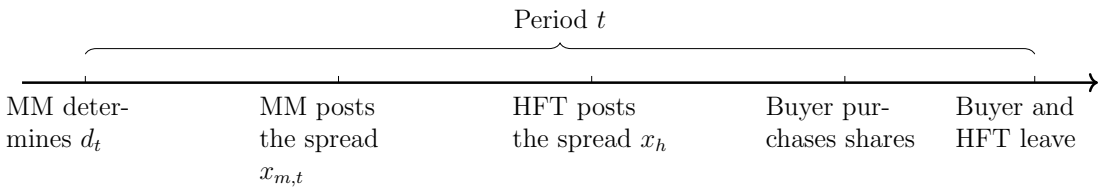


Figure 2.1: A Concise Time-line

market maker and the HFT (if enters) may sell the remaining shares at the fair price 1 back to the inter-dealer market. This concludes a period.

I make the following assumptions on the distributions of q_b and v: $v-1$ follows a distribution supported on $[0, \hat{x}]$ with CDF F. q_b follows a distribution with finite expectation, positive support and CDF G. F and G are independent and continuously differentiable. I further assume that F has non-decreasing hazard rate; i.e., $\frac{f(x)}{1-F(x)}$ is non-decreasing, or equivalently, f is log-concave.

Several specific assumptions are worth more discussion. First, the buyer's demand q_b is inelastic when spreads are lower than $v-1$. In practice, this corresponds to the buyer posting a limit order with quantity q_b at price v. Since the market maker and the HFT do not observe q_b when setting spreads, higher spreads reduce the probability of trade. Thus, although the demand curve of each buyer is inelastic, from the market maker's and the HFT's perspectives, the demand curve is downward sloping. Second, in this model, the HFT is a short-run player with an exogenous entry probability π and a fixed shareholding q_h. This assumption by no means denies the possibility of the HFT being a long term market participant in practice. Instead, it means to reflect two features of high-frequency trading: (1) The HFT's entry and holding heavily depend on exogenous market conditions; (2) the HFT focuses on

short term trading and only carries positions for a short period of time. Third, I only consider a one-sided market; i.e., the market maker and the HFT only sell shares to other investors. This is without loss of generality given that the market maker can adjust his position with no cost in the inter-dealer market. Considering a two-sided market setting leads to similar qualitative predictions.

2.3.2 Measures of Market Quality

Liquidity is one of the most important indicators of market quality. In this section, I define liquidity, the main measure of market quality in this model, and briefly discuss other market quality measures. Formally, L_t, liquidity in period t, is the expected number of shares sold to the buyer in period t. Since I focus on the steady state, where the market maker's pricing and capital commitment decisions are time invariant, I drop the time subscript and define liquidity (in the steady state) to be

$$L = \pi \underbrace{E(min(q_b, q_m I_{\{x_m \leq v-1\}} + q_h I_{\{x_h \leq v-1\}}))}_{\text{Expected selling volume with HFT}} + (1-\pi) \underbrace{E(min(q_b, q_m I_{\{x_m \leq v-1\}}))}_{\text{Expected selling volume without HFT}}.$$

Further define $L(v)$ to be the expected number of shares sold to the buyer with buying threshold v. It captures the market's capacity to satisfy buyers with buying thresholds higher than v. Specifically, the fill rate at buying threshold v can be measured by $L(v)/E(q_b)$. It is also worthwhile to examine the average spread. Define it to be the expected profit of liquidity providers (the market maker and the HFT) divided by liquidity.

Several features of this liquidity definition worth discussing. First, this definition

incorporates both price and quantity information of the market. If spreads are high, the buyer's buying probability would be low. Then even with a large supply, liquidity would be low due to the lack of buyer. On the other hand, low spreads alone does not imply high liquidity. If the aggregate supply is small due to the low profit margin, liquidity would still be low since only a small portion of the buyer's demand can be satisfied. Second, this measure is closely related to (the buyer's) welfare.[22] Since the buyer has a higher valuation for each share, holding everything else equal, higher liquidity indicates better welfare. This definition differs from the buyer's surplus, by putting equal weights on each share sold. These two measures bear similarity in the sense that the buyer's surplus and liquidity almost always change in the same direction in the comparative statics. Moreover, liquidity is more feasible than the buyer's surplus as a market quality measure since trading volume is easier to observe in practice.

2.3.3 Equilibrium Definition

Two facts suggest that the market maker's net worth, w, should be considered as the state variable. First, net worth constraint is the only constraint faced by the market maker. Second, given the market maker's strategy, the HFT has no incentive to relate her action to the history of the game. Thus, equilibrium can be defined as follows:

Definition 1. *Consider a infinite horizon game* (w_0, q_h, π) *where the market maker*

[22]Here I follow the tradition of the literature by regarding the market maker and the HFT as integrated parts of the financial market and focusing on the buyer's welfare. The discussion on the social planner's problem is delegated to the Appendix.

starts with net worth w_0 and the HFT enters the market with probability π and q_h shares.

1. *An equilibrium in a sequential pricing game is a triple (q_m, x_m, x_h) such that: (i) Given q_m and x_m, $x_h(q_m, x_m)$ maximizes the expected payoff of the HFT. (ii) Given $x_h(q_m, x_m)$, $\{q_{m,t} = w_t - d_t\}_{t=0}^{\infty}$ and $\{x_{m,t} = x_m(q_{m,t})\}_{t=0}^{\infty}$ maximize $E_0(\sum_{t=0}^{\infty} \delta^t d_t)$.[23] (iii) $0 \leq q_{m,t} \leq w_t$ for all t.*

2. *An equilibrium in a simultaneous pricing game is a triple (q_m, x_m, x_h) such that: (i) Given q_m, $x_h(q_m)$ maximizes the expected payoff of the HFT. (ii) Given $x_h(q_m)$, $\{q_{m,t} = w_t - d_t\}_{t=0}^{\infty}$ and $\{x_{m,t} = x_m(q_{m,t})\}_{t=0}^{\infty}$ maximize $E_0(\sum_{t=0}^{\infty} \delta^t d_t)$. (iii) $0 \leq q_{m,t} \leq w_t$ for all t.*

I focus on the steady state capital commitment and spread to characterize the long term market quality. The formal definition of a steady state equilibrium is as follows:

Definition 2. *An equilibrium is a steady state equilibrium if there exists q_m, x_m and x_h such that $q_{m,t} = q_m$, $x_{m,t} = x_m$ and $x_{h,t} = x_h$ for all t.*[24]

Intuitively, in a steady state equilibrium, the market maker's capital commitment $q_{m,t}$, spread $x_{m,t}$ and the HFT's spread $x_{h,t}$ are time invariant. Since the focus of this paper is on capital commitment rather than capital constraint, I assume that the market maker always starts the game with a sufficiently large net worth w_0.

[23] Notice that the distribution of w_{t+1} can be uniquely determined by w_t and the equilibrium strategies. Given w_0, the dynamic of w_t is well-defined.

[24] In the simultaneous moving game, x_m and x_h might be distributions rather than numbers.

2.4 Baseline Models

2.4.1 Benchmark Case with No HFT

First consider the situation with no HFT (or equivalently, $\pi = 0$). The market maker's value function satisfies the following equation:

$$\begin{aligned} V(w) = max_{d,x_m} d + \delta \underbrace{F(x_m)V(w-d)}_{\text{Buyer not buying}} \\ +\delta[\underbrace{(1-F(x_m))\int_0^{w-d} V(w-d+x_m q)g(q)dq}_{\text{Buyer buying with small demand}} \\ +\underbrace{(1-F(x_m))(1-G(w-d))V((1+x_m)(w-d))}_{\text{Buyer buying with large demand}}] \end{aligned} \tag{2.1}$$

with the budget constraint

$$0 \leq d \leq w \ . \tag{2.2}$$

Let

$$k(s) = E_G(min(q_b, s)) \ .^{25}$$

This function measures expected trading volume when s shares are within the buying threshold. Since the buyer's demand is random, this function is strictly concave and the marginal value of capital commitment is decreasing to zero.[26] In other words, at

[25]The subscript emphasizes that the expectation is over the buyer's demand.

[26]Alternatively, if q_b is deterministic, when the market maker's capital commitment is lower than q_b, at any fixed spread, each unit of capital committed has the same marginal value. Then the capital commitment problem become trivial since the market maker would choose to commit either q_b or 0.

a certain point, the market maker would find it more profitable to payout dividend. Specifically, there exists a steady state equilibrium with the capital commitment $q_m = \bar{q}$ and the spread $x_m = x^*$. The market maker pays out $w_0 - \bar{q}$ in period 0 and his profit in subsequent periods as dividend. The steady state can be characterized by the following theorem:

Theorem 1. *With no HFT, there exists a unique steady state equilibrium where the market maker set* $q_{m,t} = \bar{q}$ $(d_t = w_t - \bar{q})$ *and* $x = x^*$ *for all* t. x^* *satisfies*

$$x^* = argmax_x(1 - F(x))x \ .$$

$\bar{q}$ *satisfies*

$$\frac{\delta}{1-\delta}(1 - F(x^*))x^*(1 - G(\bar{q})) = 1 \ .^{27}$$

The market maker's expected payoff is

$$V(w_0) = \frac{\delta}{1-\delta}(1 - F(x^*))x^*k(\bar{q}) + (w_0 - \bar{q}) \ .$$

Liquidity at the steady state is

$$L = (1 - F(x^*))k(\bar{q}) \ .$$

The average spread is x^*.

Proof. See Appendix. □

[27] If no such $\bar{q}$ exists, the optimal strategy is to liquidate $(d = w_0)$ at $t = 0$.

This theorem has a clear economic interpretation. Since the buyer's demand is random, each additional share is less likely to be sold at any given spread. Thus, the market maker's capital commitment has decreasing marginal value. Conversely, the marginal value of dividend payout is constant. This implies that in the equilibrium, the market maker would commit capital up to a unique level where the marginal value of capital commitment equals the marginal value of dividend payout. In the steady state, the market maker maintains his capital commitment level and pays out the profit. This makes him act like a short-run monopolist, setting the spread to maximize the expected profit.

With no HFT, the market has high supply at a high spread. Indeed, from the pricing perspective, x^* is the highest possible spread set by any rational liquidity supplier.[28] From the capital commitment perspective, the marginal value of capital commitment is the highest for the market maker facing no competition from the HFT. With HFT's presence, the market maker's steady state capital commitment is always lower than $\bar{q}$.

2.4.2 Sequential Pricing Game

In the sequential pricing game, the HFT observes the market maker's shareholding q_m and spread x_m before posting her spread x_h. In practice, this corresponds to the situation where the HFT has a superior trading technology and can undercut the market maker before the market maker is able to adjust the spread.

[28] At a spread higher than x^*, the loss from selling with lower probability dominates the benefit from selling at a higher price.

To characterize the steady state, it is helpful to first consider a one-shot game with fixed capital commitment/shareholding. The reason is clear: In the steady state, the market maker's capital commitment is constant over time and he pays out his profit from the previous period as dividend. Thus, in the steady state, the market maker would set spread as if he is a one-shot profit maximizer.

Consider a one-shot game (q_m, q_h, π) where the market maker holds q_m shares and the HFT enters with probability π holding q_h shares. The market maker sets spread x_m first and the HFT, if enters, sets spread x_h after observing x_m. Each player aims for maximizing his/her expected profit and can sell shares back to the inter-dealer market at the end of the game at price 1. Equilibrium of this one-shot game can be defined as follows:

Definition 3. *An equilibrium of a one-shot sequential pricing game (q_m, q_h, π) is a pair $(x_m, x_h(x_m))$. Given the market maker's spread x_m, the HFT's spread $x_h(x_m)$ maximizes her expected payoff. Given the HFT posting her spread according to $x_h(x_m)$, the market maker's spread x_m maximizes his expected payoff.*

First consider the HFT's pricing problem. If the HFT sets her spread $x_h \leq x_m$, her shares would be purchased first but at a lower price. Conversely, if $x_h > x_m$, the HFT would earn higher profit per share sold. Yet she would only receive the residual demand. Within each pricing region ($x_h \leq x_m$ or $x_h > x_m$), the HFT only faces the trade-off between earning higher unit profit and losing the buyer with low valuation.[29] This trade-off is characterized by the term $(1 - F(x))x$, the expected

[29]Notice that for a buyer with buying threshold higher than $1 + x_h$, his expected demand does not depend on x_h. This relies on the independence assumption of the buyer's buying threshold and demand.

marginal value of supplying a share at spread x within the buyer's demand. Since I assume F, the CDF of the buyer's buying threshold, has non-decreasing hazard rate, $(1 - F(x))x$ is increasing in x for $x \leq x^*$ and decreasing in x for $x \geq x^*$. This simplifies the HFT's optimal pricing strategy.

Lemma 1. *Given the market maker's capital commitment q_m and spread x_m, the HFT's optimal spread is either $x_h = x_m$ or $x_h = x^*$.*

Proof. See Appendix. □

Next consider the market maker's pricing problem. If the market maker sets a spread such that the HFT chooses $x_h = x_m$, the market maker would be better off setting $x_m = x^*$. If the market maker sets a spread such that the HFT chooses $x_h = x^*$. Since $(1 - F(x))x$ is increasing in x for $x \leq x^*$, the market maker would optimally set $x_m = \underline{x} < x^*$ such that the HFT is indifferent between setting $x_h = \underline{x}$ and setting $x_h = x^*$. All other pricing strategies are dominated by either of the aforementioned two strategies. To simplify the notation, define

$$a(x) = \frac{(1 - F(x))x}{(1 - F(x^*))x^*} \leq 1 \ .^{30}$$

The following lemma characterizes the market maker's optimal pricing strategy:

Lemma 2. *The market maker's optimal spread is either $x_m = x^*$ or $x_m = \underline{x} < x^*$. $\underline{x}$ is pinned down by the HFT's indifference condition*

$$a(\underline{x})k(q_h) = k(q_m + q_h) - k(q_m) \ .$$

[30]Note that $x^* = argmax_x(1 - F(x))x$.

Proof. See Appendix. □

By Lemma 1 and 2, I can pin down the equilibrium by comparing the market maker's payoffs with pricing strategies $x_m = \underline{x}$ and $x_m = x^*$.

Proposition 1. *If $k(q_m) > \pi k(q_h)$, the unique equilibrium is $x_m = x_h = x^*$. If $k(q_m) < \pi k(q_h)$, the unique equilibrium is $x_m = \underline{x}$ and $x_h = x^*$. When $k(q_m) = \pi k(q_h)$, both equilibria exist.*

Proof. See appendix. □

By Proposition 1, the market maker has two possible pricing strategies against the potential HFT. I name $x_m = x^*$ to be the wide spread strategy. This strategy yields a high expected profit without HFT presence. When the HFT enters, however, the market maker will be undercut and only receives the residual demand. The effectiveness of this strategy depends on the HFT's entry probability π and shareholding q_h. $x_m = \underline{x}$ is the tight spread strategy. Under this strategy, the market maker receives a lower expected profit when the HFT does not enter. Yet when the market maker uses the tight spread strategy, it is unprofitable for the HFT to undercut the market maker. Thus, the buyer would always buy shares from the market maker first and the HFT's entry does not affect the market maker's expected profit.

Another observation is that the HFT always sets spread $x_h = x^*$ in the equilibrium. However, this does not imply that the HFT always sells shares at a higher spread. When the market maker is using the wide spread strategy, the HFT's pricing strategy should be understood as undercutting the market maker at $x_h = x^* - \epsilon$ with

an infinitesimally small ϵ.[31]

Steady State Characterization

In this section, I solve for the steady state equilibrium of the infinite period game. Let $M(q)$ be the market maker's expected profit in the one-shot game with $q_m = q$. Let $\hat{x}_m(q)$ and $\hat{x}_h(q)$ correspond to the market maker and the HFT's equilibrium spreads in the one-shot game.[32] If the game reaches a steady state in period 0 with capital commitment q, the market maker's expected payoff is

$$\frac{\delta}{1-\delta}M(q) + (w_0 - q) \ .$$

$\frac{\delta}{1-\delta}M(q)$ is the present value of a perpetuity paying out the market maker's expected profit starting from period 1. $w_0 - q$ is the market maker's dividend payout in period 0 to reach the steady state. An obvious candidate of the market maker's steady state capital commitment is

$$q_m = argmax_{q\in[0,\bar{q}]}\frac{\delta}{1-\delta}M(q) + (w_0 - q) \ .$$

The following theorem validates that q_m is indeed the market maker's capital commitment in the steady state equilibrium.

Theorem 2. *Let* $q_m = argmax_{q\in[0,\bar{q}]}\frac{\delta}{1-\delta}M(q) + (w_0 - q)$.

[31] If a minimum tick size exists, then the HFT would post a lower spread than the market maker in this situation.

[32] I suppress the dependency of these functions on q_h and π.

1. *$q_{m,t} = q_m$, $x_m = \hat{x}_m(q_m)$, $x_h = \hat{x}_h(q_m)$ consists a steady state equilibrium. The market maker's expected payoff in the equilibrium is*

$$V(w_0) = \frac{\delta}{1-\delta} M(q_m) + (w_0 - q_m) \ .$$

2. *If the market maker uses the wide spread strategy in the equilibrium, market liquidity is*

$$L = (1 - F(x^*))[\pi k(q_m + q_h) + (1 - \pi)k(q_m)] \ .$$

   The average spread is x^*.

3. *If the market maker uses the tight spread strategy in the equilibrium, market liquidity is*

$$L = (1 - F(x_m))k(q_m) + \pi(F(x^*) - F(x_m))(k(q_m + q_h) - k(q_m)) \ .$$

   The average spread is lower than x^*.

Proof. See appendix. □

I now discuss some important corollaries.

Corollary 1. *For $\pi > 0$, $q_m < \bar{q}$.*

Corollary 1 states that the market maker commits less capital facing competition from the HFT. The competition reduces the marginal value of the market maker's capital commitment. This is either due to the HFT's undercut or the market maker

using a lower spread. Lower marginal value leads to less capital commitment in the equilibrium.

Corollary 2. *If $\bar{q} > 0$, $q_m > 0$. In other words, the market maker never fully exit the market in the steady state equilibrium. Moreover, q_m, the market maker's steady state capital commitment, satisfies the following conditions:*

1. *If the market maker uses the wide spread strategy,*

$$\frac{\delta}{1-\delta}(1-F(x^*))x^*[(1-\pi)(1-G(q_m))+\pi(1-G(q_m+q_h))] = 1 \ .$$

2. *If the market maker uses the tight spread strategy,*

$$\frac{\delta}{1-\delta}(1-F(\underline{x}))\underline{x}(1-G(q_m)) > 1 \ .$$

Proof. See appendix. □

Corollary 2, derived from first order conditions of the market maker, is useful for comparative statics in π. Using the wide spread strategy, the market maker's marginal value of committing capital equals to the marginal value of dividend payment fixing x_h. On the contrary, under the tight spread strategy, the market maker's marginal value of committing capital is larger fixing x_h. This means using the tight spread strategy, the market maker refrains from committing more capital because the market maker needs to maintain a low spread to prevent the HFT from undercutting.

Comparative Statics on π

In this section, I analyze how the steady state equilibrium and market quality change with π, the HFT's entry probability. Higher π indicates fiercer competition from the HFT. The market maker would adjust his capital commitment and pricing strategies accordingly and thus changes market quality.

First, consider the one-shot game. Importantly, the HFT's pricing decision does not depend on π. Thus, regardless of π, the market maker's candidates for the optimal spread, i.e., x^* and $\underline{x}$, are the same. Furthermore, if $x_m = x^*$, the market maker's expected payoff is decreasing in π due to the HFT's undercut. Conversely, the market maker's expected payoff does not depend on π when $x_m = \underline{x}$. Consequently, the tight spread strategy becomes more attractive with higher HFT entry probability. The comparative statics for one-shot games can be characterized by the following proposition:

Proposition 2. *Consider two one-shot games (q_m, q_h, π_1) and (q_m, q_h, π_2) with $\pi_2 > \pi_1$.*

1. *If the market maker adopts the tight spread strategy in the equilibrium in game (q_m, q_h, π_1), then he would also adopt the tight spread strategy in game (q_m, q_h, π_2). His expected profits in two games are the same.*

2. *If the market maker adopts the wide spread strategy in the equilibrium in game (q_m, q_h, π_2), then he would also adopt the wide spread strategy in game (q_m, q_h, π_1). His expect payoff is higher in game (q_m, q_h, π_1).*

Proof. Since q_m and q_h are fixed, the equilibrium strategy choices are implied by proposition 1.

Note that the tight spread $\underline{x}$ is determined by the equation

$$k(q_h + q_m) - k(q_m) = a(\underline{x})k(q_h) \ ,$$

which does not depend on π. Thus, the market maker's expected payoff when adopting the tight spread strategy, $(1 - F(\underline{x}))\underline{x}k(q_m)$, does not depend on π.

The market maker's expected net profit of adopting the wide spread strategy is

$$(1 - F(x^*)x^*)[\pi(k(q_h + q_m) - k(q_h)) + (1 - \pi)k(q_m)] \ .$$

This quantity is decreasing in π since

$$k(q_h + q_m) < k(q_h) + k(q_m) \ .$$

□

Now consider the infinite period game in two markets with different HFT entry probabilities. If the market maker uses the tight spread strategy in both markets, then he would make identical pricing and capital commitment decisions and enjoy the same expected payoffs. On the other hand, by Corollary 2, if the market maker uses the wide spread strategy in both markets, in the market with high HFT entry probability, he commits less capital and achieves a lower expected payoff. Combining these observations leads to the following result:

Theorem 3. *There exists $\hat{\pi} \in (0, 1]$ such that in the steady state equilibrium, $x_m = x^*$ when $\pi < \hat{\pi}$ and $x_m = \underline{x}$ when $\pi > \hat{\pi}$. Denote $[0, \hat{\pi})$ to be the wide spread region and $(\hat{\pi}, 1]$ to be the tight spread region.*

1. *In the wide spread region, the market maker's expected payoff $V(w_0)$ and equilibrium capital commitment q_m is decreasing in π; liquidity L's change in π is ambiguous.*

2. *In the tight spread region, the market maker's expected payoff $V(w_0)$ and equilibrium capital commitment q_m remain constants; Liquidity L is increasing in π.*

3. *The market maker's equilibrium capital commitment is smaller in the tight spread region comparing to any equilibrium capital commitment in the wide spread region.*

4. *In the wide spread region, the average spread is x^*. In the tight spread region, the average spread is lower than x^* and increasing in π.*

Proof. See appendix. □

By Theorem 3, the steady state equilibrium can be categorized into two regimes depending on π. In the wide spread region, the market maker sets the monopolistic spread $x_m = x^*$ and responds to the competition by cutting capital commitment. In this region, the competition between the market maker and the HFT does not benefit low-valuation buyers since both the market maker and the HFT set the monopolistic spread. Instead, when the HFT enters the market, she improves market

quality by increasing the market's capacity to satisfy high-valuation buyers' demands. Conversely, when the HFT does not enter, the market's capacity to satisfy large demand is lower and decreasing in π since the market maker's capital commitment is decreasing in π in this region.

In the tight spread region, low-valuation buyers benefit from the competition since the market maker's spread is lower then the monopolistic spread. However, to deter the HFT from undercutting, the market maker keeps his capital commitment at a lower level. This impairs the market's capacity to satisfy large demands and the market becomes shallower. Indeed, although shares become cheaper, the supply is limited. When the buyer's demand is large, either the price per share would jump to the monopolistic price with the HFT's presence or no enough supply exists to fulfill the order.[33] Moreover, in this region, an increase in the HFT's entry probability improves market quality since the market maker's capital commitment and spread are not changing in π. A higher HFT entry probability increases the market's capacity to satisfy buyers with large demands.

This theorem also demonstrates why the average spread and the implementation shortfall may fail to faithfully characterize market quality. Since higher average spread indicates higher implementation shortfall in this model, I only focus on the average spread in the following discussion. In the wide spread region, although liquidity (and thus the buyer's welfare) is changing with π, the average spread remains the same since both the market maker and the HFT set the monopolistic spread x^*. In the tight spread region, higher π leads to better market quality. Yet the average

[33]In this model, I do not consider other liquidity providers. Yet in reality it can be the case that the rest of the order are fulfilled by other suppliers at a higher price.

spread is also increasing because the HFT's spread is higher. With a higher HFT entry probability, a larger proportion of shares are sold at the higher spread. This drives up the average spread.

How liquidity changes with π is ambiguous in the wide spread region. Under some mild assumptions, more competition from the HFT is not always beneficial to the market. When the wide spread region is large enough, there is always a region where the liquidity is decreasing with the level of competition.

Proposition 3. *Suppose the wide spread region is* $[0, 1]$*; i.e., the market maker uses the wide spread strategy when* $\pi = 1$*. Then either there exists a region where* L *is strictly decreasing in* π *or* L *is constant over* $[0, 1]$*.*

Proof. See appendix. □

The reason behind this result is simple. If the market maker uses the wide spread strategy at $\pi = 1$, from the first order condition, $q_m + q_h = \bar{q}$. In other words, from a buyer's perspective, the market is identical to the monopolistic market and thus has the same liquidity. By continuity, if L is not constant over π, there exists a region where L is strictly decreasing in π. Importantly, when the HFT's shareholding q_h is small, the assumption of Proposition 3 holds. Intuitively, with low q_h, the HFT's undercut is not much of a concern for the market maker. It is optimal for the market maker to set the monopolistic spread regardless of the HFT's entry probability. Thus, when q_h is low enough, there is always a region where liquidity is decreasing in π.

Assumptions may also be imposed on the distribution of the buyer's demand q_b. If G follows the exponential distribution (which has constant hazard rate), liquidity

is not changing in π over the wide spread region. If G has increasing hazard rate (or equivalently, g is log-concave),[34] there always exists a region where L is decreasing in π.

Proposition 4. *If G follows an exponential distribution, liquidity is a constant with respect to π in the wide spread region.*

Proof. See appendix. □

The discussion above leads to the following theorem regarding the non-monotonicity of liquidity L over the level of competition π:

Theorem 4. *If G has increasing hazard rate or q_h is small, L is non-monotonic with respect to π on $[0, 1]$.*

Proof. See appendix □

This theorem, albeit simple, bears important implications for both empirical analysis and policy debate over high-frequency trading. In many high-frequency trading empirical research, when market quality as a welfare indicator is the dependent variable, there is an independent variable highly correlated to the HFT entry probability. For example, it can be high-frequency trading volume, frequency of order submission and cancellation, etc. If liquidity, as a measure of market quality, is not changing monotonically with respect to the HFT entry probability, the linear regression model might not deliver accurate prediction over high-frequency trading's effects over market quality.

[34] Many distributions satisfy this property including uniform distribution, gamma distribution with $\alpha > 1$, truncated normal distribution, etc.

From the policy making perspective, this theorem suggests that policy makers cannot only rely on past observations of how high-frequency trading changes the market to predict the welfare and market quality effects of high-frequency trading regulation. The reason is that regulations' would have huge effects on the HFT entry probability. Without monotonicity, the welfare and market quality effects might "flip signs". A theoretical framework is necessary to achieve a critical stance over high-frequency trading policy making.

2.4.3 Simultaneous Pricing Game

In this section I analyze the situation where the HFT only observes q_m (but not x_m) before setting her spread x_h. This corresponds to the market maker and the HFT having similar trading technologies and the HFT cannot undercut the market maker easily. This is related to two real world scenarios. First, some HFTs may become designated market makers.[35] With a better trading technology, the market maker can flicker quotes fast enough to avoid the HFT's detection. Second, the HFT might be constrained by exchange policies or regulation requirements such that she can no longer observe the price information ahead of other traders or undercut other traders easily.

We first analyze a one-shot simultaneous pricing game (q_m, q_h, π). In this game, the market maker's shareholding is q_m and the HFT enters the market holding q_h shares with probability π. Similar to the sequential pricing game, the buyer would

[35]Actually, two out of four NYSE's major designated market makers, Citadel Securities LLC and Virtu Americas LLC, are considered also as high-frequency trading firms.

purchase shares from the HFT first if the HFT and the market maker post the same spread.[36]

Definition 4. *An equilibrium of a one-shot simultaneous pricing game (q_m, q_h, π) is a pair of cumulative distribution function (H_m, H_h) such that x_m has CDF H_m and x_h has CDF H_m. Let the support of x_m (x_h) be a measurable set X_m (X_h). The equilibrium satisfies following conditions:*

1. *Given that the HFT posts spreads according to H_h, the market maker posting spreads according to H_m maximizes his expected payoff.*

2. *Given that the market maker posts spreads according to H_m, the HFT posting spreads according to H_h maximizes her expected payoff.*

3. *Given H_h, any $x_m \in X_m$ yields the same expected payoff for the market maker; this expected payoff is weakly higher than the expected payoff by posting a spread $x_m \notin X_m$.*

4. *Given H_m, any $x_h \in X_h$ yields the same expected payoff for the market maker; this expected payoff is weakly higher than the expected payoff by posting a spread $x_h \notin X_h$.*

The following proposition characterizes candidates of equilibrium.

Proposition 5. *No pure strategy equilibrium exists. Let the infimum of $X_m(X_h)$ be $\underline{x}_m(\underline{x}_h)$ and the supremum of $X_m(X_h)$ be $\bar{x}_m(\bar{x}_h)$. In any mixed strategy equilibrium,*

[36]The only purpose of this assumption is to make the simultaneous pricing case comparable to the sequential pricing case. The specific tie-breaking rule does not matter.

$\underline{x}_m = \underline{x}_h = \underline{x}$, $\bar{x}_m = \bar{x}_h = x^*$. *$X_m$ and X_h are dense in $[\underline{x}, x^*]$. There exists no $x_m(x_h) \in [\underline{x}, x^*)$ such that $x_m(x_h)$ is posted with positive probability in the equilibrium.*

Proof. See appendix. □

By Proposition 5, without loss of generality, I consider equilibrium where X_m and X_h are intervals. The equilibrium can be pinned down by the market maker and the HFT's indifference conditions.

Proposition 6. *There exists a unique equilibrium in the one-shot game (q_m, q_h, π) satisfying the following conditions:*

1. *If $k(q_m) \geq \pi k(q_h)$, in the equilibrium the market maker posts spread $x_m = x^*$ with positive probability $\bar{P}_m = 1 - \frac{\pi k(q_h)}{k(q_m)}$.*

   $\underline{x}$ is uniquely determined by

$$(1-\pi)k(q_m) + \pi(k(q_m + q_h) - k(q_h)) = a(\underline{x})k(q_m) \ . \tag{2.3}$$

   The market maker's mixed strategy satisfies

$$H_m(x) = (1 - \frac{a(\underline{x})}{a(x)}) \cdot \frac{k(q_h)}{k(q_m) + k(q_h) - k(q_m + q_h)} \quad \forall x \in [\underline{x}, x^*) \ . \tag{2.4}$$

   H_m satisfies $H_m(\underline{x}) = 0$, $\lim_{x \to x^{*-}} H_m(x) = 1 - \bar{P}_m$.

   The HFT's mixed strategy satisfies

$$H_h(x) = \frac{1}{\pi}(1 - \frac{a(\underline{x})}{a(x)}) \cdot \frac{k(q_m)}{k(q_m) + k(q_h) - k(q_m + q_h)} \quad \forall x \in [\underline{x}, x^*) \ . \tag{2.5}$$

H_h *satisfies* $H_h(\underline{x}) = 0$, $\lim_{x \to x^{*-}} H_h(x) = 1$.

2. *If* $k(q_m) \leq \pi k(q_h)$, *in the equilibrium the HFT posts spread* $x_h = x^*$ *with positive probability* $\bar{P}_h = 1 - \frac{k(q_m)}{\pi k(q_h)}$.

   $\underline{x}$ *is uniquely determined by*

$$k(q_m + q_h) - k(q_m) = a(\underline{x})k(q_h) \ . \tag{2.6}$$

   H_m *satisfies Equation (2.4). Moreover,* $H_m(\underline{x}) = 0$, $\lim_{x \to x^{*-}} H_m(x) = 1$.

   H_h *satisfies Equation (2.5). Moreover,* $H_h(\underline{x}) = 0$, $\lim_{x \to x^{*-}} H_h(x) = 1 - \bar{P}_h$.

Proof. By Proposition 5, X_m and X_h are dense in $[\underline{x}, x^*]$. Thus, in any "regular" equilibrium, $(\underline{x}, x^*) \in X_m$; $(\underline{x}, x^*) \in X_h$. Then the uniqueness naturally follows from the equilibrium construction.

I only prove the first part of the theorem here since the calculation for the second part is similar. The only difference is that x^* is not in the support of X_m since the payoff of posting x^* is strictly lower than posting $x^* - \epsilon$ for a small ϵ.

The HFT's indifference condition implies

$$(1 - \bar{P}_m)(k(q_m + q_h) - k(q_m)) + \bar{P}_m k(q_h) = a(\underline{x})k(q_h) \ . \tag{2.7}$$

The market maker's indifference condition implies

$$(1-\pi)k(q_m) + \pi(k(q_m+q_h) - k(q_h)) = a(\underline{x})k(q_m) \ . \tag{2.8}$$

By equation (2.7) and (2.8),

$$\bar{P}_m = \frac{a(\underline{x})k(q_h) + k(q_m) - k(q_m+q_h)}{k(q_h)+k(q_m)-k(q_m+q_h)} = 1 - \frac{\pi k(q_h)}{k(q_m)} \ . \tag{2.9}$$

H_m can be pinned down by the HFT's indifference condition:

$$a(x)[H_m(x)(k(q_m+q_h)-k(q_m)) + (1-H_m(x))k(q_h)] = a(\underline{x})k(q_h) \quad \forall x \in [\underline{x}, x^*) \ . \tag{2.10}$$

H_h can be pinned down by the market maker's indifference condition for all $x \in [\underline{x}, x^*)$:

$$a(x)\{(1-\pi)k(q_m) + \pi[H_h(x)(k(q_m+q_h)-k(q_h)) + (1-H_h(x))k(q_m)]\} = a(\underline{x})k(q_m) \ . \tag{2.11}$$

Notice that $a(x)$ is increasing with x for $x \in [0, x^*]$ and $k(q_m+q_h) < k(q_h)+k(q_m)$. Thus, existence and uniqueness of H_m and H_h is guaranteed by the intermediate value theorem. For the market maker (HFT), the indifference condition guarantees any strategy in support X_m (X_h) yields the same expect profit. From the proof of Proposition 5, no player has incentive to deviate to a spread smaller than $\underline{x}$ or larger than x^*. □

An important corollary of Proposition 6 is that the market maker's expected

payoffs are the same in both the sequential pricing game and the simultaneous pricing game. Since the market maker acts as if a short term payoff maximizer in the steady state, the same one-shot payoff induces the same capital commitment decision. This observation simplifies the comparison of market quality under two settings.

Corollary 3. *In any one-shot game (q_m, q_h, π), the market maker's expected profits are the same under sequential pricing and simultaneous pricing.*

Proof. If $k(q_m) > \pi k(q_h)$, the market maker would use the wide spread strategy in the sequential pricing game with expected profit

$$(1 - F(x^*))x^*[(1-\pi)k(q_m) + \pi(k(q_m + q_h) - k(q_h))] \ .$$

This equals the expected profit in the simultaneous pricing game when $k(q_m) > \pi k(q_h)$.

If $k(q_m) < \pi k(q_h)$, in the sequential pricing game, the market maker would use the tight spread strategy to achieve the expected payoff $(1 - F(\underline{x}))\underline{x}k(q_m)$ where the tight spread $\underline{x}$ is determined by

$$k(q_m + q_h) - k(q_m) = a(\underline{x})k(q_h) \ .$$

This equals the expected profit in the simultaneous pricing game when $k(q_m) < \pi k(q_h)$. □

Steady State Characterization

The following theorem relates equilibria in one-shot games to the steady state equilibrium of the infinite period game. Moreover, this theorem offers comparison over the market maker and the HFT's expected payoffs in the sequential pricing game and the simultaneous pricing game.

Theorem 5. *Let* $q_m = argmax_{q\in[0,\bar{q}]} \frac{\delta}{1-\delta} M(q) + (w_0 - q)$.

1. *Let* $x_m(q_m)$ *and* $x_h(q_m)$ *follow the mixed strategy defined in Proposition 6. Then* q_m, $x_m(q_m)$ *and* $x_h(q_m)$ *determines a steady state equilibrium.*[37] *In this equilibrium, the market maker's expected payoff is*

$$V_m(w_0) = \frac{\delta}{1-\delta} M(q_m) + (w_0 - q_m) \ .$$

2. *The market maker's expected payoffs and steady state capital commitments are the same in both sequential pricing and simultaneous pricing games.*

3. *The HFT is strictly better off in the sequential pricing game if* π *is in the wide spread region. The HFT's expected payoffs are the same under both settings if* π *is in the tight spread region.*

[37] When G has non-decreasing hazard rate, this equilibrium can be micro-founded by considering a model where the HFT does not observe δ and the market makers signals δ with capital commitment. There exists a perfect Bayesian equilibrium that shares the same on path property as this steady state equilibrium.

4. In a simultaneous pricing game, the steady state liquidity is

$$\begin{aligned} L =&(1-F(x^*))[\pi k(q_m+q_h)+(1-\pi)k(q_m)]+\pi\int_{\underline{x}}^{x^*}[H_m(z)H_h(z)k(q_m+q_h) \\ &+(1-H_m(z))H_h(z)k(q_h)+H_m(z)(1-H_h(z))k(q_m)f(z)dz \\ &+(1-\pi)\int_{\underline{x}}^{x^*}H_m(z)k(q_m)f(z)dz]\ . \end{aligned}$$

Proof. See appendix. □

It is informative to compare market qualities under the sequential pricing game and the simultaneous pricing game. By Theorem 5, the market maker's equilibrium capital commitments are the same in both settings. Pricing decisions of the market maker and the HFT drive the difference in market qualities. The following proposition summarizes liquidity comparison results.

Proposition 7. *Denote the steady state liquidity in the sequential pricing game and the simultaneous pricing game to be L_{se} and L_{sim}.*

1. *$L_{sim} > L_{se}$ if π is in the wide spread region.*

2. *$L_{sim} - L_{se}$ is constant for any π in the tight spread region.*

3. *L_{sim} and L_{se} is increasing in π in the tight spread region.*

Proof. See Appendix. □

To understand the intuition, first consider the wide spread region. In the sequential pricing game, all shares are supplied at the monopolistic spread x^* while in

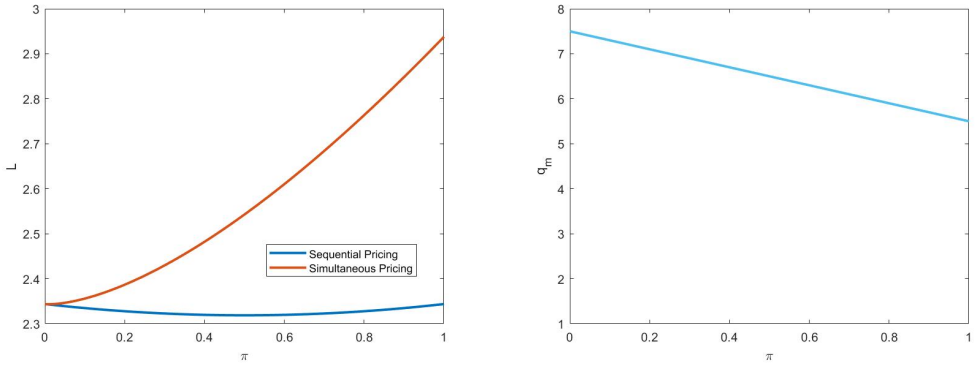
the simultaneous pricing game, spreads are lower than x^* with positive probability. Since the market maker makes the same capital commitment decisions, liquidity is higher in the simultaneous pricing game. Moreover, since the HFT cannot undercut the market maker at the spread x^* in the simultaneous pricing game, the HFT's expected payoff is lower. In other words, the HFT in the simultaneous pricing game is willing to pay a small cost to trade faster than the market maker. Conversely, since the market maker's expected payoffs are the same in both settings, in the sequential pricing game, the market maker has no incentive to upgrade his technology to trade at the same speed as the HFT. This means the HFT has stronger incentive to upgrade trading technology than the market maker. Yet as discussed above, this incentive is detrimental to market quality.

In the tight spread region, liquidity comparison between two settings is ambiguous. In the sequential pricing game, more shares are supplied at a low price because the market maker fixes a tight spread. Yet for a buyer with buying threshold between $1+\underline{x}$ and $1+x^*$, only the market maker's supply is available. On the other hand, in a simultaneous pricing game, a buyer with buying threshold $1+\underline{x}$ will not purchase any share with probability one. Yet for a buyer with buying threshold slightly lower than $1 + x^*$, in expectation he would be able to purchase more shares in a simultaneous pricing game. This ambiguity does not impose much difficulties in the quantitative analysis. I show that the liquidity difference between the sequential pricing game and the simultaneous pricing game is not changing in π in the tight spread region. With specific assumptions on distributions of the buyer's buying threshold and demand, I can achieve a clear comparison over liquidity under two settings in the tight spread

region.

2.4.4 Numerical Examples

This section contains numerical examples to visualize results in sections 2.4.2 and 2.4.3. In all examples, the buyer's buying threshold v follows a uniform distribution. The difference lies in the distribution of the buyer's demand q_b and the magnitude of HFT's shareholding q_h.

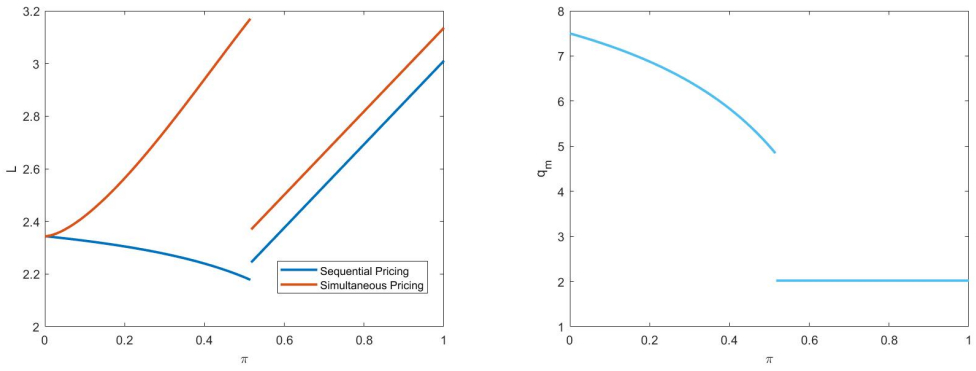


(a) Liquidity **(b)** Capital Commitment

Figure 2.2: Uniform Demand with Small HFT

Figure 2.2 depicts liquidity and the market maker's equilibrium capital commitment under different HFT entry probabilities when the buyer's demand q_b follows a uniform distribution and the HFT's shareholding q_h is small. With small q_h, even when $\pi = 1$, the market maker still sets the monopolistic spread in the equilibrium; i.e., the wide spread region is $[0, 1]$. As shown in Figure 2.2b, with no regime change, the market maker's equilibrium capital commitment is decreasing continuously with π.

The blue line in Figure 2.2a shows how steady state liquidity changes with π in the sequential pricing game. There exists a region where liquidity is decreasing in π. In this example, the region is $\pi \in [0, \frac{1}{2}]$. The red line in Figure 2.2a shows how liquidity changes with π in the simultaneous pricing game. As predicted by Proposition 7, liquidity in the simultaneous pricing game is higher.



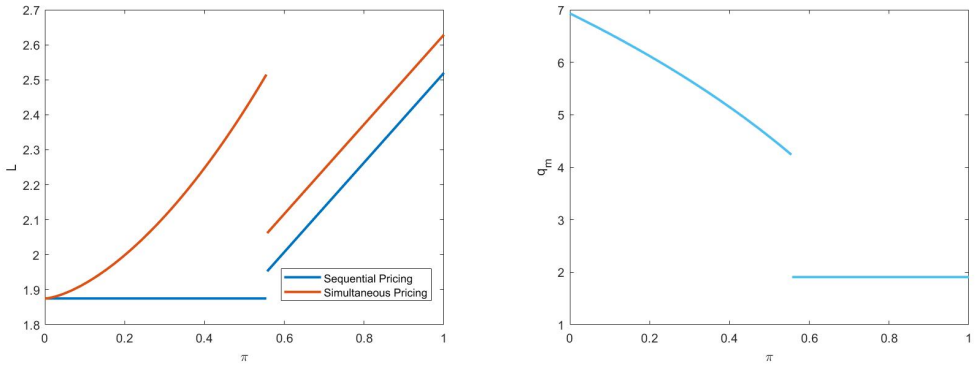
(a) Liquidity **(b)** Capital Commitment

Figure 2.3: Uniform Demand with Large HFT

Figure 2.3 shows liquidity and the market maker's capital commitment when q_b follows a uniform distribution and q_h is large. When π is large, the market maker would use the tight spread strategy in the equilibrium. This leads to the liquidity jump in Figure 2.3a and the capital commitment jump in Figure 2.3b. Since the market maker secures his payoff against the HFT entry in the tight spread region, the equilibrium capital commitment is not changing in π.

Another important observation can be made by comparing liquidity with $\pi \in [0.5, 0.6]$ and liquidity with $\pi = 0$ under the sequential pricing setting. Obviously, the average spread is lower in the tight spread region than in the monopolistic market.

However, the liquidity when $\pi \in [0.5, 0.6]$ is lower. The reason is that the market maker cut capital commitment facing the HFT's competition. This implies that pricing information alone cannot fully reflect market quality.



(a) Liquidity **(b)** Capital Commitment

Figure 2.4: Exponential Demand

Figure 2.4 shows liquidity and the market maker's capital commitment when the buyer's demand follows an exponential distribution. This serves as a robustness check by demonstrating a similar comparative statics. The only difference is that liquidity remains constant in the wide spread region in the sequential pricing game. This follows from the constant hazard rate property of the exponential distribution.

2.5 Costly High-Frequency Trading Participation

In this section, I consider an extension where the HFT needs to pay a fixed cost C to participate in high-frequency trading. Specifically, after observing the market maker's capital commitment q_m (and spread x_m in the sequential pricing game), the

HFT chooses whether to participate in high-frequency trading with cost C. If the HFT participates, she successfully enters the market with probability π. The cost C is paid regardless of the HFT successfully entering the market or not.[38] The HFT's profit is normalized to zero if she does not participate. This extension partially endogenizes the HFT's entry decision.[39]

2.5.1 Sequential Pricing Game

In the sequential pricing game, the HFT observes the market maker's shareholding q_m and spread x_m before making the entry decision. Consider a one-shot game with high frequency trading cost C. Since the HFT observes the market maker's shareholding and spread before posting her spread, I focus on the pure strategy equilibrium.

Definition 5. *An equilibrium of a one-shot sequential pricing game* (q_m, q_h, π, C) *is a triple* (x_m, η, x_h)*;* $\eta \in \{0, 1\}$ *indicates the HFT's participation decision. The HFT's participation (non-participation) of high-frequency trading is denoted by* $\eta = 1$ $(\eta = 0)$*.*

1. *Given the market maker's spread* x_m *and holding* q_m*,* x_h *maximizes the HFT's expected payoff.* $\eta = 1$ *if and only if the HFT's expected payoff is greater than* C*.*

[38] Another way to model costly participation is to assume that the HFT only pays the cost C upon successfully entering the market. Yet assuming the HFT always pays the cost is in line with the regulatory measures taken in practice. For instance, the German High Frequency Trading Act of 2013 requires exchanges to charge excessive system usage fees, including both order amendments and order cancellations. France and EU also have similar requirements on charging order cancellation fee. For examples of exchange policies complying these regulations, see Eurex. (2016) and Eurex. (2019).

[39] The endogenous entry is a special case of this setting with $\pi = 1$.

2. *Given the HFT posts spreads according to $x_h(x_m)$ and makes entry decisions according to η, x_m maximizes the market maker's expected payoff.*

It is useful to compare one-shot games with and without participation cost. Condition on participating in trading, the HFT's optimal pricing strategies and thus the market maker's pricing strategies in two games are the same. On the other hand, with participation cost, the HFT takes her entry probability π into account. Specifically, the HFT would lose money if she participates in high-frequency trading but cannot enter the market. This gives the market maker an additional strategic advantage.

Let the deterring spread $\underline{x}^d \leq x^*$ satisfy

$$\pi(1 - F(\underline{x}^d))\underline{x}^d k(q_h) = C \ .$$

If the market maker sets the deterring spread $\underline{x}^d$, participating in trading and undercutting the market maker is (weakly) not the optimal strategy for the HFT since doing so cannot cover the participation cost C. From the market maker's perspective, he would set $x_m = \underline{x}^d$ only when $\underline{x}^d \geq \underline{x}$. Facing the tight spread $\underline{x}$, the HFT is indifferent between setting the wide spread x^* and undercutting the market maker. Thus, when the market maker optimally sets $x_m = \underline{x}^d > \underline{x}$, it must be that participating in trading and setting $x_h = x^*$ cannot cover the participation cost, either. Thus, when the participation cost is high, the market maker may deter the HFT from participating by setting the deterring spread. Given the equilibrium strategy in a one-shot game without participation cost, the only additional decision for the market

maker in the similar game with $C > 0$ is whether to post the deterring spread $\underline{x}^d$. This strategy becomes more profitable with higher participation cost C. Formally, the market maker and the HFT's pricing decisions in a one-shot game (q_m, q_h, π, C) can be characterized as follows:

Proposition 8. *Consider a one-shot game (q_m, q_h, π, C). Let*

$$\bar{C}(\pi) = \pi(1 - F(x^*))x^* k(q_h) \ .$$

If $C \geq \bar{C}$ the market maker posts $x_m = x^*$ and the HFT does not participate in high-frequency trading ($\eta = 0$). For $C < \bar{C}$:

1. *If **(i)** $k(q_m) < \pi k(q_h)$ and*

$$C > \pi(1 - F(x^*))x^*[k(q_m + q_h) - k(q_m)] \ ,$$

   *or **(ii)** $k(q_m) > \pi k(q_h)$ and*

$$C > \frac{\pi k(q_h)}{k(q_m)}(1 - F(x^*))x^*[\pi(k(q_m + q_h) - k(q_h)) + (1 - \pi)k(q_m)] \ ,$$

   the market maker posts the deterring spread $\underline{x}^d$ and the HFT does not participate in high-frequency trading ($\eta = 0$).

2. *If $k(q_m) < \pi k(q_h)$ and*

$$C \leq \pi(1 - F(x^*))x^*[k(q_m + q_h) - k(q_m)] \ ,$$

the market maker posts the tight spread $\underline{x}$ and the HFT participates ($\eta = 1$). Upon a successful entry, the HFT sets $x_h = x^*$.

3. *If $k(q_m) > \pi k(q_h)$ and*

$$C \leq \frac{\pi k(q_h)}{k(q_m)}(1 - F(x^*))x^*[\pi(k(q_m + q_h) - k(q_h)) + (1 - \pi)k(q_m)] \ ,$$

the market maker posts the wide spread and the HFT participates ($\eta = 1$). Upon a successful entry, the HFT posts $x_h = x^*$ to undercut the market maker.

Proof. See appendix. □

Now consider the steady state in the infinite period game. A similar analysis guarantees the existence of a steady state equilibrium. The following result considers the comparative statics on C.

Theorem 6. *There exists $\hat{C}(\pi, q_h) \in (0, \bar{C})$ such that:*

1. *For $0 < C \leq \hat{C}$, the steady state equilibrium is the same as the steady state equilibrium with no participation cost ($C = 0$).*

2. *For $\hat{C} < C \leq \bar{C}$, the market maker sets the deterring spread $x_m = \underline{x}^d$ and the equilibrium capital commitment satisfying*

$$\frac{\delta}{1 - \delta}(1 - F(x_m))x_m(1 - G(q_m)) = 1 \ .$$

The HFT does not participate in high-frequency trading.

3. *For $C > \bar{C}$, the steady state equilibrium is the same as the monopolistic steady state equilibrium. The HFT does not participate in high-frequency trading.*

Proof. See appendix. □

This result is intuitive. When the participation cost is low, it is unprofitable for the market maker to deter the HFT from participating in trading.[40] In this case, the HFT's expected payoff is larger than the participation cost C. Thus, the HFT always participates and the steady state equilibrium is the same as the equilibrium with no participation cost. If the participation cost is high enough, the market maker deters the HFT's participation with the deterring spread. Moreover, the market maker optimally commits capital to the level such that the marginal value of capital commitment equals 1, the marginal value of dividend payout. The HFT in this situation does not participate in high-frequency trading. Finally, with an extremely high participation cost $C > \bar{C}$, the HFT never breaks even participating in high-frequency trading regardless of the market maker's spread. The market maker becomes a monopolist.

2.5.2 Simultaneous Pricing Game

In the simultaneous pricing game, the HFT only observes q_m, the market maker's shareholding, before making the participation decision. Consider a one-shot game (q_m, q_h, π, C). Similar to the simultaneous pricing game with no participation cost,

[40]The market maker may still prevent the HFT from undercutting with a tight spread strategy as in the baseline model

no pure strategy equilibrium exists. A mixed strategy equilibrium can be defined as follows.

Definition 6. *An equilibrium of a one-shot simultaneous pricing game* (q_m, q_h, π, C) *is a triple* (H_m, η, H_h). $\eta \in [0, 1]$ *is the HFT's participation probability.* x_m *follows CDF* H_m *and* x_h *follows CDF* H_h. *Let the support of* $x_m(x_h)$ *be* $X_m(X_h)$. *The equilibrium satisfies the following conditions:*

1. *Given that the HFT posts spreads according to CDF* H_h *and tries to enter according to* η, *the market maker posting spreads according to CDF* H_m *maximizes his expected payoff.*

2. *Given that the market maker posts spreads according to CDF* H_m, *the HFT posting spreads according to CDF* H_h *and tries to enter according to* η *maximizes her expected payoff.*

3. *Given* H_h *and* η, *any* $x_m \in X_m$ *yields the same expected payoff for the market maker; this expected payoff is weakly higher than the expected payoff by posting a spread* $x_m \notin X_m$.

4. *Given* H_m, *any* $x_h \in X_h$ *yields the same expected payoff for the market maker; this expected payoff is weakly higher than the expected payoff by posting a spread* $x_h \notin X_h$.

To find out the equilibrium pricing strategy of the one-shot game (q_m, q_h, π, C), consider $(q_m, q_h, \pi, 0)$, a one-shot game with no participation cost. If the HFT's expect profit in the equilibrium of game $(q_m, q_h, \pi, 0)$ is greater than C, in the game

(q_m, q_h, π, C), the HFT participates with probability 1 and both players use the same pricing strategy as in game $(q_m, q_h, \pi, 0)$. Conversely, if the HFT's expected equilibrium profit in game $(q_m, q_h, \pi, 0)$ is lower than C, she would mix in participation decision. This mixing has two effects. First, it reduces the expected participation cost. Second, by entering the market with a lower probability, the HFT improves her strategic position against the market maker in the pricing game. The participating probability η can be uniquely determined by the HFT's indifference condition over participation.

Proposition 9. *Consider a one-shot simultaneous pricing game (q_m, q_h, π, C). Define $a(\underline{x})(\pi)$ as in Proposition 6. That is, if $k(q_m) \geq \pi k(q_h)$,*

$$a(\underline{x})(\pi) = 1 - \pi + \pi \frac{k(q_m + q_h) - k(q_h)}{k(q_m)} \ ;$$

if $k(q_m) < \pi k(q_h)$,

$$a(\underline{x})(\pi) = \frac{k(q_m + q_h) - k(q_m)}{k(q_h)} \ .$$

1. *If*

$$\pi(1 - F(x^*))x^* a(\underline{x})(\pi)k(q_h) \geq C \ ,$$

the HFT chooses $\eta = 1$. The equilibrium of game (q_m, q_h, π, C) coincides with the equilibrium of game $(q_m, q_h, \pi, 0)$ characterized in Proposition 6.

2. *If*

$$\pi(1 - F(x^*))x^* a(\underline{x})(\pi)k(q_h) < C \ ,$$

there exists a unique $\eta \in (0,1)$ *such that*

$$\pi(1 - F(x^*))x^* a(\underline{x})(\eta\pi)k(q_h) = C \ .$$

In the equilibrium, the HFT participates with probability η *and receives zero expected payoff if enters. The equilibrium of game* (q_m, q_h, π, C) *coincides with the equilibrium of game* $(q_m, q_h, \eta\pi, 0)$.

Proof. See appendix. □

An important implication of this proposition is as follows:

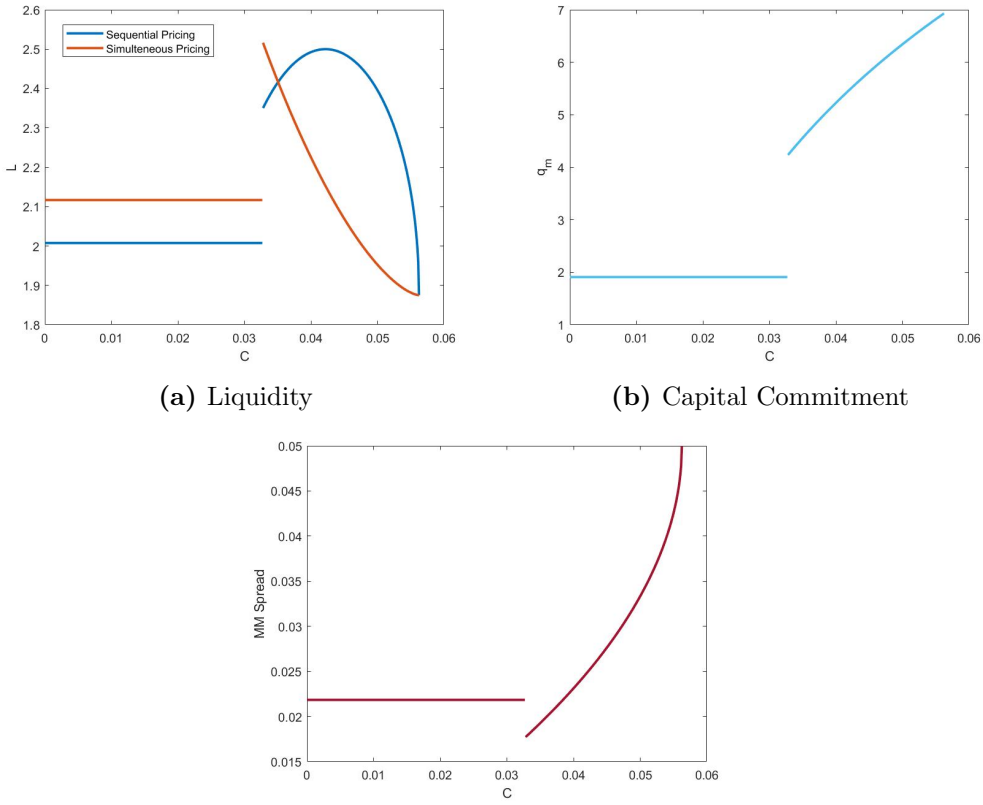
Corollary 4. *For any game* (q_m, q_h, π, C)*, the market maker's equilibrium payoffs are the same under both the sequential pricing and the simultaneous pricing settings.*

Proof. See appendix. □

Since the market maker receives the same expected payoffs in game (q_m, q_h, π, C) in the sequential pricing game and the simultaneous pricing game, the market maker's steady state capital commitments is both games are the same.

Proposition 10. *In the steady state, the market maker commits the same amount of capital in both the sequential and the simultaneous pricing game.*

Proof. This proof is similar to the no participation cost case and is thus omitted. □



(a) Liquidity **(b)** Capital Commitment

(c) MM's Spread (Sequential Game)

Figure 2.5: Comparative Statics on Participation Cost

2.5.3 Numerical Examples

This section presents a numerical example to illustrate how market quality changes with the HFT's participation cost. In this example, the HFT's entry probability π is fixed. The buyer's buying threshold v follows a uniform distribution while his demand q_b follows an exponential distribution. The market maker uses the tight spread strategy in the steady state when $C = 0$.

The equilibrium can be divided into three regions. With a low participation

cost, it is profitable for the HFT to participate with probability one. Thus, the market is the same to a market with no participation cost. As the participation cost increases, the market maker's deterring strategy becomes more profitable. Moreover, the marginal value of capital commitment also increases. Thus, the market maker's capital commitment is increasing with participation cost. One observation is that in the sequential game, the market maker's spread jumps downward when transiting into the deterring region. The reason is that when using the tight spread strategy, the spread is decreasing with the market maker's capital commitment. On the other hand, when the market maker is deterring the HFT with the deterring spread, this effect does not exist. Thus, when the market maker is indifferent between using the tight spread strategy and the deterring spread strategy, the deterring spread must be smaller. Finally, with a high participation cost, the market becomes a monopolistic market since it is never profitable for the HFT to participate.

2.6 Policy Implications

In this section, I collect results developed in previous sections to discuss effects of regulations over high-frequency trading. Taking the sequential game as the benchmark, this paper examines three types of regulations: altering the HFT's entry probability π, leveling the trading technology difference between the HFT and the market maker and imposing high-frequency trading participation cost.

2.6.1 Altering the HFT's Entry Probability

In practice, the HFT's entry probability hinges on the HFT's ability to detect other investor's orders and acquire shares in a timely manner. Regulations changing the HFT's detecting and purchasing capacities affect the HFT's entry probability. For instance, banning co-location and integrating financial markets would decrease the HFT's entry probability. Upgrading exchange's trading system without further restricting high-frequency trading would increase the HFT's entry probability.

This model predicts that in a market with high HFT entry probability (tight spread region), encouraging the HFT's entry is beneficial to market quality. The reason is that the market maker is setting a tight spread facing a fierce competition and the HFT is fulfilling the residual demand. An increase in HFT's entry probability leads to more liquidity supply from the HFT without changing the market maker's incentive to commit capital. On the other hand, in a market with low HFT entry probability, the market maker responses to the competition by committing less capital in market making. Liquidity would increase with the HFT's entry probability only if the benefit from more HFT supply outweighs the market maker's cut in capital commitment. Moreover, this model predicts that banning high-frequency trading does not necessarily deteriorates liquidity. Yet the spread would become higher due to the lack of competition.

2.6.2 Leveling the Trading Technology

This type of regulation "levels the playground" by making the market maker's trading technology comparable to the HFT's. For instance, the regulator can encourage HFTs to become designated market makers or incentivize existing market makers to upgrade their trading technologies. The batch auction proposed by Budish et al. (2015) also achieves this goal since the market maker would have a chance to revise his order.

This model predicts that this policy is beneficial when the HFT's entry probability is low (wide spread region). Without a superior technology, the HFT mixes in posting spreads rather than undercuts the market maker at the monopolistic spread. This drives down the average price and improves market quality. When the HFT's entry probability is high, this model predicts that leveling the trading technology leads to less shares for low-valuation buyers (because the market maker now mixes rather than stick to the tight spread) and more shares for high-valuation buyers (because the HFT now mixes rather than stick to the monopolistic spread). The overall effect can be ambiguous.

2.6.3 Imposing High-frequency Trading Participation Cost

The third type of regulation imposes a participation cost over high-frequency trading. For example, regulations in France and Germany require a fee to be charged based on both executed and canceled orders. Regulation in Germany further requires all traders to tag algorithm generated orders. These regulations essentially induce

participation costs on high-frequency trading.

This model predicts that low participation cost would not change the market quality. On the other hand, if the cost is high, the HFT would (at least partially) exit the market. The market maker's spread increases with the cost but he also commits more capital in market making. The directional change of liquidity depends on which effect dominates. Yet it is certain that extremely high participation cost always hurts the market.

2.7 Flipping

In this section, I consider the situation where the HFT can flip orders by first purchasing shares from the market maker and then reselling them at a higher spread. The HFT observes the market maker's capital commitment q_m and spread x_m before making flipping and pricing decisions. There are two implicit assumptions. First, the HFT is not capital constrained.[41] Second, the market maker does not have enough time to acquire additional shares from the inter-dealer market after the HFT purchases shares from him. For the ease of notation, I assume that G has an unbounded support. When G has a bounded support, the qualitative results are essentially the same. All proofs in this section are delegated to the appendix.

First consider the HFT's flipping and pricing decisions in a one-shot game (q_m, q_h, π). If the HFT flips shares, her spread must be higher than the market maker's spread. This implies her optimal spread is $x_h = x^*$. If the market maker holds q_m shares and

[41]The HFT's shareholding q_h reflects the exogenous market condition. Thus, the HFT cannot expand her shareholding even she is not capital constrained.

his spread is $x_m < x^*$, the HFT's expected payoff when buying q_f shares from the market maker is

$$r(q_f) = (1 - F(x^*))x^*[k(q_m + q_h) - k(q_m - q_f)] - x_m q_f \ . \tag{2.12}$$

The first term of the right hand side is the expected gain from selling $q_h + q_f$ shares at spread x^* when the market maker is left with $q_m - q_f$ shares at a lower spread x_m. The second term of the right hand side is the premium paid by the HFT. The HFT pays $1 + x_m$ for each flipped share. If the buyer does not purchase these shares, the HFT only receives 1 by selling each share left to the inter-dealer market. By purchasing shares from the market maker, the HFT reduces the market maker's supply and thus the competition. Since the market maker's spread is lower and the demand is uncertain, the marginal benefit of flipping is increasing in q_f. Thus, the HFT would follow an "all or nothing" flipping strategy.

Proposition 11. *The HFT either purchases the market maker's entire shareholding* q_m *or nothing. In other words,* $q_f = q_m$ *or* 0.

Then consider the market maker's pricing problem. At a low enough price, by Proposition 11, the HFT would purchase all shares from the market maker upon entry. Thus, comparing to the baseline case, the market maker has an additional option to strategically lower his spread to induce flipping. The highest possible spread that induces flipping, x_m^f, can be pinned down by the HFT's indifference condition: Buying all shares from the market maker should be more profitable than the optimal pricing strategy without flipping. This can be summarized by the following lemma:

Lemma 3. *x_m^f satisfies*

$$(1 - F(x^*))x^* k(q_m) \geq x_m^f q_m \tag{2.13}$$

and

$$(1 - F(x^*))x^* k(q_m + q_h) \geq x_m^f q_m + (1 - F(x_m^f))x_m^f k(q_h) \ . \tag{2.14}$$

At least one inequality is binding. Moreover, if Inequality (2.14) binds, the flipping strategy dominates the tight spread strategy.

The wide and tight spread strategies are still available to the market maker. Specifically, if the market maker uses a wide spread, his expected payoff is

$$(1 - F(x^*))x^*[\pi(k(q_m + q_h) - k(q_h)) + (1 - \pi)k(q_m)] \ .$$

If $\underline{x} > x_m^f$, the market maker's expected payoff from the tight spread strategy is

$$(1 - F(\underline{x}))\underline{x} k(q_m) \ .$$

If the market maker posts x_m^f, his expected payoff is

$$\pi x_m^f q_m + (1 - \pi)(1 - F(x_m^f))x_m^f k(q_m) \ .$$

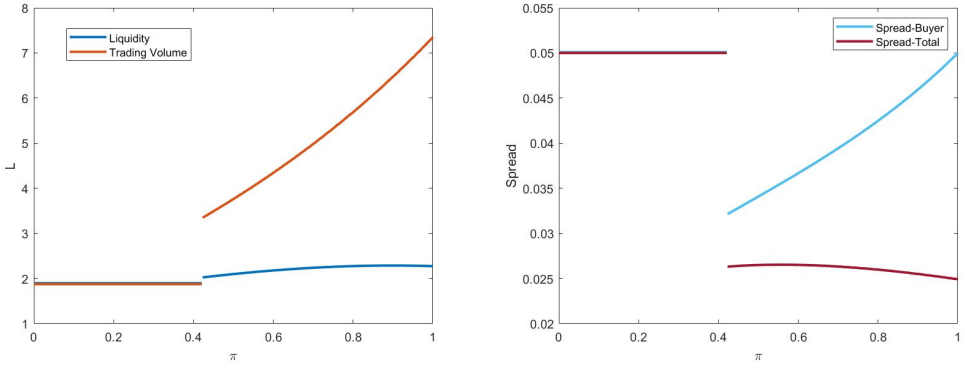
An important observation is that, if the market maker expects the HFT to flip shares, the market maker's expected payoff is increasing in π. With flipping, the HFT is providing insurance for the market makers. When the HFT entry probability is large,

the market maker would always induce flipping.

Proposition 12. *Under any q_m and q_h, if π is high enough, the market maker sets spread x_m^f in the equilibrium.*

In the infinite period game, although the market maker can be insured by the HFT, he does not have the inventive to increase capital commitment indefinitely. This is because under any capital commitment level, the expected payoff is upper-bounded by the monopolistic payoff. Moreover, when the capital commitment is large, the spread to induce flipping becomes close to zero. This implies an upper-bound exists for the market maker's capital commitment in the steady state equilibrium.[42]

Proposition 13. *For large enough w_0, a steady state equilibrium exists.*



(a) Liquidity v.s. Volume **(b)** Buyers' v.s. Average Spread

Figure 2.6: Equilibrium Volume and Price with Flipping

[42]The upper-bound may not be $\bar{q}$ since the market maker's expected payoff with shareholding $\bar{q}$ might be lower than the monopolistic payoff.

Figure 2.6 presents a numerical example showing the equilibrium liquidity and average spread when the HFT is able to flip orders. At low HFT entry probability, whether the HFT can flip shares or not does not change the equilibrium outcome. The reason is that from the market maker's perspective, the benefit of using a low spread to induce flipping is not large enough.

When the HFT's entry probability becomes large, the market maker sets a low spread to induce flipping. A large portion of transactions happens between the market maker and the HFT since the HFT purchases all low price shares upon entering. The buyer only benefits from the market maker's low spread when the HFT fails to enter the market. This suggests that it is important to separate trades between liquidity suppliers (the market maker and the HFT) and trades from liquidity suppliers to the buyer. As shown in Figure 2.6a, the expected trading volume and the average spread do not accurately reflect the market quality. The expected shares sold to the buyer only increase modestly in π comparing to the expected trading volume. Moreover, the average spread remains low while the buyer is facing a much higher spread increasing in π. This is because the majority of low price shares are purchased by the HFT. As the HFT becomes more likely to enter the market, the buyer becomes less likely to purchase cheap shares. When the HFT can flip shares, the entry of HFT only has limited benefits for the buyer. If we only look at the overall trading data, the benefit of high-frequency trading will be overestimated.

2.8 Supply Schedule and Limit Order Book

Up till now I assume the market maker sells all shares at one spread. In this section, I analyze an extension where the market maker can submit a supply schedule to sell shares at different spreads. To keep the problem tractable, I maintain the assumption that the HFT sells all her shares at one spread and determines her spread after observing the market maker's capital commitment q_m and supply schedule.

Formally, given the market maker's capital commitment q_m, his pricing strategy can be represented by a supply schedule Ψ. The amount of shares supplied by the market maker with spreads less or equal to x is $q_m\Psi(x)$. In the steady state, the market maker posts the supply schedule to maximize his expected profit in each period. Thus, it is suffice to first solve for the optimal Ψ in a one-shot game under any q_m and then determine the steady state capital commitment with the market maker's indifference condition.

2.8.1 No HFT

Consider a one-shot game where the market maker holding q_m shares maximizes expected profit in a single period. With no HFT, even though the market maker may sell shares at different spreads, he optimally supplies all shares at the monopolistic spread x^*. Formally, we have the following proposition:

Proposition 14. *Given any q_m, in a one-shot game, the market maker would optimally set the supply schedule to be $\Psi(x) = I_{\{x \geq x^*\}}$.*

Proof. See appendix □

A direct implication of Proposition 14 is that, in a infinite period game with no HFT, the steady state equilibrium is the same as the equilibrium in the baseline model. In other words, with no potential competition from the HFT, the market maker has no incentive to submit a non-degenerate supply schedule.

Corollary 5. *When no HFT exists, the steady state equilibrium is the same as the baseline model. Moreover, the market maker does not pay dividend when his net worth is smaller than the steady state capital commitment* $\bar{w}$.

Proof. The first statement is a straightforward result from Proposition 14. For the second statement, if the dividend payout is non-zero, the market maker can always achieve a higher payoff by refraining from paying dividend and supply the extra amount of shares at the spread x^* and payout the total return from the extra shares in the next period. □

2.8.2 With HFT

When the HFT may enter the market, the market maker's pricing strategy is non-degenerate. Specifically, it is never optimal for the market maker to sell all shares at one spread. The intuition behind this result is simple. Given any single spread pricing strategy, the market maker can always sell a small amount of shares at another spread without changing the HFT's pricing strategy. If the market maker is using the wide spread strategy, he can improve his payoff by selling some shares before the HFT at a spread close to the monopolistic spread. If the market maker is using the tight spread strategy, he can sell some shares at a higher spread without the HFT

undercutting him.

Proposition 15. *For any q_h and $\pi > 0$, supplying all shares at any spread x is not the optimal pricing strategy for the market maker in the steady state equilibrium.*

Proof. See appendix □

Moreover, with the ability to flexibly sell shares, an immediate lower bound $\underline{q}$ exists for the market maker's capital commitment in the steady state. If the capital commitment level is below $\underline{q}$, the market maker can always improve his expected payoff by committing more capital and sell additional shares at the spread x^*.

Corollary 6. *The market maker would commit at least $\underline{q} > 0$ unit of capital, as long as his capital commitment with no HFT is non-zero. Specifically, $\underline{q}$ is the solution of*

$$\frac{\delta}{1-\delta}(1-F(x^*))x^*[\pi(1-G(\underline{q}+q_h))+(1-\pi)(1-G(\underline{q}))] = 1 \ .$$

Notice that $\underline{q}$ is also the market maker's equilibrium capital commitment level in the wide spread region of the baseline model. Thus, allowing the market maker to submit a supply schedule improves liquidity in the wide spread region. Liquidity change in the tight spread region when the market maker can submit a supply schedule is ambiguous. However, the following proposition guarantees that given any specific set of parameters, the market maker's supply schedule can be easily computed. Then the change in spreads and liquidity can be characterized through numerical calculation.

Proposition 16. *The market maker's equilibrium pricing strategy* $\Psi(x)$ *satisfies three conditions:*

1. $\Psi(x^*) = 1$.
2. $\Psi(\cdot)$ *has no mass point for* $x < x^*$.
3. *The HFT achieves the same expected payoff by setting any* $x_h \in [\underline{x}, x^*]$ *where* $\Psi(\underline{x}) = 0$.[43]

Proof. See appendix □

With this result, the market maker's equilibrium capital commitment q_m and pricing strategy $\Psi(x)$ can be numerically computed with the following algorithm: (i) Fix a q_{ml}, the amount of shares sold by the market maker with spreads lower than x^*. (ii) If $q_{ml} \leq \underline{q}$, $q = \underline{q}$; i.e., the market maker sells $\underline{q} - q_{ml}$ shares at the monopolistic spread x^*. Otherwise, $q = q_{ml}$. (iii) Given q_{ml} and q, $\Psi(x)$ is pinned down by

$$(1 - F(x))x[k(\Psi(x)q + q_h) - k(\Psi(x)q)] = (1 - F(x^*))x^*[k(q_{ml} + q_h) - k(q_{ml})]$$

for $x \in [\underline{x}, x^*)$ and $\Psi(x^*) = 1$. (iv) As in the baseline case, let $M(q)$ be the expected per-period payoff of the market maker with capital commitment q. If $q = \underline{q}$, define $M(\underline{q})$ to be the maximum expected payoff for $q_{ml} \in [0, \underline{q}]$. (v) The market maker's equilibrium capital commitment is

$$q_m = max_{q \in [\underline{q}, \bar{q}]} \frac{\delta}{1 - \delta} M(q) + (w_0 - q) \ .$$

[43]Notice that the second result is implied by the third result. If there is a mass point at a spread $x < x^*$, the indifference condition cannot hold everywhere.

The market maker's pricing strategy is then pinned down by the procedure above.

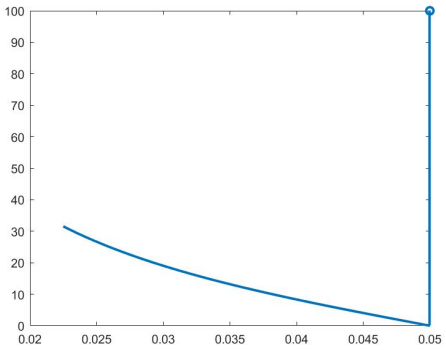


Figure 2.7: Supply Schedule of the Market Maker

When the buyer's demand q_b follows an exponential distribution, the market maker's supply schedule can be explicitly characterized. Specifically, let $\psi(x) = \Psi'(x)$. Then for $x \in [\underline{x}, x^*)$, $\psi(x) \propto \frac{1}{x} - \frac{f(x)}{1-F(x)}$. Figure 2.7 provides a visual illustration of the market maker's supply schedule under a further assumption that the buyer's spread tolerance $v - 1$ follows a uniform distribution. The x-axis represents the spread while the y-axis represents the density of the market maker's supply schedule. The density of the market maker's supply is decreasing to zero approaching the monopolistic spread x^*. Moreover, the line at rightmost of Figure 2.7 demonstrates that the market maker is supplying a positive quantity at the spread x^*.

2.8.3 Discussion

This extension analyzes the change in market quality when the market maker can sell shares at different spreads. With no HFT, the limit order book is degenerate in the sense that all shares are still supplied at the monopolistic spread x^* as in

the baseline model. Conversely, when the HFT might enter the market, the market maker would supply shares at continuum of spreads. This improves liquidity in the wide spread region. The liquidity change in the tight spread region is ambiguous.

Moreover, this extension illustrates how asymmetric competition between the market maker and the HFT determines the shape of the limit order book.[44] Intuitively, fixing the HFT's pricing strategy, the market maker has incentive to increase spreads of some shares for higher expected profit. Yet to prevent the HFT from undercutting, the market maker needs to supply enough amount of shares at low spreads. This trade-off uniquely determines the shape of the limit order book. In any steady state equilibrium, the market maker would choose a supply schedule such that the HFT is indifferent between undercutting the market maker at any spread in the schedule and posting the monopolistic spread x^*.

2.9 Conclusion

My paper studies how high-frequency trading changes market quality through affecting the traditional market maker's capital commitment and pricing decisions. I consider a long-run market maker facing competition from short-run HFTs in providing liquidity. In the steady state, the long-run market maker responses to the competition by reducing his spread and committing less capital in market making. The latter effect impairs market quality. Thus, when taking the market maker's capital commitment channel into consideration, high-frequency trading does not necessar-

[44] Roşu (2009) analyzes a similar problem under the assumption that each market participant supplies one unit of share to the market and all market participants have the same trading speed.

ily improves market quality though it always (weakly) reduces the average spread. Moreover, in my model, the difference in trading technologies between the HFT and the market maker affects market quality. When the HFT's entry probability is low, "leveling the playground" by making the market maker and the HFT trade at the same speed improves market quality.

I further consider three extensions. The first extension introduces high-frequency trading participation cost to endogenize the HFT's participation. When the HFT trades faster than the market maker and the participation cost is low, market quality remains the same. On the other hand, when the participation cost is high, the market maker optimally sets a spread to deter the HFT from entering the market. Although the HFT does not participate in trading when the participation cost passes a certain threshold, the cost level still affects the market quality since the market maker's deterring strategy depends on the cost. When the HFT and the market maker trade at the same speed, the model's prediction is similar except that the HFT mixes in participation facing high participation cost.

In the second extension, the HFT can "flip shares" by purchasing shares from the market maker and resupplying them at a higher spread. With high HFT entry probability, the market maker would induce flipping by posting a low spread since flipping effectively insures the market maker. Yet the buyer does not benefit much from the low spread since most of the cheaper shares are mainly acquired by the HFT. This extension demonstrates the importance to exclude the trading between liquidity suppliers when evaluating market quality. Otherwise, market quality would be overestimated with an overestimation of the expected trading volume and an

underestimation of the average spread.

The third extension investigates implications on the shape of the limit order book when the market maker can sell shares at different spreads. Specifically, with no HFT, the market maker would still sell all shares at the monopolistic spread. Facing the competition from the HFT, the market maker would sell shares at a continuum of spreads. This extension demonstrates how competition between the market maker and the HFT determines the shape of the limit order book.

I want to emphasize several important insights of this model. First, the price information alone cannot fully reflect market quality; the volume information is equally important. Second, more high-frequency trading does not necessarily improves market quality since it reduces the market maker's willingness to commit capital in market making. Third, the relative trading speed between the market maker and the HFT affects market quality. When the HFT's entry probability is low, letting the market maker and the HFT trade at same speed improves market quality. Fourth, it is important to separate the trades among liquidity suppliers to avoid overestimations on market quality and the high-frequency trading's welfare effect.

Chapter 3

Do Electronic Markets Improve Execution If You Cannot Identify Yourself?[1]

3.1 Introduction

Over the last thirty years, the most significant change to the operation of financial markets has been the shift from human-based trading to electronic trading. While it is clear that small traders have benefited from electronic trading based on a narrower bid-ask spread (Hendershott et al., 2011b), the impact on large traders is less certain, with some research suggesting that more frequent trading is associated with additional costs for these traders. This sentiment was shared by Warren Buffet, who stated in a CNBC interview that high-frequency traders have made it more costly for large traders but have made small investors better off (Crippen, 2014). This is somewhat counterintuitive, since large traders are more likely to have the sophistication and resources to take advantage of faster, more complex electronic markets.

[1]This chapter is based off of a paper of the same name co-authored with John Shim (Shim and Wang (2021)).

One underemphasized byproduct of replacing humans with computers has been a move towards anonymity, which is of particular relevance to large traders who, in the context of floor trading, used their reputations to reduce execution costs (Battalio et al., 2007). One way in which traders used reputations was through preannouncement, defined by Admati and Pfleiderer (1991) as announcing and committing to trading in the future. While preannouncement has previously been identified as a way of reducing execution costs by revealing a trader is uninformed, in this paper we focus on the commitment power aspect of preannouncement, which is distinct from the informational channel. In essence, commitment power was what allowed preannouncements to carry weight – a trader's reputation credibly signaled that the trader would act as promised. This suggests that replicating commitment power in anonymous electronic markets, where preannouncements cannot be verified, may be difficult. However, electronic markets have also significantly increased trading frequency, which may allow large traders to reduce overall price impact by spreading out trading over time, and offset the loss of commitment power utilized in non-electronic markets.

In this spirit, we develop a model to study the optimal execution problem of large traders with and without commitment power, and with different trading frequencies. Specifically for commitment power, we have in mind the ability to commit to an entire sequence of future orders, which we refer to as scheduling, that lends itself more naturally to a setting with many trading periods.[2] We highlight two main find-

[2]We think of scheduling as the analogue to preannouncement when there are multiple trading periods. If there is only one trading period, preannouncement and scheduling are conceptually identical. When there are multiple periods, scheduling assumes that the trader announces her entire schedule of trades before any trades take place.

ings. First, the benefits of frequent trading without scheduling relative to infrequent trading with scheduling are relatively small and can even be negative, depending on the level of competition amongst market makers. This suggests that traders who previously utilized reputation-based execution strategies on the floor may not benefit much, if at all, from faster electronic markets if they gave up scheduling as part of the transition. Second, increasing trading frequency improves execution costs for both the trader that does and does not utilize scheduling, but the benefits are significantly larger in magnitude for scheduling traders. Thus, while the implementation of scheduling becomes more complex in moving from the floor to anonymous electronic markets – reputations no longer being sufficient – the benefits also increase considerably.

In the context of frequent trading, our model suggests that scheduling with frequent trading may have important benefits, and these benefits may be large enough to overcome at least some challenges and costs associated with implementing scheduling in modern, anonymous markets. We raise the possibility that scheduling in electronic markets can be achieved by using transparent, automated execution algorithms. While execution algorithms do not directly reveal trader identities, they can be made open-source and imprint a recognizable trading signature. Moreover, since algorithms behave in precise, pre-programmed ways and run autonomously without human intervention, market makers can have confidence that future trades will be executed predictably.[3] We acknowledge that there may be important costs associated with revealing intentions to trade. In particular, "predatory" trading strategies

[3]It is also important that either algorithms run so quickly as to make it unlikely or impossible for humans to intervene, or that humans are disincentivized from intervention.

have been a primary concern of many traders in electronic markets.[4] To address this concern, we extend the model to incorporate a stylized predatory trader and confirm that the benefits of scheduling are large enough to overcome this particularly salient source of cost.

The model we present in this paper has a single risky asset with uncertain payoff following a normal distribution, and two types of agents: a risk-neutral trader with an exogenous need to sell exactly z units of a risky asset with a sequence market orders, and several large risk-averse market makers.[5] The trader chooses how much to sell in each period to minimize total execution costs. The market makers are identical, initially maintain no position in the asset, and each provides liquidity by submitting a demand schedule every period while taking into account his own impact on prices, i.e., there is imperfect competition amongst market makers. We use the number of trading periods, N, as a measure of trading frequency. Since we study the execution problem of a large trader over a relatively short period of time, such as an hour or a day, faster electronic markets mean there can be more frequent interactions within that time. Thus, we think of small N as corresponding to slower human-based markets and large N as corresponding to electronic markets, with trade occurring over the same clock time. Scheduling corresponds to the trader's ability to credibly commit to a specific sequence of future market orders. The asset's value is realized after all trading periods, and there is no asymmetric information – the trader's total desired selling quantity, the distribution of the asset's value, and the market makers'

[4]Some of this concern may be attributable to both the air of secrecy surrounding high-frequency traders, who operate with different objectives than investors, and media reports.

[5]The analysis is symmetric for a trader that needs to buy.

initial inventory are all common knowledge.

We show that, for a given trading frequency, scheduling always results in smaller execution costs for the large trader compared to when scheduling is not utilized. Moreover, in nearly all cases, scheduling leads to smaller execution costs than a market with perfectly competitive market makers.[6] The intuition behind these results stems from two observations. First, market depth (or the inverse of price impact) is decreasing as trading approaches the final trading period as in Rostek and Weretka (2015). Second, the equilibrium price in any given trading period depends not only on the quantity traded in that period but also the quantity that will be traded in all future periods.[7] For example, selling in the second period pushes prices down in both the second and first periods. These two observations are key to understanding the trader's optimal execution strategies with and without scheduling.

To see the intuition behind the optimal strategies, consider a two period version of the model. Given the features highlighted above, it might seem intuitive to sell more in periods where market depth is greater. Indeed, the optimal selling sequence without scheduling is to spread out trading across periods by selling more in the first period, which has greater market depth. However, scheduling leads to a different pattern: overselling. That is, the trader oversells in the first period and buys back the excess in the second period. She does this because buying in the second period pushes up prices in both periods – committing to buying in the second period leads

[6]The only case in our model where scheduling does not outperform the case with perfectly competitive market makers is when trading frequency and the number of market makers are both at their minimum value. For more details on this case, please see Section 3.4.3.

[7]This echos the insight in Kyle and Obizhaeva (2016) that we should focus on the total execution quantity rather than analyzing individual trades separately.

to greater selling prices in the first period. In addition, since the market is shallower in the second period, she has greater impact on prices with fewer units traded.[8]

We also show that while the scheduling and non-scheduling trader both benefit from an increase in trading frequency, the benefit is significantly larger for the scheduling trader. This is because the scheduling trader is better able to take advantage of time-varying price impact by using later more price-sensitive periods to buy and raise prices in earlier periods. Thus, the pattern of overselling and buying back is exacerbated and, in addition, leads to an exponential increase in trading volume.[9] And although non-scheduling traders also benefit from an increase in trading frequency, they still prefer a market with perfectly competitive market makers, even as the number of trading periods approaches infinity. On the other hand, as trading frequency increases, the scheduling trader's cost approaches the first best, which can be achieved in a dealer market where the trader can make take-it-or-leave-it offers to each market maker. In fact, scheduling with very frequent trading is in some ways superior to trading in this particular dealer market as scheduling does not require knowledge of the value of market makers' risk aversion, whereas making take-it-or-leave-it offers in a dealer market does.

The model can help understand how trading costs have changed during the transition from human to electronic trading. First, the model shows that traders that never

[8]Vayanos (2001) also presents a case where there is overselling, but it stems from asymmetric information about who is trading.

[9]Broadly speaking, as the frequency of trade in equity markets has increased over time, volume has also increased. While our model does not incorporate many other motivations for trade (i.e., speculation, informed trading, small noise traders, etc.), the positive relationship between trading frequency and volume over time are broadly consistent with the model.

utilize scheduling still benefit from a move to more frequent, electronic trading. This suggests that traders who never utilized reputation-based commitment strategies on the floor are better off from an increase in trading frequency. Second, traders that utilized scheduling on the floor but moved to non-scheduling strategies when trading became more frequent, perhaps due to the difficulty in translating reputation-based scheduling strategies to anonymized electronic markets, are likely to have seen a small decrease in trading costs, or even an increase in trading costs. Lastly, traders that have been able to implement scheduling with frequent electronic trading are likely to have seen the biggest reduction in execution costs. This is consistent with the notion that, on average, execution costs for large traders seem to have gone down as trading has shifted moved from humans to computers. Moreover, the model's results are also consistent with significant heterogeneity in the change in large trader execution costs over the transition, with some traders reporting significant improvement and others reporting only minor reductions or even greater costs.[10]

As mentioned above, a natural concern with scheduling is that announcing intentions to trade in the future may lead to "predatory" trading. Specifically, there may be concerns that speculators will profit from anticipated future trading and price pressure, wiping out gains from scheduling. We study this case in an extension of our model with a single predatory trader who starts and ends with zero inventory of the risky asset and only seeks to make trading profits. We find that the predator is

[10]In a Liquidnet survey, more than two-thirds of asset managers responded that they believe their trading costs were rising as a result of high-frequency trading, though the specific reason was not provided (Grant and Demos, 2011). Comerton-Forde et al. (2011) provide empirical evidence consistent with the claim that many uninformed traders choose to trade non-anonymously if they are able to do so.

able to earn positive profits at the expense of the trader, but not enough to outweigh the benefits of scheduling. That is, the predictability of scheduling allows the predator to make profits but despite this, the trader is still better off with scheduling than without. While modern traders seemingly fear leakage about trading intentions, our results suggest that the costs associated with revealing intentions may not necessarily outweigh the benefits if implemented correctly.[11]

While scheduling reduces execution costs for traders, the optimal scheduling sequence of orders may, on the surface, appear to resemble manipulation. Indeed, the investor does cause prices to move based on trading units with no intention of holding those units (the amount oversold), and takes advantage of time-variation in price impact. This type of trading activity has been described as "market depth arbitrage" (Black, 1995). However, this type of activity in general does not lead to a reduction in price or allocative efficiency (Kyle and Viswanathan, 2008). In fact, in our setting the overall price impact is reduced with scheduling in that prices are closer the the expected asset value. In addition, the excessive units traded come at a loss, serving only to improve prices for inframarginal units.

The rest of the paper is structured as follows: in Section 3.2 we review related literature; in Section 3.3 we introduce the model with two trading periods; in Section 3.4 we present the full model with an arbitrary number of trading periods; in Section 3.5 we provide an extension of the model with a predatory trader; in Section 3.6 we

[11]One caveat is that our paper studies a setting with no asymmetric information about fundamentals: large uninformed traders may benefit from committing to a sequence of orders. We also emphasize that our results hold for the optimal sequence of trading and not any predictable sequence of trading (algorithms targeting execution at the volume-weighted average price, or VWAP, would be an example of the latter).

conclude.

3.2 Related Literature

From the theoretical perspective, this paper can be viewed as a multi-period extension of Kyle (1989) with a strategic trader submitting market orders and large risk-averse market makers absorbing orders with demand schedules. Rostek and Weretka (2015) point out that when large risk-averse traders share risk via demand schedules, aggregate supply is constant, and there is no additional information, market depth becomes shallower as periods approach the trading deadline. This paper further demonstrates how a risk-neutral trader can optimize liquidation over multiple periods with time-varying market depth by adjusting the aggregate asset supplied in each period. The benefit of commitment power is also clear based on this perspective. When the trader can commit to the aggregate supply for all periods, she can pick the optimal schedule to maximize total liquidation value. When the trader does not have commitment power, she is limited to supply sequences that give her no incentive to deviate and improve total liquidation value. On the other hand, the fact that the trader does not have an incentive to deviate implies that greater trading frequency, i.e., more trading periods in the model, always improves execution. Du and Zhu (2017) also consider the implication of trading frequency, but in a dynamic double auction framework. They focus on analyzing trades motivated by private information and the relationship between trading frequency and allocative efficiency, whereas we focus on optimal execution in the absence of asymmetric information

and how it is affected by trading frequency.

Our findings also contribute to the literature on strategic execution. Vayanos (2001) considers a large trader's liquidation problem when facing endowment shocks. In Vayanos (2001), market makers are perfectly competitive and patterns in price and trading are driven by the inability of market makers to distinguish orders from the large trader and small traders. Bertsimas and Lo (1998), Obizhaeva and Wang (2013), and Carlin et al. (2007), among other papers, also study optimal trader execution. In these models, price processes are exogenously specified while in our model price is endogenously determined by the equilibrium. Three related recent papers are Pritsker (2009), Fardeau (2020) and Glebkin et al. (2020). Pritsker (2009) analyzes risk sharing of risk-averse traders competing with market orders to explain time-varying liquidity's effect on asset prices. Fardeau (2020) considers risk sharing between imperfectly competitive traders as price takers under anticipated demand shocks to explain the pattern of price movements following the shock. Glebkin et al. (2020) consider a single period risk sharing problem among large traders with multiple assets and non-Gaussian asset payoffs.

This paper also relates to research on the role of anonymity in trading. Simaan et al. (2003) argue that anonymous quotes can improve price competition. Foucault et al. (2007), Rindi (2008), Comerton-Forde et al. (2011) and Meling (2020) consider the informational effect of trading without anonymity.[12] This paper complements the existing research by providing a *non-informational* channel through

[12]In this sense, the role of non-anonymity in these papers is similar to the role of pronouncement in Admati and Pfleiderer (1991). Other empirical research on the liquidity implications of anonymity includes Battalio et al. (2007), Hachmeister and Schiereck (2010), Friederich and Payne (2014), Dennis and Sandås (2020), Pham et al. (2016), etc.

which anonymity can affect trading. Specifically, trading anonymously could lead to inferior execution for large traders since they lose the commitment power associated with their reputation.

This paper also brings new insight into predatory trading. Existing literature mainly focuses on how high-frequency traders (HFT) or other speculators may take advantage of a trader by learning of future trading intentions. Brunnermeier and Pedersen (2005) and Carlin et al. (2007) show that speculators may exploit other traders' need to sell an asset by first selling the asset in anticipation and then buying back. Yang and Zhu (2020) show theoretically how speculators may engage in "back running," which entails inferring fundamental information from past order flow and trading on it. Hirschey (2020) presents evidence that some high-frequency traders may use anticipatory strategies that resemble of electronic front running. Our focus is a bit different in that we study the potential benefits of predictable trading, the kind of trading that is thought to be susceptible to predators. In the predatory trader extension of our model, we highlight the difference between implementing a predictable trading strategy that may entice predators and scenarios without predators. In the context of our model, we find that the benefits of using a predictable trading strategy, i.e., scheduling, outweigh the risk of attracting predators.

3.3 Two-Period Model

In this section, we present our model studying optimal execution, but with two trading periods. This allows us to emphasize the intuition behind scheduling, much

of which can be shown with the two-period model. In Section 3.4, we extend the model to N trading periods where we study the effect of more frequent trading on execution.

The model has a single risky asset and two types of agents: a risk-neutral trader and $I \geq 3$ risk-averse market makers. The risk neutral trader needs to sell $z \geq 0$ units of the risky asset before a fixed trading deadline.[13] The analysis is identical for a trader that needs to purchase as there is no asymmetric information about the quantity or side on which the trader needs to execute. Market makers have CARA preferences with the same risk aversion coefficient ρ. We assume that each market maker is large in the sense that each recognizes his own price impact. Market makers start with no position in the asset and they compete to purchase the asset with demand schedules. The asset's value is realized immediately after the trading deadline and follows a normal distribution with mean μ and variance σ^2.[14] Since execution takes place in a relatively short period of time, we assume that agents do not discount. Also, there is no asymmetric information – the trader's desired selling quantity z, the distribution of the asset value, the number of trading periods, and market makers' initial positions are common knowledge.

As mentioned above, we introduce the notion of scheduling, which allows the trader to announce and commit to an entire sequence of market orders before trading commences. If the trader does not engage in scheduling, she submits a market order

[13]Note that z represents the trader's selling quantity – a positive value of z indicates the trader needs to sell z units.

[14]We avoid the dynamic inconsistency issues raised by Basak and Chabakauri (2010) since uncertainty is resolved once after all trading periods.

each period and market makers submit demand schedules in each period without observing the trader's market order, i.e, the trader and market makers submit in each period simultaneously. All agents can observe all market orders and demand schedules in past periods. Denote the size of the trader's sell market order in period t by z_t and market maker i's demand schedule in period t by $x_{it}(p_t)$.[15] To be clear, z_t is the trader's selling quantity in period t: a positive value of z_t represents the number of units sold, a negative value of z_t represents the number of units purchased. As is standard in the literature, we focus on symmetric linear equilibria in which all market makers' demand schedules are the same affine function of the current period's asset price. We find equilibria in our model when the trader does and does not utilize scheduling as defined below, and the definition applies to both the two-period model presented in this section and the N-period model in the next section.

Definition 7. A linear symmetric equilibrium with imperfect competition is a quadruple $\{z_t, \hat{z}_t, x_t(p_t), \lambda_t\}$ satisfying:

1. *On the equilibrium path, each market maker correctly anticipates other market makers demand schedules, and his period t conjecture for the trader's market sell order, $\hat{z}_t$, matches the trader's order z_t; for market maker i, the demand schedule $x_{i,t}(p_t)$ maximizes*

$$\mu\left(\sum_{k=1}^{t} x_{ik} + \sum_{j=t+1}^{N} x_{ij}\left(p_j^*\right)\right) - \frac{\rho\sigma^2}{2}\left(\sum_{k=1}^{t} x_{ik} + \sum_{j=t+1}^{N} x_{ij}\left(p_j^*\right)\right)^2$$
$$-x_{it}p_t - \sum_{j=t+1}^{N} x_{ij}\left(p_j^*\right)p_j^*$$

[15] Here we highlight the demand schedule's dependence on p_t. The demand schedule could in principle depend on other variables.

for all t.

2. In period t, p_t^* is determined by the market clearing condition $\sum_{i=1}^{I} x_{it}\left(p_t^*\right) = z_t$; for any future period $j > t$, p_j^* is determined through the market clearing conjecture $\sum_{i=1}^{I} x_{ij}\left(p_j^*\right) = \hat{z}_j$.

3. $x_{it}\left(p_t\right)$ is an affine function of p_t.

4. When each market maker submits the specified demand schedule, the implied linear price impact is $\lambda_t > 0$.

5. *Intertemperal Market Clearing at earh period t:* $\sum_{j=1}^{t-1} z_j + \sum_{j=t}^{N} \hat{z}_j = z$

6. The trader maximizes:
   (a) With scheduling, the trader determines the entire sequence of market sell orders $z_1, \ldots, z_N$ at the beginning of period 1 to maximize the total selling value $\sum_{t=1}^{N} p_t^* z_t$.
   (b) Without scheduling, the trader's sequence of market sell orders maximizes her total selling value in any proper sub-game given market makers' strategies.

We note that we maximizing selling value is equivalent to minimizing execution costs (which is the difference between prices transacted and the expected asset price μ).[16]

In presenting the model, we first describe market makers' demand schedules anticipating the trader's sequence of market orders. We then analyze the trader's optimal market order submission problem, highlighting the differences between when

[16]We also use the terms liquidation value and execution value interchangeably with selling value.

the trader engages in scheduling and does not. The two period model provides the following insights: (1) the trader's execution costs are smaller when implementing scheduling; (2) with scheduling, the trader oversells in the first period and buys back in the second; (3) without scheduling, the trader splits her desired selling quantity over the two periods.

3.3.1 The Market Maker's Problem

We focus on symmetric equilibria, and conjecture that market maker i's demand schedule in each period t is

$$x_{it} = a_{it} - b_{it} p_t. \tag{3.1}$$

Like much of the literature, we assume that price impact is linear in quantity traded and takes the form

$$p_t = \tilde{p}_t + \lambda_{it} x_{it}$$

where $\tilde{p}_t$ is a constant term and λ_{it} captures market maker i's price impact.

We consider the demand schedule of market maker i and solve backwards by starting with period 2. The market maker chooses x_{i2} to maximize

$$(x_{i1} + x_{i2})\,\mu - \frac{\rho}{2}\,(x_{i1} + x_{i2})^2\,\sigma^2 - (\tilde{p}_2 + \lambda_{i2} x_{i2})\,x_{i2}.$$

From the first order condition, we have

$$x_{i2} = \frac{\mu - \rho\sigma^2 x_{i1} - p_2}{\lambda_{i2} + \rho\sigma^2}.$$

Since all market makers are identical, we have by symmetry that each market maker i has $b_{i2} = b_2$ and $\lambda_{i2} = \lambda_2$. Comparing parameters with equation 3.1 yields

$$b_2 = \frac{\gamma}{\rho\sigma^2}$$
$$\lambda_2 = \frac{1-\gamma}{\gamma}\rho\sigma^2,$$

and, following the notation of Rostek and Weretka (2015), we define $\gamma = 1 - \frac{1}{I-1} < 1$, which can be interpreted as the degree of competition amongst market makers – as I, the number of market makers, increases, γ approaches 1. Given market maker i's inventory for the first period x_{i1}, his demand schedule in the second period is

$$x_{i2}(p_2) = \frac{\gamma}{\rho\sigma^2}(\mu - p_2) - \gamma x_{i1}.$$

By market clearing, the equilibrium price p_2^* is

$$p_2^* = \mu - \frac{\rho\sigma^2}{\gamma}\frac{z_2}{I} - \rho\sigma^2\frac{z_1}{I}. \tag{3.2}$$

The equilibrium demand is

$$x_{i2}^* = x_{i2}(p_2^*) = \frac{z_2}{I} + \gamma\left(\frac{z_1}{I} - x_{i1}\right).$$

Now consider market maker i's problem in the first period. Similar to the analysis

above, market maker i chooses x_{i1} in the first period to maximize

$$(x_{i1} + x_{i2}^*)\,\mu - \frac{\rho}{2}\left(x_{i1} + x_{i2}^*\right)^2 \sigma^2 - p_2^* x_{i2}^* - (\tilde{p}_1 + \lambda_{i1} x_{i1})\, x_{i1},$$

which yields the equilibrium demand schedule

$$x_{i1}(p_1) = \frac{\gamma}{(1-\gamma)^2 \rho\sigma^2}(\mu - p_1) - \frac{(2-\gamma)\,\gamma}{(1-\gamma)}\frac{z_2}{I} - \frac{(2-\gamma)\,\gamma^2}{(1-\gamma)^2}\frac{z_1}{I}, \tag{3.3}$$

and, by market clearing, the equilibrium price is

$$p_1^* = \mu - (2-\gamma)\,\rho\sigma^2\frac{z_2}{I} - \left[(2-\gamma)\,\gamma + \frac{(1-\gamma)^2}{\gamma}\right]\rho\sigma^2\frac{z_1}{I} \tag{3.4}$$

and the first period demand is $x_{i1}^* = \frac{z_1}{I}$.

We wish to point out a few important observations from the equilibrium demand schedules and prices. First, given the first period market sell order z_1, the equilibrium price in the second period, p_2^*, does not depend on market makers' inventory from the first period, x_{i1}.[17] However, market makers' equilibrium demand schedules do. Second, examination of equation 3.3 reveals that market maker i's forecast of the second period market sell order, z_2, affects the level of his first period demand schedule and thus the first period equilibrium price. Since $\frac{1}{2} \leq \gamma < 1$, the market order in the second period has a greater impact on first period equilibrium price than the first period market order. This suggests that, given the total desired selling quantity z, the trader can increase the first period price by selling less of the asset

[17]This stems from the fact that the assumption that market makers are symmetric and we focus on the symmetric equilibrium.

in the second period and more of the asset in the first period.

3.3.2 The Trader's Problem without Scheduling

We first solve for the optimal sequence of market orders if the trader does not utilize scheduling. Without scheduling, the equilibrium condition imposes restrictions over the credible market order sequences. Specifically, in the equilibrium, under market makers' conjectures on the trader's market orders z_1 and z_2, the trader has no incentive to deviate. This requirement is trivially satisfied in the second period since the aggregate selling amount z is common knowledge and all agents observe past orders. However, in the first period, given the market makers' conjectures, the trader may deviate by allocating more/less demand to the second period. This no deviation constraint uniquely determines the market order sequences in the equilibrium and thus the trader's execution cost. From the trader's perspective, the equilibrium requirement limits the possible market order sequences for execution and, ultimately, reduces the value the trader can receive from selling.

If (z_1, z_2) is an equilibrium selling sequence, the trader has no incentive to deviate to $z_1 - k$ and $z_2 + k$ for any $k \neq 0$. If the trader deviates with size k, equilibrium prices would be

$$\hat{p}_2^* = \mu - \frac{\rho\sigma^2}{\gamma}\frac{z_2 + k}{I} - \rho\sigma^2\frac{z_1 - k}{I}$$
$$\hat{p}_1^* = \mu - (2-\gamma)\,\rho\sigma^2\frac{z_2}{I} - (2-\gamma)\,\gamma\rho\sigma^2\frac{z_1}{I} - \frac{(1-\gamma)^2\,\rho\sigma^2}{\gamma}\frac{z_1 - k}{I}.$$

These reflects that market makers, not knowing the trader's deviation, maintain the

equilibrium demand schedule from equation 3.3 in the first period. Yet the realized market order in the first period is $z_1 - k$, which changes the market clearing price in the first period (and accounts for the single $z_1 - k$ term in the expression for $\tilde{p}_1^*$). Observing the first period market order, market makers know the second period market order will be $z_2 + k$ since the aggregate liquidation size is z. Since both the market makers' inventories and the second period market order quantity change, the market clearing price in the second period also differs.

Given market makers' conjectures, the trader's optimal deviation k is determined by

$$k^* = \arg\max_k \ (z_2 + k)\,\tilde{p}_2^* + (z_1 - k)\,\tilde{p}_1^*.$$

In the equilibrium, since the trader does not deviate and market maker conjectures are correct, $k^* = 0$. Equivalently,

$$\left(\frac{2}{\gamma} - 3 + \gamma\right) z_2 = \left[\frac{2\,(1-\gamma)^2}{\gamma} + (2-\gamma)\,\gamma - 1\right] z_1.$$

And, since $z_1 + z_2 = z$, we have that $z_1 = \frac{1}{2-\gamma} z$ and $z_2 = \frac{1-\gamma}{2-\gamma} z$. This implies that in equilibrium, prices are

$$p_1^* = p_2^* = \mu - \frac{\rho\sigma^2}{\gamma\,(2-\gamma)}\frac{z}{I}$$

and the trader's value from selling z units without scheduling, V_{ns}, is

$$V_{ns} = \mu z - \frac{\rho\sigma^2}{\gamma\,(2-\gamma)}\frac{z^2}{I}.$$

This equilibrium features selling in each period ($z_1, z_2 > 0$) and constant prices.[18] Specifically, the trader splits up her total desired selling quantity across the two periods, with the amount split between the two periods increasing as the degree of competition between market makers γ decreases.

3.3.3 The Trader's Problem with Scheduling

We now analyze the case where the trader implements scheduling. That is, when the trader announces and commits to her entire sequence of market sell orders before the first trading period.

The trader engaged in scheduling determines the market orders in each period, z_1 and z_2, to maximize $p_1^* z_1 + p_2^* z_2$ such that $z_1 + z_2 = z$. Moreover, we assume that the trader announces her sequence of orders and all know with certainty that she will follow through with her announced sequence exactly. Since all quantities are determined ex-ante, the problem reduces to a simple maximization problem, and solving yields $z_1 = \frac{1}{2(1-\gamma)} z$ and $z_2 = \frac{1-2\gamma}{2(1-\gamma)} z$. Thus, given the expressions in equations 3.4 and 3.2, the equilibrium prices with scheduling are

$$p_1^* = \mu - \frac{\left[2 - \gamma - (1-\gamma)^2\right] \rho\sigma^2}{2\gamma} \frac{z}{I} \tag{3.5}$$

$$p_2^* = \mu - \frac{\rho\sigma^2}{2\gamma} \frac{z}{I}, \tag{3.6}$$

[18] Without scheduling, prices are constant in the two period model but this is not the case in general as the number of trading periods grows. See Section 3.4.2 for more details.

and the trader's liquidation value for selling z units while utilizing scheduling, V_s, is

$$V_s = \mu z - \frac{2+\gamma}{4\gamma}\rho\sigma^2\frac{z^2}{I}. \tag{3.7}$$

The proposition below summarizes the equilibria.

Proposition 17. *With two trading periods, market maker i's equilibrium demand schedules in the first and second periods, respectively, are*

$$x_{i1}(p_1) = \frac{\gamma}{(1-\gamma)^2\rho\sigma^2}(\mu - p_1) - \frac{(2-\gamma)\gamma}{1-\gamma}\frac{z_2}{I} - \frac{(2-\gamma)\gamma^2}{(1-\gamma)^2}\frac{z}{I}$$
$$x_{i2}(p_2) = \frac{\gamma}{\rho\sigma^2}(\mu - p_2) - \gamma x_{i1}.$$

The trader's selling sequence is as follows:

- *Without scheduling, she sells $z_1 = \frac{1}{2-\gamma}z$ and $z_2 = \frac{1-2\gamma}{2-\gamma}z$ in the first and second periods, and the equilibrium prices are*

$$p_1^* = p_2^* = \mu - \frac{\rho\sigma^2}{\gamma(2-\gamma)}\frac{z}{I},$$

  and the trader's liquidation value is

$$V_{ns} = \mu z - \frac{\rho\sigma^2}{\gamma(2-\gamma)}\frac{z^2}{I}.$$

- *With scheduling, she sells $z_1 = \frac{1}{2(1-\gamma)}z$ and $z_2 = \frac{1-2\gamma}{2(1-\gamma)}z$ in the first and second*

periods, and the equilibrium prices are

$$p_1^* = \mu - \frac{\left[2 - \gamma - (1-\gamma)^2\right] \rho\sigma^2}{2\gamma} \frac{z}{I}$$
$$p_2^* = \mu - \frac{\rho\sigma^2}{2\gamma} \frac{z}{I},$$

and the trader's liquidation value is

$$V_s = \mu z - \frac{(2+\gamma)\,\rho\sigma^2}{4\gamma} \frac{z^2}{I}.$$

3.3.4 Comparing Execution with and without Scheduling

We highlight several important observations regarding the two equilibria. Most notably, the optimal second period selling quantity with scheduling is negative since $\frac{1}{2} \leq \gamma < 1$. That is, the trader submits a buy market order in the second period despite needing to sell in total. In order to sell z units, the scheduling trader sells more than z units in the first period and buys the oversold amount back in the second period. By scheduling to buy units in the second period, the trader raises the the execution price in the first period. As mentioned above, examining the expressions for equilibrium prices in equations 3.2 and 3.4 reveals that the second period order has a greater impact on the first period price than the first period order.

However, there is a limit to this strategy. While the trader's buy order in the second period raises the price in the first period, it also raises the price in the second period. And since the second period order has a greater impact on the second

period price than the first period price, eventually the losses on the oversold amount outweigh the benefits of raising the selling price on the inframarginal units sold in period 1. To some, this strategy may seem like manipulation. Specifically, the trader engages in excessive trading to alter prices and may be interpreted as "market depth arbitrage" Black (1995), which Kyle and Viswanathan (2008) argue is innocuous in that it does not make prices less efficient and does not reduce allocative efficiency. In the context of our model, scheduling causes prices to be closer to the expected asset value μ in every period.

The non-scheduling trader implements an equilibrium strategy more aligned with conventional wisdom by splitting her total selling amount into smaller sell orders. The exact splitting method is uniquely pinned down by the no deviation condition. Specifically, the trader cannot profit from deviating from market makers' conjectured market orders in periods 1 and 2. To see this, suppose market makers mistakenly conjecture that the trader is utilizing scheduling and will trade according to the sequence of orders from the scheduling equilibrium. In this case, the first period price is expected to be high due to conjectured purchases in the second period, which drives up prices in both periods. This means that the trader would have a profitable deviation by buying less in the second period and, since the total selling amount is fixed, selling less in the first period while still taking advantage of greater first period market depth.[19] This is also the sense in which the set of feasible market order sequences without scheduling is limited.

[19]Most will think of market depth as the inverse of price impact or the price impact coefficient, i.e., Kyle's Lambda (Kyle, 1985a, 1989). In this particular example, market depth shifts because of the constant term in the demand schedule.

Since the scheduling trader is able to implement a more effective execution strategy, her execution value is greater than the trader that does not utilize scheduling, i.e., $V_s > V_{ns}$. It is also useful to characterize another benchmark: a perfectly competitive market. Specifically, we model the perfectly competitive market as one where market makers do not take their own price impact into account.[20] The value received by a trader in this setting is given by $V_c = \mu z - \rho\sigma^2 \frac{z^2}{I}$; we provide full details of this case in Appendix B.1.1. We can see that with two periods, the non-scheduling value is smaller than the value in a competitive market but the scheduling value is greater than both, i.e., $V_s > V_c > V_{ns}$.

3.4 General Model

In this section, we extend the two period model to incorporate arbitrarily many trading periods. Specifically, we normalize the start of the game to begin at clock time 0 and the trading deadline to be at clock time 1. Let η be the length of time needed to complete one period of trading – smaller η indicates that trading can occur more quickly. We think of smaller η more generally as representing the speed and efficiency improvements that come from the adoption of technology by exchanges and market participants. The trader in our model can trade $N = \left\lfloor \frac{1}{\eta} \right\rfloor$ periods before the trading deadline. Since our focus is on the implications of trading speed within a fixed period of time, we use the parameter N in the analysis, which is equivalent

[20] This corresponds to the case where there are infinitely many market makers and their price impact is negligible because of their infinitesimal size.

in our setting to trading frequency.[21] Thus, we interpret the analysis in this section as extending the analysis of Section 3.3 to incorporate more frequent trading but within the same interval of clock time.

Just as above, we solve for the representative market maker's optimal pricing schedule, then solve the model for the trader that utilizes scheduling and for one that does not. The N period model builds on the intuition of the two period model and provides the following new insights: (1) Qualitatively, both the scheduling and non-scheduling trader submits orders in a similar pattern as in the two period model; (2) The trader benefits from an increase in trading frequency regardless of whether she utilizes scheduling; (3) The trader almost always prefers to trade in a slow market with scheduling than a fast market without scheduling; (4) As trading becomes very frequent, the trader that utilizes scheduling achieves a selling value approaching the first best, which is achieved in a dealer market. Equilibrium in the N period model follows Definition 7 stated above.

3.4.1 The Market Makers' Problem

Consider each market maker's optimal demand schedule under any given sequence of market sell orders $z_1, \ldots, z_N$. In this case, we call a demand schedule an equilibrium demand schedule if Properties (1)-(4) in Definition 7 are satisfied fixing $z_1, \ldots, z_N$. As in the two-period model, we focus on symmetric equilibria and assume that price impact in period t is affine in the quantity traded in period t. The following

[21] N can also be interpreted as the inverse of the time to execute a trade, η. We sometimes refer to a market with small N as a slow market and large N as a fast market.

proposition characterizes the optimal demand schedule.

Proposition 18. *For a given sequence of market sell orders $z_1, \ldots, z_N$, there exists a linear symmetric equilibrium demand schedule where market maker i's demand schedule in period t is*

$$\begin{aligned} x_{i,t}(p_t) =& b_t(\mu - p_t) - \gamma \sum_{j=1}^{t-1} x_{i,j} - \gamma \left[\frac{1}{(1-\gamma)^{2(N-t)}} - 1 \right] \frac{z}{I} \\ & - \sum_{j=t+1}^{N} \left[\gamma + \frac{\gamma(1-\gamma)}{(1-\gamma)^{2(j-t)}} \right] \frac{z_j}{I} \end{aligned}$$

where $\gamma = 1 - \frac{1}{I-1}$ and

$$\begin{aligned} b_t &= \frac{\gamma}{(1-\gamma)^{2(N-t)} \rho \sigma^2} \\ \lambda_t &= \frac{1-\gamma}{b_t}, \end{aligned}$$

and the equilibrium asset price in period t is

$$p_t^* = \mu - \rho\sigma^2 \frac{z}{I} - \frac{\lambda_t z_t}{I} - \gamma \sum_{j=t+1}^{N} \frac{\lambda_j z_j}{I}.$$

The functional form of the demand schedule is worth discussion. Other than the current price p_t, market maker i's demand schedule also depends on his inventory of the risky asset accumulated over all previous periods, $\sum_{j=1}^{t-1} x_{ij}$, the trader's total desired selling quantity z, and the anticipated market sell orders in future periods z_j for $j > t$. The fact that the current demand schedule depends on future market orders suggests that scheduling will play an important role. Similar to the two period

example, without scheduling some sequences of market orders may not be credible since the trader has an incentive to deviate. This additional restriction reduces the value at which the trader is able to sell her units. The above proposition also shows that the equilibrium price in period t depends on both current and future selling by the trader. That is, future sales reduce future prices as well as the current period price. [22]

3.4.2 The Trader's Problem without Scheduling

We first analyze the model when the trader does not engage in scheduling, i.e., she cannot or does not commit to a sequence of market orders ex-ante. We show that the trader sells gradually over time, just as in the two period model. The following proposition characterizes the trader's optimal sequence of market sell order quantities and resulting equilibrium prices.

Proposition 19. *There exist two sequences* $\{\alpha_n\}_{n=1}^{\infty}$ *and* $\{A_n\}_{n=1}^{\infty}$*, where* $\alpha_1 = A_1 = 1$*. In an N-period execution problem, there exists a unique symmetric linear equilibrium in which a trader's sub-game perfect Nash equilibrium market sell order quantity sequence is* $z_t = \frac{\alpha_{N-t+1}}{A_N} z$ *for all* t*.* $\{\alpha_t\}_{t=2}^{\infty}$ *and* $\{A_t\}_{t=2}^{\infty}$ *are determined inductively*

[22] Note that our assumed specification for price impact restricts the effect of future trading activity on current prices to the intercept term. Also note that our price impact specification is identical to the one used by Rostek and Weretka (2015), though in their paper aggregate supply in every period is fixed at zero.

by equations

$$\begin{aligned} a_t =& \frac{2}{A_{t-1}(2-\gamma)} \sum_{k=1}^{t-1} \alpha_k \left[\sum_{j=1}^{k-1} \frac{\gamma(1-\gamma)}{(1-\gamma)^{2(t-j)}} \alpha_j + \frac{1-\gamma}{(1-\gamma)^{2(t-k)}} \alpha_k \right] \\ & - \sum_{k=1}^{t-1} \frac{\gamma(1-\gamma)}{(1-\gamma)^{2(t-k)}(2-\gamma)} \alpha_k \\ A_t =& \sum_{i=1}^{t} a_t. \end{aligned}$$

Given z and $z_1, \ldots, z_{t-1}$, each market maker conjectures $z_j = \frac{\alpha_{N-j+1}}{A_N - A_{t-1}} \left(z - \sum_{k=1}^{t-1} z_k \right)$ for $j = t, \ldots, N$. Market maker demand schedules and equilibrium asset prices are as given in Proposition 18.

One important observation is that the trader's sequence of market orders is sub-game invariant. Specifically, if the trader has the same number of trading periods and the same amount of units left to liquidate in two sub-games, the trader's equilibrium selling sequence in both sub-games is the same. An immediate implication of this is that, even without scheduling, the trader can achieve a greater total liquidation value with more frequent trading. To see this, imagine that a trader faces N periods and deviates once by selling nothing in the first period. In this case, because the optimal sequence is sub-game invariant, her liquidation value and sequence would be the same as the equilibrium liquidation value when she can trade for only $N-1$ periods. Thus, the fact that there is no profitable deviation means that she must receive a greater liquidation value in a game with N trading periods compared to one with $N-1$ periods.

Corollary 7. *The trader that does not utilize scheduling has liquidation value V_{ns} which is increasing in the number of trading periods N.*

Another important implication of Proposition 19 is that the non-scheduling trader does not flip the sign of her orders. The trader in this case takes advantage of more frequent trading by breaking her large sell order, z, into many smaller ones. This is consistent with the conventional wisdom that more frequent trading has led to large traders breaking up their orders into smaller pieces to reduce overall price impact.

Corollary 8. *In the non-scheduling equilibrium, $z_t \geq 0$ for all t, i.e., the trader never purchases shares, i.e., $\sum_t |z_t| = z$.*

Although the trader is able to increase her total selling value with more frequent trading, the value she is able to obtain is always bounded above by the liquidation value when she trades in a market with perfectly competitive market makers V_c. To see this, notice that in any period t, the equilibrium asset price without scheduling (p_t^* from Proposition 18) is less than the (constant) price from a perfectly competitive market, $\mu - \rho\sigma^2 \frac{z}{I}$, for all t if if $z_t > 0$. Closer examination also reveals that the bound holds regardless of the value of N.

Corollary 9. *The non-scheduling trader achieves a greater liquidation value in a market with perfect competition amongst market makers regardless of trading frequency. That is, $V_{ns} \leq \mu z - \rho\sigma^2 \frac{z^2}{I} = V_c$ for any N.*

3.4.3 The Trader's Problem with Scheduling

Just as in the two period model, when the trader utilizes scheduling she solves

$$\max_{z_1,\ldots,z_N} \sum_{t=1}^{N} p_t^* z_t.$$

The following proposition characterizes the trader's optimal sequence of market sell order quantities and the corresponding liquidation value.

Proposition 20. *When the trader utilizes scheduling, the equilibrium market sell order quantity submitted in period t is*

$$z_t = \frac{(1-t\gamma)\, z}{N\,(1-\gamma)^{N-t+1}},$$

the equilibrium asset price in period t is

$$p_t^* = \mu - \frac{\left(2-\gamma-(1-\gamma)^{N-t+1}\right)\rho\sigma^2}{N\gamma}\frac{z}{I},$$

and the trader's final liquidation value is

$$V_s = \mu z - \frac{[2+(N-1)\,\gamma]\,\rho\sigma^2}{2N\gamma}\frac{z^2}{I}$$

which is increasing in N.

Similar to the two period model, scheduling leads to excessive trading due to initially overselling and then buying back the excess in later periods. [23] That is,

[23] With only one large trader, the trader only sells in the first period. Generally, when there are

$\sum_t |z_t| > z$. The price of the asset is also increasing over trading periods. This pattern of overselling may seem surprising at first given prices are increasing and the trader is essentially selling at the lowest price of any of the N periods. But similar to the two period model, it is critical to observe that the first period's equilibrium price explicitly depends on market orders in all future periods. So while the first period price may seem low relative to the price in subsequent periods, it is high relative to what the price would have been if future market orders were sales as opposed to buys. Equivalently, market makers are willing to buy at a relatively high price in the first period only because the trader commits to repurchase some of the shares in future periods. The following corollary summarizes.

Corollary 10. *When there are many trading periods and the trader engages in scheduling, the optimal sequence of market orders involves overselling and buying back. Specifically, the trader sells in the first period ($z_1 > 0$) and buys in subsequent periods ($z_2 \leq 0$ and $z_t < 0$ for $t \geq 3$). The equilibrium price is increasing in t.*

We next discuss the trader's liquidation value with very frequent trading. Proposition 20 reveals that when $N \to \infty$, the trader's total liquidation value converges to $V_d = \mu z - \frac{\rho\sigma^2}{2}\frac{z^2}{I}$, which is the highest possible liquidation value. This value is achieved only in a dealer market where the trader can make take-it-or-leave-it offers in the form of a price-quantity pair to each of the I market makers. Since market makers all have the same level of risk-aversion, the trader would optimally choose to offer $\frac{z}{I}$ shares to each market maker at a price of $\mu\frac{z}{I} - \frac{\rho\sigma^2}{2}\frac{z^2}{I^2}$, which is where

multiple large traders, they may oversell in the first few periods.

each market maker is just indifferent between trading and not trading (see Appendix B.1.1 for full details on the dealer market).

There are a few points to highlight in the scenario where trading becomes extremely frequent and the trader is able to nearly achieve the first-best execution value. First, the expression for the optimal sequence of market orders reveals that total trading volume, $\sum_{t=1}^{N} |z_t|$, approaches infinity as $N \to \infty$. This comes from the increasingly excessive pattern of overselling and buying back in order to better take advantage of time-varying price impact and competition between market makers. This provides an complimentary explanation for the historical relationship between the rise in trading volume and the rise in trading frequency. Second, it is also important to point out that the trader needs less information to achieve the optimal selling value, V_d, with scheduling compared to a dealer market. In order to make effective take-it-or-leave-it offers to each of the I market makers, the trader needs to know the market makers' common risk aversion coefficient ρ. However, when the trader engages in scheduling and can trade sufficiently frequently, her liquidation value becomes arbitrarily close to the optimal value even without knowledge of ρ – knowing all market makers share a common risk aversion coefficient is sufficient. In a sense, with very frequent trading, the trader is able to use competition between market makers more effectively and push each market maker up against the limit of what he is willing to accept for trading.

Corollary 11. *As $N \to \infty$, the trader's liquidation value with scheduling, V_s, approaches $\mu z - \frac{\rho\sigma^2}{2}\frac{z^2}{I}$, which is the highest possible liquidation value given market maker preferences. The total trading volume also approaches infinity, i.e., $\sum_{t=1}^{N} |z_t| \to \infty$.*

To better understand how multiple trading rounds benefits the trader that implements scheduling, consider the environment where market makers are perfectly competitive, as described above, in that they do not take their own price impact into account. In this environment, multiple trading rounds do not improve liquidation. To see why, first notice that the equilibrium asset price in this scenario must be the same in all periods. If a market maker expects that the asset price is higher in the future, he would have incentive to buy an infinite amount of the asset in the current period and sell an infinite amount of the asset in the future. Moreover, the equilibrium asset price is determined by equalizing price and marginal benefit in the last period. This implies that the asset price in the last period, and thus the equilibrium asset price in every period, must be identical regardless of the number of trading periods (specifically equal to $\mu - \rho\sigma^2 \frac{z}{I}$). That is, trading frequency does not improve execution value in a competitive market.

Conversely, consider again when market makers recognize their price impact but the trader only has one period to liquidate. In this case, the equilibrium asset price is $p^* = \mu - \frac{\rho\sigma^2}{\gamma}\frac{z}{I}$ which is lower than the equilibrium asset price when market makers are perfectly competitive. Put differently, the trader prefers to trade with perfectly competitive market makers with a single period, but with more frequent trading and scheduling, she would prefer to face imperfectly competitive market makers with price impact. Specifically, the scheduling trader strictly prefers imperfect competition when $N \geq 3$, or when $N = 2$ and $I > 3$.

Corollary 12. *With a single trading period, the trader prefers a perfectly competitive market over a market with imperfect competition. With relatively frequent trading,*

the trader prefers a market with imperfect competition.

We acknowledge that there may be some concerns surrounding overselling and buying back as a realistic strategy, both because of the possible appearance of manipulation and excessive transaction costs. One simple restriction that we impose is that the trader is restricted from overselling, i.e., she does not buy in any period. This can be interpreted as a case where the trader has an exceedingly cautious interpretation of regulations governing manipulation. This case can also be interpreted as a simple, reduced-form way of capturing a desire to avoid excessive trading to reduce transaction costs.

The constraint on overselling leads the trader to sell her entire quantity, z, in the first period. There are three additional important observations. First, when overselling is banned, the trader prefers a market with competitive market makers, i.e., market makers who do not take their own price impact into account. This is, of course, in contrast to the unconstrained scheduling trader, who almost always prefers market makers with imperfect competition over perfect competition. The reason for this is that, in our model, overselling and buying back is the only way to achieve total selling values above what can be achieved with perfect competition. Second, for any given N, the scheduling trader still achieves a higher total selling value relative to the non-scheduling trader. Third, even though the constrained trader only trades for one period, her total selling value is still increasing in trading frequency, N.

Proposition 21. *When the trader utilizes scheduling with the constraint that $z_t \geq 0$ for all t, the equilibrium sequence of orders is $z_1 = z$ and $z_t = 0$ for $t = 2, \ldots N$. The trader's liquidation value is $\mu z - \rho\sigma^2 \frac{z^2}{I} - \frac{(1-\gamma)^{2N-1}\rho\sigma^2}{\gamma}\frac{z^2}{I}$, which is lower than*

the value with perfectly competitive market makers but higher than the value without scheduling.

3.4.4 Discussion

In this section, we compare execution with and without scheduling. For a given trading frequency N, the total selling value is strictly greater when the trader utilizes scheduling relative to when the trader does not. One way to see this is that the equilibrium selling sequence without scheduling could be implemented with scheduling, but is not because it is inferior. While the scheduling and non-scheduling trader both experience greater selling values as trading becomes more frequent, the scheduling trader typically achieves a greater value with only two periods (and always with three periods) than the non-scheduling trader with N periods. The exception is when there are two periods and only three market makers ($I = 3$), in which case the perfectly competitive value (the upper bound value for non-scheduling traders) yields a greater value than scheduling. This is the sense in which traders who utilized scheduling in infrequently traded floor-based markets with little competition between market makers may be worse off if they gave up scheduling to trade in frequently traded electronic markets.

In addition to scheduling values always exceeding non-scheduling values for a given trading frequency, the value increase from an increase in N is also greater for scheduling compared to non-scheduling. We can see this by comparing the difference between the upper-bound value and the value when $N = 2$ for each trader.[24]

[24]We use the upper-bound value because the value in each case approaches its respective upper-

The value increase for the scheduling trader is $\frac{\rho\sigma^2}{2}\frac{z^2}{I}\left(\frac{(2+\gamma)}{2\gamma}-1\right)$ and for the non-scheduling trader it is $\rho\sigma^2\frac{z^2}{I}\left(\frac{1}{\gamma(2-\gamma)}-1\right)$. These values represent how much each type of trader benefits in moving from infrequent to very frequent trading, and we think is representative of the move from floor-based to electronic trading while keeping the choice of scheduling fixed. The ratio of the scheduling to non-scheduling value increases is $\frac{(2-\gamma)^2}{4(1-\gamma(2-\gamma))} \geq \frac{9}{4}$, indicating that the improvement scheduling traders enjoy is more than twice as large, at a minimum, than non-scheduling traders. So, not only do scheduling traders have greater selling values than non-scheduling traders when trading is infrequent (i.e., $N = 2$), they realize even bigger gains with increased trading frequency.

3.5 Predatory Trader Extension

In this section, we present an extension of the model which introduces a predatory trader into the N-period model with scheduling from above. This case attempts to incorporate one particularly salient criticism of scheduling – the revealing of future trading intentions – in a stylized way into the model.

We introduce I_p risk-neutral predators into the N-period scheduling model. Specifically, each of the identical predators begin with zero units held of the risky asset and must end trading with zero units of the risky asset. This captures the idea that predatory traders only seek short-term profits from knowledge of future trading and not from holding risk. Just like the trader, the predator is also able to engage in

bound value as trading frequency becomes arbitrarily large.

scheduling.

We focus on the symmetric equilibrium where all predators submit the same sequence of market orders; otherwise the equilibrium definition is as provided in Definition 7. We explicitly solve for the trader and the predator's selling sequences with a single predator – please see Appendix B.2.1 for the general I_p predator case. The following proposition gives a closed-form expression of the trader's and the predator's selling sequences.

Proposition 22. *When $I_p = 1$, there exists a symmetric equilibrium where the trader's market order at period t is*

$$z_t = \left[\frac{\frac{(3-\gamma)^t}{(3-2\gamma)^t} - 1}{\frac{(3-\gamma)^N}{(3-2\gamma)^N} - 1} + \frac{1-(1-\gamma)^t}{1-(1-\gamma)^N} \right] \frac{z}{2(1-\gamma)^{N-t}} - \left[\frac{\frac{(3-\gamma)^{t-1}}{(3-2\gamma)^{t-1}} - 1}{\frac{(3-\gamma)^N}{(3-2\gamma)^N} - 1} + \frac{1-(1-\gamma)^{t-1}}{1-(1-\gamma)^N} \right] \frac{z}{2(1-\gamma)^{N-t+1}}$$

and the predator's market order at period t is

$$q_t = \left[\frac{\frac{(3-\gamma)^t}{(3-2\gamma)^t} - 1}{\frac{(3-\gamma)^N}{(3-2\gamma)^N} - 1} - \frac{1-(1-\gamma)^t}{1-(1-\gamma)^N} \right] \frac{z}{2(1-\gamma)^{N-t}} - \left[\frac{\frac{(3-\gamma)^{t-1}}{(3-2\gamma)^{t-1}} - 1}{\frac{(3-\gamma)^N}{(3-2\gamma)^N} - 1} - \frac{1-(1-\gamma)^{t-1}}{1-(1-\gamma)^N} \right] \frac{z}{2(1-\gamma)^{N-t+1}} .$$

We wish to point out that the predator can always achieve zero profits by not participating in trading. Thus, it must be that the predator can make a positive profit when he actively trades. Furthermore, in all symmetric equilibria, market

makers achieve perfect risk sharing in the end of the game by holding $\frac{z}{I}$ units of asset, just as in the model without a predator. Thus, the game is zero sum in the sense that what the trader pays in execution costs is equal to what the market maker earns in revenues. And, since the predator holds no position at the end of the game, the combination of the trader's and the predator's sequence of market orders yields a feasible sequence that the trader could implement without the existence of the predator. But since the aggregated sequence of the predator and the trader differs from the sequence the trader implements without the predator, it must be that the trader has a strictly lower execution value with the predator. Thus, market makers are better off with a predator, and both the predator and the market maker benefit from increased costs paid by the trader. Despite this, we confirm numerically that the scheduling trader with a predator achieves a liquidation value greater than the liquidation value achieved without scheduling. This follows from the fact that the oversell-and-buyback strategy is still implemented, which allows the trader to achieve a value greater than the non-scheduling upper bound, i.e. the value with perfectly competitive market makers.

Corollary 13. *The addition of a predator to the N-period model with scheduling results in a lower liquidation value for the trader. This lower liquidation value benefits the market makers – who make more than what they would in the model without a predator – and the predator. However, the scheduling liquidation value with the addition of a predator still exceeds the value achieved by a non-scheduling trader.*

3.6 Conclusion

The adoption of technology has led to faster and more sophisticated financial markets which has significantly altered trading. Our model identifies two important elements of this transition for large traders. First, the benefits of implementing reputation-based commitment strategies from the human-era of trading (e.g., preannouncement) in modern, fast electronic markets are significant. Second, traders who gave up human reputation-based commitment strategies in exchange for faster markets may only see a small benefit in execution or may even be worse off. These results suggest that there are important benefits of reimplementing what conceptually resemble human-based reputation strategies in the modern electronic era of trading.

While conceptually simple, there may be significant challenges in implementing electronic reputation-based strategies beyond the concerns over predatory trading (which we study). First, it is natural for humans to reveal their type to other humans through repeated face-to-face transactions but much more challenging for computers to reveal their type to other computers, largely because anonymous electronic trading makes it difficult to identify participants. And while trading signatures or patterns may be recognizable and allow some trader identities to be inferred, these patterns are easily imitated by others that seek to gain from beneficial terms of trade. Second, the set of skills required for human and electronic commitment strategies are quite different. Human commitment strategies rely on relationships and reputations, whereas electronic commitment strategies require technology and methods that prevent or disincentivize human intervention. Third, electronic trading has adopted a culture of secrecy and confidentiality. At least some of this culture may

be attributable to high-frequency trading firms that seek to protect small advantages in speed or other proprietary trading models. This culture may have spilled over to uninformed investors, who also appear to closely guard execution algorithms though they lack proprietary speed technology or information on fundamentals. Our paper suggests that these investors should at least consider the benefits that transparency and predictability provide in improving execution costs.

Chapter 4

Insider Trading When There May Not Be an Insider[1]

4.1 Introduction

Since the seminal work of Kyle (1985b), it has been extensively studied how market prices reveal inside information about asset values. Yet, an equally, if not more, important issue for both practioners and researchers is, whether such private information, or equivalently, insiders of a given asset, exist in the first place. After all, the prices cannot reveal information that does not exist. Then, the question is, how prices reveal the mix of private information about asset values and the information about its existence. This paper addresses this question.

We construct a dynamic market equilibrium model in which there are two types of private information, both can potentially be revealed by the asset price. One is the standard inside information, which is about the asset value; the other is information

[1]This chapter is based off of a paper of the same name co-authored with Dai Liang and Ming Yang (Dai et al. (2020b)).

about the existence of inside information; i.e., about the existence of the insider. Specifically, we consider an otherwise standard Kyle-Back model of insider trading, in which an insider exists with probability less than one. There are (potentially) three agents in the market: a market maker, an insider and a representative liquidity trader. If the insider exists, she observes the true value of the asset (either 0 or 1) and can trade strategically to profit from her inside information. The other two agents definitely exist. Similar to Anderson and Smith (2013), a monopolistic market maker[2] sets the asset price to minimize his[3] loss to the (potential) insider. The liquidity trader is a non-strategic agent, whose exogenous and random orders camouflage those from the (potential) insider. The liquidity trader's cumulative orders over time are modeled as a Brownian motion.

As our main result, we obtain an essentially unique equilibrium,[4] in which only the usual inside information is acquired by the informationally disadvantageous party (i.e., the market maker), and is reflected in the asset price, while the information about the existence of inside information is not revealed by the order flows. Specifically, the equilibrium has the following three characteristics. First, the pricing strategy of the market maker does not depend on the insider's existence probability. Instead, it is driven by the ratio of the probabilities of the insider receiving different information about the asset value (0 or 1) given her existence. This is because the market maker minimizes his loss to the insider and thus focuses on the situation

[2]In Section 4.5, we also analyze the situation where the market maker faces perfect competition.

[3]In this paper, we use "he" to refer to the market maker and "she" to refer to the insider.

[4]This equilibrium is unique in the sense that the payoffs of both the market maker and the insider in any other equilibrium (if there is any) are the same as in this one.

where the insider exists. Second, given the existence of the insider, the expected trading rate of (i.e., the average net order submitted by) the different types of insider is zero at each instant. To see this, suppose in equilibrium that the expected trading rate is positive[5] (i.e., different types of insider are buying the asset on average). Then the market maker would set the asset price to 1 to exploit the insider in the sale. The insider, at least the type who buys more than the other type, cannot make any profit at all at this price. In addition, she would be willing to reduce her trading rate so that she can divert the market maker's belief away from her type and thus benefit in the long run. This incentive to deviate contradicts the optimality of the equilibrium strategy. Third, the market maker never updates his belief about the probability of the insider's existence, since he is not able to statistically distinguish the insider's existence from her non-existence (where, by construction, the expected trading rate is also zero).

Our model provides an alternative perspective on the impact of stock market regulations on market liquidity. In reality, regulation measures increase the legal risk of insider trading and deter insiders from entering the market. To study their impact, models in which insiders definitely exist may not suffice. In our model, a regulation is naturally interpreted as an unexpected negative shock to the insider's existence probability,[6] and the market liquidity of an asset is measured by the inverse of price

[5]In this paper, we assign a positive sign to buy orders and a negative sign to sell orders. A symmetrical argument applies when the expected trading rate is negative.

[6]Kacperczyk and Pagnotta (2019) use data on SEC's investigations of illegal trading and show that insiders do internalize the legal risk. Facing a greater legal risk, an insider is more likely to refrain from insider trading or to trade on inside information for a shorter time. This motivates our modeling approach featuring a reduction in insiders' participation in trading activities, instead of an alternative in which all insiders still trade as strategically as before but are less likely to be

impact (Kyle's λ). As previously discussed, the price and thus the liquidity of the asset are driven by the ratio of the probabilities of the insider receiving different information about the asset value (0 or 1) given her existence. As a result, a regulation that merely reduces the probability of the insider's existence without affecting this ratio has no effect on market liquidity, while a regulation that disproportionately reduces the existence probabilities of the two types of insider may increase or decrease the ratio. Thus, depending on how this ratio is affected, a decrease in the insider's overall existence probability may improve or worsen market liquidity. This reconciles the seemingly puzzling empirical findings that a regulation may have opposite effects on liquidity for stocks in different exchanges and for stocks with different attributes.

Our model also yields an interesting prediction concerning the impact of competition among market makers, and complements the literature with an alternative rationale for the monopoly power of market makers in reality, such as specialists of NYSE.[7] As a special case of our model, when the insider definitely exists, Anderson and Smith (2013) establish that monopolistic and competitive market making yield the same equilibrium outcomes. However, we show that this conclusion no longer holds if there is uncertainty about the existence of the insider. Recall that in this case, the monopolist market maker would set the price to the expectation of asset value conditional on the insider's existence. And the insider, if she exists, can only achieve finite profit at that price. However, if the market maker faces perfect compe-

informed when facing such regulations. Different implications of these two approaches are discussed in Section 4.3.3.

[7] Existing empirical work, e.g., Venkataraman and Waisburd (2007), indeed finds that firms with designated dealers exhibit better market quality, that the introduction of designated dealers improves upon the terms of trade offered by public limit order books, and that the announcement of such introduction experience a positive and statistically significant average abnormal return.

tition as in many existing market microstructure models, then he instead has to set the price to the unconditional expectation of asset value. Since the insider's trading rate is unbounded, if she exists, she would be able to make infinite profit at that price. Thus, an equilibrium fails to exist, indicating the breakdown of the market.

We proceed as follows. Section 4.1.1 reviews relevant literature. Section 4.2 sets up the model. Section 4.3 solves for the equilibrium and analyzes its characteristics. Section 4.4 discusses how to use this model to analyze stock market regulations and their impact on market liquidity. Section 4.5 discusses extensions with competitive market making. Section 4.6 summarizes our main findings. All proofs are relegated to the Appendix.

4.1.1 Literature Review

Our paper is conceptually related to Banerjee and Breon-Drish (2020b) and Banerjee and Breon-Drish (2020a), who consider a dynamic setup in which an existing strategic trader, with no inside information initially, can decide when to acquire costly information. Whether an insider is informed is only known to herself in Banerjee and Breon-Drish (2020a), but is observable to the market maker in Banerjee and Breon-Drish (2020b). In contrast, our focus is on how the insider's existence, rather than whether she is informed, affects market making and the (potentially existing) insider's trading patterns. Allen and Gale (1992) and Back et al. (2017) also consider the possibility of an existing but uninformed strategic trader. If all agents are risk neutral, an insider with no inside information is equivalent to an insider who knows that the asset value equals its prior expectation. In this sense, their models incor-

porate the uncertainty due to inside information (about the asset value) but not the uncertainty due to information about the existence of inside information, while we incorporate the latter as well.

Several previous papers also address related topics. Chakraborty and Yilmaz (2004a) and Chakraborty and Yilmaz (2004b) consider uncertainty about the insider's existence in a discrete-time model in which a strategic trader definitely exists but can be either an informed insider or a liquidity trader. They show that the insider is bluffing in every equilibrium. That is, the insider may choose to trade against her information in order to hide her identity. Instead of studying whether the insider would bluff in equilibrium, our focus is on the implications for asset pricing and liquidity. Avery and Zemsky (1998) consider the uncertainty about the existence of a sequence of short-run insiders in a discrete-time framework with a publicly observable trading history. Uncertainty about the existence of the short-run insiders creates room for liquidity-driven trading to camouflage the content of inside information. The insider may trade according to the observed trading history rather than her private signal when uncertainty about the asset value is multidimensional. Cipriani and Guarino (2014) apply a model similar to that of Avery and Zemsky (1998) to financial market transaction data and provide an empirical methodology to gauge the importance of herding in actual financial markets. Unlike short-run insiders who submit orders of a fixed size, the insider in our model can choose any trading rate at any instant. This enables us to better explore the liquidity implications of uncertainty about the insider's existence.

Concerning the effects of such regulations on market liquidity, the existing lit-

erature provides mixed evidence. For example, Eleswarapu et al. (2004) and Chiyachantana et al. (2004) document an increase in market liquidity after Regulation Fair Disclosure (FD) came into effect, due to the reduction of the information component in bid-ask spreads. In contrast, Sidhu et al. (2008) document that after controlling for other factors, the adverse selection cost increases by 36%, suggesting a decrease in liquidity after the same regulation.[8] Both findings are supported by convincing stories. On one hand, the regulation makes insiders less likely to enter the markets and thus mitigates information asymmetry, suggesting an improvement in market liquidity. On the other hand, this may also give existing insiders greater informational advantages and reduce market liquidity. We offer a new perspective from which to evaluate the potentially asymmetric impacts of regulations on different types of inside information, together with a reconciliation of the discrepancy in existing empirical results concerning the effect of regulations on market liquidity, and some testable implications.

Lastly, many micro-structure models, including Glosten (1989), Kyle (1989), Viswanathan and Wang (2002), Back and Baruch (2013), etc., consider market making without assuming perfect competition among market makers. This paper offers an rationale for this practice. Specifically, when the insider may not exist, it is not feasible to assume perfect competitive market making since it may allow the insider to achieve infinite profit.

[8]In the data, the overall bid-ask spread decreases following the regulation. Yet the authors argue that this is because of the reduction in the inventory holding cost, while the adverse selection cost component of the bid-ask spread rises.

4.2 The Model

We consider a continuous-time insider trading model with uncertainty about the existence of the insider. A risky asset is traded continuously in the market from time 0. The true value of the asset is $\upsilon \in \{0, 1\}$, which is announced to the public at a random time τ, following an exponential distribution with parameter r. After the announcement, the price of the asset jumps to its true value and the game is essentially over.

4.2.1 Market Participants

There are (potentially) three agents in the market: a market maker, an insider and a liquidity trader. They are all risk neutral and discount future cash flows at a rate of zero. The insider may or may not exist in the market, and her ex ante probability of non-existence is $\hat{\pi}_u \in [0, 1]$. If the insider exists, she knows the true value of the asset and can trade strategically to profit from her information. We say that the insider, if she exists, is of type-υ if the true value of the asset is $\upsilon \in \{0, 1\}$. The liquidity trader's order flow is exogenous, modeled as $\sigma \cdot Z_t$, a Brownian motion with volatility σ. The market maker knows neither the true value of the asset nor whether the insider exists. He starts with a prior belief that the insider does not exist with probability $\hat{\pi}_u$, and that when no insider exists, the asset's true value is 1 with probability $\widehat{p}$; i.e., $\widehat{p} = \Pr(\upsilon = 1 | \textit{no insider}) \in [0, 1]$ at time 0. The history of the aggregate order flow (the sum of order flows from the liquidity trader and the potentially existing insider) is publicly observable. The market maker updates his

beliefs about the existence of the insider and the true value of the asset according to the history of the aggregate order flow. At each instant, the insider, if she exists, can submit market buy/sell orders of any size, and the market maker sets the price of the asset.

4.2.2 Market Making

The market maker acts as a monopolist and plays a zero-sum game with the potential insider.[9] We analyze an alternative setting of competitive market making in Section 4.5. In particular, the market maker and the potentially existing insider play a stage game at each instant of time. For ease of exposition, we present players' payoffs here and leave the details of the stage game and the discussion of the setting to Section C.1 in the Appendix. In the stage game at instant t, the market maker sets the price of the asset to $p_t \in [0, 1]$. And the insider, if she exists, chooses a trading rate $\theta_t \in (-\infty, +\infty)$, where $\theta_t > 0$ means "buying" and $\theta_t < 0$ means "selling". This results in an instantaneous payoff of $-(\upsilon - p_t)\theta_t dt$ to the market maker. If there is no insider, the market maker's payoff is 0. Here we can represent the insider's strategy with a trading rate $\theta_t \in \mathbb{R}$ because otherwise, a jump in the aggregate order flow reveals the insider's existence and is suboptimal for her. This representation simplifies our derivation and presentation.

The assumption that the market maker aims at minimizing his loss against the

[9]This assumption is meant to capture the direct conflict of interest between the market maker and the insider. The payoffs do not have to be exactly "zero-sum". All the analysis carries over as long as the sum of the payoffs is a constant. Hence, although in equilibrium the market maker could be losing money to the insider, we can always adjust the sum of the payoffs upward such that the market maker's expected payoff is positive, making him willing to participate.

potential insider rather than maximizing his profit against both the insider and the liquidity trader is justified based on several perceptions. First, for our purpose, this highlights the tension between the insider and the market maker due to asymmetric information. Second, if the market maker is instead assumed to maximize his trading profit, he can generate an infinite profit by setting $p = 1$ when he observes a buying order and $p = 0$ when he observes a selling order, which is trivial and contradicts common sense. Third, in practice, market makers are established to improve liquidity. Regulations prohibit them from front-running investors, and stock exchanges offer them rebates to encourage liquidity provision.[10] Finally, as will be discussed in Section 4.5, this setting nests the setting of competitive market making, which is standard in the literature, in the sense that they yield the same equilibrium outcomes when the insider definitely exists.

4.2.3 State Space

In principle, there are four possible scenarios: the insider exists and $v = 1$; the insider exists and $v = 0$; the insider does not exist and $v = 1$; and the insider does not exist and $v = 0$. The last two scenarios are not distinguishable from the market maker's perspective, because all orders come from the liquidity trader and hence the observation conveys no information about the asset. We therefore combine them and denote the combined scenario by u. We denote the first and the second scenarios by 1 and 0, respectively. In Scenario u, the market maker has to rely on his prior

[10]For instance, NYSE rebates its designated market makers \$0.0027 per share for providing National Best Bid and Offer and \$100 per stock per month if some requirements are met. NASDAQ has a similar policy for specialists.

belief of the asset's expected value $\Pr(v = 1|no\ insider) = \widehat{p}$. We denote the set of all scenarios by $\Omega = \{1, 0, u\}$. The market maker's belief over Ω at time t can be represented by an ordered pair $(\pi_{0,t}, \pi_{1,t})$, the probability assigned to Scenario 0 and 1, respectively. Then, $\pi_{u,t} = 1 - \pi_{0,t} - \pi_{1,t}$ is the probability assigned to Scenario u at time t. We will suppress the time subscripts if no confusion is engendered. Note that θ_t is the instantaneous trading rate chosen by the insider (if she exists) at time t. Then, the total of the cumulative orders up to time t is $Y_t = \int_0^t \theta_s ds + \sigma Z_t$, where $\theta_t = 0$ for all $t \geq 0$ in Scenario u. We require that θ is adapted to the filtration generated by Y and the scenario $\omega \in \Omega$. The market maker's belief over Scenario $\omega \in \Omega$ at $t < \tau$ is given by

$$\pi_{\omega,t} = \mathbb{E}\left[1_\omega | \left(Y_s\right)_{s \leq t}\right] \ ,$$

where 1_ω is the indicator function for Scenario ω. The equilibrium concept we consider here is a Markov equilibrium, in which π_0 and π_1 are the state variables. Specifically, the state space is

$$\Pi = \{(\pi_0, \pi_1) : \pi_0 \geq 0,\ \pi_1 \geq 0, \text{ and } \pi_0 + \pi_1 \leq 1\} \ ,^{11}$$

as shown by the shaded area in Figure 4.1. Note that the event that the insider exists with certainty; i.e., $\pi_0 + \pi_1 = 1$, is a subspace of it, represented by the -45-degree diagonal in Figure 4.1.

[11]Note that we do not have to include the market participants' positions in the set of state variables because of the risk neutrality assumption.

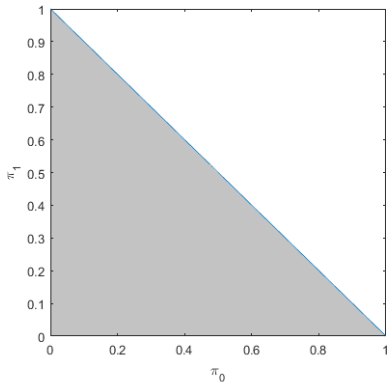


Figure 4.1: State Space

4.3 Equilibrium Analysis

In this section, we solve the model and obtain an equilibrium with the following characteristics. First, the market maker never updates his belief regarding the existence of the insider. Second, the expected trading rate of the insider, $\pi_{0,t} \cdot \theta_{0,t} + \pi_{1,t} \cdot \theta_{1,t}$, is always zero. Finally, the equilibrium is essentially unique in the sense that the payoffs for both the market maker and the insider in any other equilibrium must be the same as in this one.

4.3.1 The Market Maker's Belief Dynamics

Suppose the market maker believes that the potentially existing insider's trading strategy is $\theta_\upsilon : \Pi \to \mathbb{R}$, where $\upsilon \in \{0, 1\}$. Given the market maker's information at time t, the expected trading rate is

$$\phi\left(\pi_{0,t}, \pi_{1,t}\right) = \pi_{0,t} \cdot \theta_0\left(\pi_{0,t}, \pi_{1,t}\right) + \pi_{1,t} \cdot \theta_1\left(\pi_{0,t}, \pi_{1,t}\right) \ , \tag{4.1}$$

and the innovation/surprise in the cumulative aggregate order process at time t is

$$dY_t - \phi\left(\pi_{0,t}, \pi_{1,t}\right) dt \ .$$

Standard filtering theory shows that the updating of the market makers' belief is proportional to the innovation in the observed cumulative aggregate order process, which is given by

$$d\pi_{0,t} = \lambda_0\left(\pi_{0,t}, \pi_{1,t}\right) \cdot \left[dY_t - \phi\left(\pi_{0,t}, \pi_{1,t}\right) \cdot dt\right] \tag{4.2}$$

and

$$d\pi_{1,t} = \lambda_1\left(\pi_{0,t}, \pi_{1,t}\right) \cdot \left[dY_t - \phi\left(\pi_{0,t}, \pi_{1,t}\right) \cdot dt\right] \ , \tag{4.3}$$

where

$$\lambda_0\left(\pi_0, \pi_1\right) = \frac{\pi_0\left[\left(1-\pi_0\right) \cdot \theta_0\left(\pi_0, \pi_1\right) - \pi_1 \cdot \theta_1\left(\pi_0, \pi_1\right)\right]}{\sigma^2} \tag{4.4}$$

and

$$\lambda_1\left(\pi_0, \pi_1\right) = \frac{\pi_1\left[\left(1-\pi_1\right) \cdot \theta_1\left(\pi_0, \pi_1\right) - \pi_0 \cdot \theta_0\left(\pi_0, \pi_1\right)\right]}{\sigma^2} \ . \tag{4.5}$$

We need to augment equations (4.2) and (4.3) by defining the values of λ_0 and λ_1 on the boundary of state space Π. Note that once the market maker assigns zero probability to a scenario $\omega \in \{0, 1, u\}$, that probability remains zero forever regardless of the arrival of new information. Hence, we obtain

$$\lambda_0\left(0, \pi_1\right) = 0 \ , \tag{4.6}$$

$$\lambda_1(\pi_0, 0) = 0 \ , \tag{4.7}$$

and

$$\lambda_0(\pi_0, \pi_1) + \lambda_1(\pi_0, \pi_1) = 0 \text{ for all } (\pi_0, \pi_1) \text{ such that } \pi_0 + \pi_1 = 1 \ . \tag{4.8}$$

4.3.2 The Equilibrium

Let $V_0(\pi_0, \pi_1)$ and $V_1(\pi_0, \pi_1)$ denote the value functions of the type-0 and the type-1 insider, respectively. A Markov equilibrium is a five-tuple

$$(\theta_0(\pi_0, \pi_1), \theta_1(\pi_0, \pi_1), p(\pi_0, \pi_1), V_0(\pi_0, \pi_1), V_1(\pi_0, \pi_1))$$

satisfying the following conditions:

a) The market maker's beliefs π_0 and π_1 obey the law of motions given by equations (4.2) and (4.3).

b) Given the market maker's pricing strategy p, θ_υ maximizes V_υ, the type-υ insider's value function.

c) Given the insider's strategy θ_υ, p maximizes the market maker's expected payoff, $-\pi_0 V_0 - \pi_1 V_1$.

d) The market maker rationally anticipates the insider's trading rate θ_υ at each instant.

The equilibrium conditions are self-explanatory. Condition a) specifies the law of motion of the market maker's belief; b) states that the potential insider's strategy is

optimal; c) is the market maker's optimality condition; and d) is the market maker's rational expectation condition.

Given the insider's trading rates θ_0 and θ_1, the market maker's beliefs about the trading strategies of the two types of insider, and the insider's beliefs about the market maker's pricing strategy p, V_0 and V_1 satisfy the following HJB equations:

$$\begin{aligned} rV_0 = max_{\widetilde{\theta}_0} - p\widetilde{\theta}_0 + \lambda_0(\widetilde{\theta}_0 - \phi)\frac{\partial V_0}{\partial \pi_0} + \lambda_1(\widetilde{\theta}_0 - \phi)\frac{\partial V_0}{\partial \pi_1} \\ + \frac{1}{2}\lambda_0^2\sigma^2\frac{\partial^2 V_0}{\partial \pi_0^2} + \frac{1}{2}\lambda_1^2\sigma^2\frac{\partial^2 V_0}{\partial \pi_1^2} + \lambda_0\lambda_1\sigma^2\frac{\partial^2 V_0}{\partial \pi_0 \partial \pi_1} \ , \end{aligned} \tag{4.9}$$

and

$$\begin{aligned} rV_1 = max_{\widetilde{\theta}_1}(1-p)\widetilde{\theta}_1 + \lambda_0(\widetilde{\theta}_1 - \phi)\frac{\partial V_1}{\partial \pi_0} + \lambda_1(\widetilde{\theta}_1 - \phi)\frac{\partial V_1}{\partial \pi_1} \\ + \frac{1}{2}\lambda_0^2\sigma^2\frac{\partial^2 V_1}{\partial \pi_0^2} + \frac{1}{2}\lambda_1^2\sigma^2\frac{\partial^2 V_1}{\partial \pi_1^2} + \lambda_0\lambda_1\sigma^2\frac{\partial^2 V_1}{\partial \pi_0 \partial \pi_1} \ . \end{aligned} \tag{4.10}$$

The first term on the right-hand side of equation (4.9) represents the type-0 insider's instantaneous gain (loss). The remaining terms on the right-hand side reflect the long-term effects on her value function if she chooses trading rate $\widetilde{\theta}_0$. The interpretation of equation (4.10) is similar.

A major technical difficulty in solving for the equilibrium is the curse of dimensionality; i.e., the fact that the value functions are bivariate, as shown in the partial differential equations (4.9) and (4.10). This originates from the fact that there are two types of asymmetric information in our model. One is inside information, about

the asset value; the other is information about the existence of the inside information (i.e., about the existence of the insider). To study the impact of regulations that may simultaneously affect both types of information, such difficulty is inevitable. As the major technical contribution of this paper, we manage to reduce the dimension of the problem to one in order to solve for the equilibrium in closed form. This is based on our conjecture that in equilibrium, only the inside information is learned by the market maker and reflected in the asset price, which we verify later.

Specifically, we conjecture that in equilibrium, the expected trading rate of the different types of insider, $\phi(\pi_{0,t}, \pi_{1,t}) \equiv \pi_{0,t} \cdot \theta_0(\pi_{0,t}, \pi_{1,t}) + \pi_{1,t} \cdot \theta_1(\pi_{0,t}, \pi_{1,t})$, is zero at every instant t. Observe that if this conjecture is true, the market maker is not able to distinguish statistically the existence of the insider from her non-existence. This is because, when the insider does not exist, her trading rate is also zero, since by construction there is no insider trading at all. Lemma 4 formally establishes this argument.

Lemma 4. *At any instant t, if $\phi(\pi_{0,t}, \pi_{1,t}) = 0$, then $d\pi_{u,t} = 0$.*

That is to say, provided that the expected trading rate of the different types of insider is zero, in equilibrium, the market maker never updates π_u, which is his belief about the probability of the non-existence of the insider; i.e.,

$$\pi_{0,t} + \pi_{1,t} \equiv 1 - \pi_{u,t} = 1 - \hat{\pi}_u \tag{4.11}$$

at any instant t, where $\hat{\pi}_u$ is the prior probability of the non-existence of the insider. This restricts the evolution of the state variables to the one-dimensional subspace

$\{(\pi_0, \pi_1) : \pi_0 + \pi_1 = 1 - \hat{\pi}_u\}$, and serves our purpose of dimensionality reduction.

Proposition 23 confirms that the condition of Lemma 4 indeed holds in equilibrium, and characterizes the equilibrium.

Proposition 23. *Given $\hat{\pi}_u$, the prior probability that the insider does not exist, this game admits an equilibrium in which the market maker never updates his belief regarding the existence of the insider; i.e., $\pi_{u,t} = \hat{\pi}_u$ for all $t \geq 0$. In this equilibrium, the potentially existing insider's expected trading rate is always zero; i.e.,*

$$\phi_t \equiv \pi_{0,t}\theta_{0,t} + \pi_{1,t}\theta_{1,t} = 0$$

for all $t \geq 0$.

The market maker prices the asset at

$$p(\pi_0, \pi_1) = \frac{\frac{\pi_1}{\pi_0}}{1 + \frac{\pi_1}{\pi_0}} \ . \tag{4.12}$$

The type-1 insider trades at rate

$$\theta_1(\pi_0, \pi_1) = \frac{\sigma r^{1/2}\left(\frac{\pi_1}{\pi_0} + 1\right)}{2 \cdot \frac{\pi_1}{\pi_0}} \Phi'\left(\Phi^{-1}\left(\frac{2 \cdot \frac{\pi_1}{\pi_0}}{1 + \frac{\pi_1}{\pi_0}} - 1\right)\right) \tag{4.13}$$

and the type-0 insider trades at rate

$$\theta_0(\pi_0, \pi_1) = -\frac{\pi_1}{\pi_0}\theta_1(\pi_0, \pi_1) \ , \tag{4.14}$$

where $\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-y^2} dy$. *The type-1 insider's value function is*

$$V_1(\pi_0, \pi_1) = \frac{\sigma}{\left(\frac{\pi_1}{\pi_0} + 1\right) r^{1/2}} \Phi^{-1}\left(\frac{2 \cdot \frac{\pi_1}{\pi_0}}{1 + \frac{\pi_1}{\pi_0}} - 1\right) + \frac{\sigma}{r^{1/2}} \Phi'\left(\Phi^{-1}\left(\frac{2 \cdot \frac{\pi_1}{\pi_0}}{1 + \frac{\pi_1}{\pi_0}} - 1\right)\right) . \tag{4.15}$$

The type-0 insider's value function is

$$V_0(\pi_0, \pi_1) = \frac{\frac{\pi_1}{\pi_0}\sigma}{\left(\frac{\pi_1}{\pi_0} + 1\right) r^{1/2}} \Phi^{-1}\left(\frac{2}{1 + \frac{\pi_1}{\pi_0}} - 1\right) + \frac{\sigma}{r^{1/2}} \Phi'\left(\Phi^{-1}\left(\frac{2}{1 + \frac{\pi_1}{\pi_0}} - 1\right)\right) . \tag{4.16}$$

Moreover, as t *goes to infinity,* π_1 *converges to* $1 - \hat{\pi}_u$ *with probability one if the type-1 insider exists, and converges to* 0 *with probability one if the type-0 insider exists. Finally, the payoffs for both the market maker and the insider in any other equilibrium (if there is any) are the same as in this one.*

Most results in Proposition 23 are self-explanatory. Notice the effect of dimension reduction — The functions from equations (4.12) to (4.16) depend on (π_0, π_1) only through $\frac{\pi_1}{\pi_0}$, the ratio of the existence probabilities of the type-1 insider and the type-0 insider. The belief convergence result (i.e., the convergence of π_1 as $t \to \infty$), indicates that the market maker will eventually learn about the inside information conditional on its existence, but not about its existence in the first place.

Figure 4.2 provides a visual illustration of Proposition 23. As in Figure 4.1, the lower triangle denotes the state space Π and each point corresponds to a state (π_0, π_1). Consider state A with coordinates $(0.1, 0.3)$ for example, where $\frac{\pi_1}{\pi_0} = 3$. In this state, the asset price is $\frac{3}{1+3} = \frac{3}{4}$ according to equation (4.12). This is because, the market maker "rescales" (π_0, π_1) to the conditional probability pair $\left(\frac{\pi_0}{\pi_0+\pi_1}, \frac{\pi_1}{\pi_0+\pi_1}\right) =$

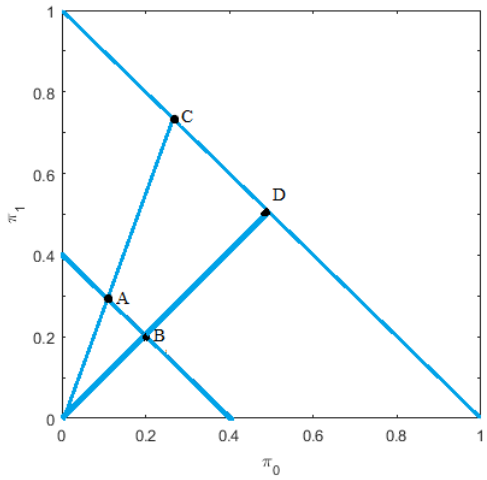


Figure 4.2: Visual Illustration of Proposition 23

$\left(\frac{1}{4}, \frac{3}{4}\right)$; i.e., state C with the same $\frac{\pi_1}{\pi_0}$ as that of state A but located on the minus 45-degree diagonal of the square corresponding to $\pi_0 + \pi_1 = 1$ (i.e., the insider definitely exists), and sets the price to the asset's expected value under it, which is $\frac{\pi_0}{\pi_0+\pi_1} \cdot 0 + \frac{\pi_1}{\pi_0+\pi_1} \cdot 1 = \frac{3}{4}$. Analogous arguments also explain why the other equilibrium objects in Proposition 23 also depend only on $\frac{\pi_1}{\pi_0}$. For each type $\upsilon \in \{0, 1\}$ of the insider, her instantaneous trading rate θ_υ and continuation value V_υ in that state are given by equations (4.13) to (4.16), respectively. It can be verified that the insider's expected trading rate, $\phi \equiv \pi_0\theta_0 + \pi_1\theta_1$, is indeed 0. Moreover, as will be clear, the same is true for our liquidity measure introduced later in Section 4.4.

In addition, Proposition 23 articulates that the market maker never updates his belief regarding the existence of the insider (i.e., $\pi_{u,t} = \hat{\pi}_u$), or equivalently, equation (4.11) holds for all $t \geq 0$. Graphically, if a state is ever reached, the belief of the market maker always evolves on the minus 45-degree line passing it, which is the

line $\pi_0 + \pi_1 = 0.4$ if that state is A. This represents the reduction of dimension due to Lemma 4. As evident from equation (4.12), on the line, the closer a state is to the horizontal axis (i.e., the smaller is $\frac{\pi_1}{\pi_0}$), the lower is the price charged by the market maker in the corresponding state, since he believes $\upsilon = 1$ to be less likely. For instance, consider state B with coordinates $(0.2, 0.2)$. Since in that state, $\frac{\pi_1}{\pi_0} = 1 < 3$, the asset price is $\frac{1}{1+1} = \frac{1}{2}$, which is lower than that in state A, $\frac{3}{4}$. Moreover, the belief convergence result in Proposition 23 indicates that if the type-1 insider exists, the market maker's belief would converge to $(0, 0.4)$ with probability one as $t \to \infty$; if the type-0 insider exists, the market maker's belief would converge to $(0.4, 0)$ with probability one as $t \to \infty$.

We present the logic of Proposition 23 in four steps. First, assuming that the insider's expected trading rate is zero, we explain the fact that the market maker's pricing decision is given by equation (4.12) and the belief convergence results. Second, we exploit the zero-sum feature of this game to show that the equilibrium payoffs for the insider and the market maker are unique. Third, we verify the assumption in the first step based on the result of the second step. Finally, we derive the insider's trading strategies in equilibrium.

Step 1: Suppose the insider's expected trading rate is zero. By Lemma 4, the market maker will not update his belief about the insider's existence in equilibrium. Moreover, since the market maker is playing a zero-sum game with the insider, he is indifferent to setting any price when the insider does not exist. Thus, with uncertainty about the insider's existence, the market maker sets the price to maximize his expected payoff, conditional on the existence of the insider. Hence, he "rescales" the

probability pair (π_0, π_1) to the conditional probability pair $\left(\frac{\pi_0}{\pi_0+\pi_1}, \frac{\pi_1}{\pi_0+\pi_1}\right)$, and sets the price to the asset's expected value under it, which is $\frac{\pi_0}{\pi_0+\pi_1} \cdot 0 + \frac{\pi_1}{\pi_0+\pi_1} \cdot 1 = \frac{\pi_1}{\pi_0+\pi_1}$.[12] Since the market maker updates his belief only in terms of $\frac{\pi_1}{\pi_0}$, the belief-convergence result follows from the standard logic that the insider trades in the direction that drives the asset price to the asset fundamental.

Step 2: To see why the equilibrium payoffs are unique, consider a general pricing strategy p^g for the market maker and a general trading strategy θ^g for the insider. The market maker's pricing strategy p^g specifies the price at time t contingent on the entire path of the order process $(Y_s)_{s \leq t}$. The insider's trading strategy θ^g specifies the trading rate at time t contingent on her type $\upsilon \in \{0, 1\}$, the entire path of the order process $(Y_s)_{s \leq t}$, her own trading history and the market maker's pricing history $(p_s)_{s \leq t}$. Consider the "one-shot" zero-sum game at time zero in which each player maximizes his/her ex ante expected payoff given her/his opponent's strategy (i.e., p^g or θ^g). Note that the Nash equilibria of this "one-shot" game do not require sequential rationality and thus nest the Markov equilibria we consider here. We show that all Nash equilibria of this "one-shot" game are payoff equivalent. Thus, all Markov equilibria of the insider trading game are also payoff equivalent to the one we obtained. Let $\mathcal{V}$ be the ex ante expected payoff for the insider in the "one-shot" game. Since this game is zero-sum, the market maker's ex ante expected payoff is $-\mathcal{V}$. If two strategy pairs (θ^g, p^g) and $(\bar{\theta}^g, \bar{p}^g)$ are both equilibria, by the optimality

[12]Section 4.5.2 discusses why this is the case.

condition of the insider in the equilibrium (θ^g, p^g),

$$\mathcal{V}(\theta^g, p^g) \geq \mathcal{V}(\bar{\theta}^g, p^g) \ .$$

The optimality condition of the market maker in the equilibrium $(\bar{\theta}^g, \bar{p}^g)$ also implies

$$-\mathcal{V}(\bar{\theta}^g, \bar{p}^g) \geq -\mathcal{V}(\bar{\theta}^g, p^g) \ .$$

Combining the above two inequalities leads to

$$\mathcal{V}(\theta^g, p^g) \geq \mathcal{V}(\bar{\theta}^g, \bar{p}^g) \ .$$

By interchanging (θ^g, p^g) and $(\bar{\theta}^g, \bar{p}^g)$, we obtain the reverse inequality

$$\mathcal{V}(\bar{\theta}^g, \bar{p}^g) \geq \mathcal{V}(\theta^g, p^g) \ .$$

Thus,

$$\mathcal{V}(\theta^g, p^g) = \mathcal{V}(\bar{\theta}^g, \bar{p}^g) \ .$$

Note that this argument holds for any prior belief (π_0, π_1). Thus, all the equilibria that we consider are essentially unique in the sense that they are payoff-equivalent.

Step 3: We argue that the insider's expected trading rate is zero by contradiction.[13] Consider the situation where the insider's expected trading rate is strictly positive (i.e., she is buying the asset on average) at some state in a Markov equi-

[13]See Lemma 10 in the Appendix for a rigorous argument.

librium. In that case, the type of insider that has the larger trading rate must be trading at a positive rate. We show that both her instantaneous payoff and continuation value can be improved by reducing her current trading rate, contradicting its optimality. Note that the market maker's pricing decision does not affect his belief-updating process. Thus, given the positive expected trading rate, it is optimal for him to set the asset price to 1 to maximize his instantaneous payoff. Thus, the instantaneous payoff for the type of insider that has the larger trading rate is negative if she is type-0 and zero if she is type-1, and can be improved by reducing her equilibrium trading rate if she is type-0 and remains unchanged if she is type-1. Regarding her continuation value, due to the uniqueness of the equilibrium payoffs established in Step 2, this equilibrium shares the same value functions as the value functions in the equilibrium characterized in Proposition 23. Thus, the insider's value functions strictly increase when the market maker's belief moves away from her true type. The insider that has the larger trading rate then strictly benefits from reducing her equilibrium trading rate, because this shifts the market maker's belief towards the other type, strictly increasing her continuation value in the next instant. This yields the desired contradiction. A symmetrical argument leads to the same contradiction if the expected trading rate is strictly negative.[14]

Step 4: To determine the insider's equilibrium trading strategies θ_0 and θ_1, notice that the insider's trading rates enter value functions (4.9) and (4.10) as linear terms. Thus, the optimality condition of the insider requires her to be indifferent

[14]Indeed, as long as the dimensionality of the problem can be reduced so that the price p can be used as the state variable, we can obtain the result that the insider's expected trading rate is zero without invoking the property of the zero-sum game, since our problem reduces to the special case studied in Theorem 1 of Back and Baruch (2004).

to trading at any rate. If not, the insider may pick an arbitrarily large trading rate to achieve infinite payoff. These indifference conditions allow us to solve for λ_0 and λ_1, the market maker's belief sensitivities to the order innovation. Notice that for a fixed state (π_0, π_1), λ_1 and λ_0 are functions of θ_0 and θ_1. We can then back out the insider's equilibrium trading strategies θ_0 and θ_1 from λ_1, λ_0, and the condition that the insider's expected trading rate is zero.

4.3.3 Discussion

Note that our equilibrium characterization in Proposition 23 is similar to those of Back and Baruch (2004) and Anderson and Smith (2013).[15] This is not a coincidence, because their models correspond exactly to our special case in which $\hat{\pi}_u = 0$; i.e., $\pi_{0,t} + \pi_{1,t}$ is fixed at 1 for all t. Yet, the belief-convergence result in Proposition 23 indicates a fundamental difference between an existing but uninformed insider and a non-existent insider.

Consider an insider that definitely exists but is uninformed. The fact that she is uninformed is her private information, so she can still trade strategically with the market maker and make a profit. Indeed, since her knowledge about the asset value υ is identical to the prior of the market maker, her trading strategy is identical to that of a (synthetic) insider who is informed that the asset value υ is $\hat{p}$. In this sense, the uncertainty about whether an existing insider is uninformed is essentially

[15]Back and Baruch (2004) consider a model with competitive market making while Anderson and Smith (2013) consider a setup equivalent to monopolistic market making. Both settings assume that the insider exists for sure. Anderson and Smith (2013) show the equivalence between the two settings in terms of their equilibrium outcomes.

equivalent to the uncertainty about whether the asset value equals the prior expectation. Therefore, such uncertainty is essentially a special case of the uncertainty about the asset value, and thus can be easily accommodated by adding one type of insider to a standard insider trading model featuring uncertainty only about the asset value. Given that, following the standard logic that the insider trades in the direction that drives the asset price to the asset fundamental, the market maker in this case eventually learns both whether the insider is informed and the asset value.

In addition, our model provides different implications concerning liquidity. If the insider definitely exists, the more likely the insider is uninformed, the less information asymmetry there is between the insider and the market maker, and the higher the average liquidity is. Our model predicts that the liquidity and the asset price remain the same if the existence probabilities of the type-1 and the type-0 insider change proportionately. The liquidity and the asset price could increase or decrease if the existence probabilities of the two types of insider change disproportionately. In Section 4.4, we elaborate this result and further study the effect of regulations on market liquidity.

Moreover, note from Proposition 23 that when $\hat{\pi}_u = 0$, the price charged by the monopolistic market maker is $p(\pi_0, \pi_1) = \pi_1 \cdot 1 + \pi_0 \cdot 0$. Were he instead facing perfect competition, he would also charge exactly the same price. Indeed, as shown in Anderson and Smith (2013), when the insider definitely exists, monopolistic and competitive market making yield the same equilibrium; i.e. the insider's trading strategy and the market maker's pricing strategy are the same. However, Proposition 23 indicates the failure of such equivalence when the insider may not exist; i.e., when

$\hat{\pi}_u > 0$. Specifically, the monopolistic market maker would price the asset at its expected value conditional on the existence of the insider, $\frac{\pi_1}{\pi_0+\pi_1}$, while a market maker facing perfect competition has to price the asset at its unconditional expected value, $\pi_1 \cdot 1 + \pi_0 \cdot 0 + (1 - \pi_1 - \pi_0) \cdot \hat{p}$. We show in Section 4.5 that such a discrepancy rules out reasonable equilibria in a competitive market making setup, indicating a market failure.

4.4 The Effect of Regulations on Market Liquidity

As the first application of our model, in this section, we analyze the effect of regulations on market liquidity. Stock market regulations aimed at reducing information asymmetry and insider trading have important implications for market microstructure. Many empirical papers have examined how market liquidity is affected by stock market regulations and have found that liquidity may change in either direction for different stocks, as reviewed in Section 4.4.1 below. However, few theoretical papers analyze this issue. One potential reason is that standard models of insider trading do not accommodate shocks to the insider's existence probability. In our model, a regulation can be naturally viewed as a negative shock to the insider's existence probability $\pi_0 + \pi_1$. Section 4.4.2 introduces the measure of liquidity in the model, explores its determinants, and provides a reconciliation between the mixed empirical findings reviewed in Section 4.4.1. In Section 4.4.3, we make additional assumptions regarding firm attributes to make directional predictions about the change in the

market liquidity of stocks.

4.4.1 Mixed Empirical Findings

Two competing hypotheses are formulated concerning the effect of regulations on liquidity. According to the first hypothesis, regulation reduces the probability of insider trading. Thus, the information content in each post-regulation order is reduced. This leads to a deeper market with better liquidity. The second hypothesis, the chilling effect, suggests that firms may disclose information more cautiously after a regulation because of the newly posted restrictions. This makes existing inside information longer-lived and more valuable, and thus reduces market liquidity.

Many empirical papers have tested these hypotheses with various regulations in major stock markets. For instance, Eleswarapu et al. (2004) and Chiyachantana et al. (2004), investigating Regulation Fair Disclosure (FD) using NYSE data, document an increase in market liquidity after the imposition of the regulation. In contrast, Sidhu et al. (2008), studying FD using NASDAQ data, find a decrease in market liquidity. Hagerman and Healy (1992), studying Security Acts Amendments in 1964, find no effect on market liquidity. Frino and Jones (2005), examining the effect of Australia's Statement on Cash Flows, document an increase in liquidity among the stocks of less transparent firms and ambiguous effects on liquidity among those of more transparent firms. Overall, it remains unclear whether stock market regulations improve market liquidity.

Admittedly, differences among empirical results can be attributed to the different nature of each regulation or the relative strength between competing forces in two

hypotheses. Yet this explanation might not be satisfactory for several reasons. First, the requirements imposed by various regulations do not differ fundamentally in terms of their mechanisms in deterring insider trading. They can be roughly classified into two categories. The first category requires potential insiders to disclose more about their trades and positions. The second category aims at reducing the "information gap" between insiders and other investors. It is puzzling why similar regulation requirements might affect market liquidity differently. Second, some regulations set mandatory requirements on the materials that must be disclosed. Thus, the chilling effect takes no bite, and the liquidity reduction of some stocks documented in empirical studies still calls for a satisfactory explanation. Third, variations in liquidity adjustments at the firm level after a regulation cannot be explained by current hypotheses.

4.4.2 A Reconciliation of Empirical Findings

Here we introduce the notion of liquidity and provide a reconciliation of the aforementioned mixed empirical findings. Following standard market microstructure models pioneered by Kyle (1985b), we use market depth, the reciprocal of price sensitivity over order innovation, as the liquidity measure in each state. Specifically, we define λ_p to be the price sensitivity over the order innovation. That is, $\lambda_p(\pi_0, \pi_1)$ satisfies

$$dp(\pi_0, \pi_1) = \lambda_p(\pi_0, \pi_1) \cdot [dY_t - \phi(\pi_0, \pi_1) \cdot dt] \ ,$$

where the expected trading rate ϕ is defined by equation (4.1). The liquidity measure, the market depth in state (π_0, π_1), is defined to be

$$L(\pi_0, \pi_1) = [\lambda_p(\pi_0, \pi_1)]^{-1}. \tag{4.17}$$

Intuitively, if λ_p, the per unit price impact of the order innovation $dY_t - \phi(\pi_0, \pi_1) \cdot dt$, is small, it must be the case that market depth; i.e., liquidity, is large.

Recall that there are two types of asymmetric information in our model. One is inside information, as captured by π_1/π_0, which is about the asset value; the other is information about the existence of inside information; i.e., about the existence of the insider, as captured by $\pi_0 + \pi_1$. These are the two potential channels through which regulations on insider trading affect market liquidity. Thus, as a key contribution, our framework, which incorporates both types of asymmetric information, provides a decomposition of the impact of such regulations on market liquidity, as stated in Proposition 24.

Proposition 24. *If two states share the same ratio of existence probabilities between the type-0 and the type-1 insider, then they share the same liquidity L. That is, $L(\pi_0, \pi_1)$ is a function of π_1/π_0. Moreover, L strictly increases in π_1/π_0 for $\pi_1/\pi_0 \geq 1$ and strictly decreases in π_1/π_0 for $\pi_1/\pi_0 \leq 1$. Thus, liquidity L achieves the minimum at $\pi_1/\pi_0 = 1$.*

By definition, market regulations aimed at driving insiders out of the market reduce the existence probability of insiders. But do they necessarily improve market liquidity? The answer is no. The first statement of Proposition 24 formulates that

conditional on π_1/π_0, the content of the inside information, changing $\pi_0 + \pi_1$, the information about the existence of insiders, has no effect on market liquidity. To see why, recall from Proposition 23 that the market maker sets the price to maximize his expected payoff conditional on the existence of the insider. This makes the asset price unresponsive to the existence probability of the insider, $\pi_0 + \pi_1$, conditional on π_1/π_0.

As an illustration, in Figure 4.2, since states A and C share the same π_1/π_0, although a regulation that transforms the market from state C to state A reduces the existence probability of the insider from 1 to 0.4, it has no impact on the price and the liquidity of the asset. An analogy applies to a regulation that transforms the market from state D to state B.

Given the first statement, the second statement of Proposition 24 predicts how market liquidity is affected by inside information π_1/π_0. The market maker is most uncertain about whether the asset value $v = 1$ or 0 when $\pi_1/\pi_0 = 1$. Therefore, his belief about the asset value, and thus the price he charges, is most responsive to order innovations, and such responsiveness falls when π_1/π_0 is further away from 1. Hence, a regulation that reduces the existence probability of the insider, $\pi_0 + \pi_1$, may lead to an increase or a decrease in π_1/π_0, and thus may result in an increase or decrease in market liquidity. This reconciles the mixed empirical findings discussed in the previous subsection.

As an illustration, in Figure 4.2, all the points on Line AB share the same existence probability of the insider, $\pi_0 + \pi_1 = 0.4$, and the minus-45-degree diagonal (i.e., Line CD) collects all the points with the insider definitely existing. All regulations

that transform the market from points on Line CD to points on Line AB reduce the existence probability of the insider from 1 to 0.4, but could have different or even opposite effects on the market price and liquidity of the asset. For example, transformation from state C to state B reduces π_1/π_0 from 3 to 1, and thus results in a decrease in the price (from 3/4 to 1/2) and market liquidity L of the asset. The effects of this transformation are exactly opposite to that from state D to state A.

It is worth noting that it is the two types of asymmetric information that jointly make such reconciliation possible. If the non-existent insider is instead modeled as existing but uninformed, then there is essentially only one type of asymmetric information, and regulations reducing the probability that the insider is informed can only improve liquidity. This is because, the more likely the insider is uninformed, the less information asymmetry there is between the insider and the market maker.

4.4.3 Predictions Based on Firm Attributes

In our model, a regulation aimed at restraining insider trading can be considered as a negative shock to the insider's existence probability. By Proposition 24, such a shock may affect market liquidity in either direction. To make directional predictions about the liquidity change after regulation, additional assumptions are needed. Specifically, we assume that the insider, if she exists, receives her inside information from a specific channel. Depending on the underlying asset's attributes, inside information from various channels has different probabilities of delivering positive news ($\upsilon = 1$) about the asset. A regulation is further interpreted as an unexpected reduction in the probability of each channel delivering inside information, by which it decreases

the insider's existence probability. This implies that if a regulation more strongly affects the channel that is more likely to deliver positive news about a firm, then its type-1 insider would be less likely to exist after the regulation relative to its type-0 insider, and vice versa.

For concreteness, henceforth we consider the following two-channel example, in which the channels are labeled channel f ("financial status") and channel r ("R&D progress"), respectively. The insider, if she exists, can receive inside information from either channel. The scenario that the insider exists and receives inside information $\upsilon \in \{0, 1\}$ from channel $j \in \{f, r\}$ is denoted by υj. Together with u, the scenario in which the insider does not exist, there are now a total of five scenarios: $0f$, $1f$, $0r$, $1r$ and u. The market maker believes at instant t that Scenario υj is true with probability $\pi^j_{\upsilon,t}$, and that Scenario u is true with probability $\pi_{u,t} = 1 - \sum_{\upsilon,j} \pi^j_{\upsilon,t}$.

Note that this is an innocuous extension of our baseline model in the sense that Scenarios υf and υr here together correspond to Scenario $\upsilon \in \{0, 1\}$ there, so that $\pi^f_{\upsilon,t} + \pi^r_{\upsilon,t} = \pi_{\upsilon,t}$. Moreover, since the market maker and the insider care only about the true value υ of the asset rather than the channel that delivers the inside information, in equilibrium, it must be the case that the insider's trading strategy does not depend on channels, so that the market maker's belief-updating processes, π_1 and π_0, are the same as in the baseline model.[16] Given that, our measure of liquidity, L, is still given by equation 4.17.

In this extension, a regulation can be further interpreted as a negative shock to the probability that each channel delivers inside information, should it be $\upsilon = 1$ or

[16]This is proved in Lemma 11 in the Appendix.

$\upsilon = 0$. Specifically, each regulation can be decomposed into two elementary ones, each reducing the probability that a specific channel $j \in \{f, r\}$ delivers news by Δ_j, whether the news is $\upsilon = 1$ or $\upsilon = 0$. Definition 8 formalizes such decomposition.

Definition 8. *For channel $j \in \{f, r\}$, a j-regulation of magnitude $\Delta_j \in [0, 1]$ at instant t is an unexpected exogenous shock to $\pi^j_{\upsilon,t}$ such that the market maker's belief right after the shock is $\pi^j_{\upsilon,t+} = (1 - \Delta_j)\,\pi^j_{\upsilon,t}$ for $\upsilon \in \{0, 1\}$. A (Δ_f, Δ_r)-regulation at instant t is the combination of an f-regulation of magnitude Δ_f and an r-regulation of magnitude Δ_r at instant t.*

Suppose that the underlying asset is a stock, and that channel f delivers inside information about the firm's financial status, while channel r delivers inside information about the firm's R&D progress. For a mature firm that is typically financially stable, a piece of positive financial information needs to be strong enough to trigger stock price fluctuations. In this sense, channel f is (believed by uninformed market participants, such as the market maker, to be) relatively less likely to produce positive news than channel r (i.e., $\pi^f_{1,t}/\pi^f_{0,t} < \pi^r_{1,t}/\pi^r_{0,t}$). For a growth firm, in contrast, channel f is relatively more likely to produce positive news than channel r (i.e., $\pi^f_{1,t}/\pi^f_{0,t} > \pi^r_{1,t}/\pi^r_{0,t}$). Proposition 25 shows that the same (Δ_f, Δ_r)-regulation at instant t has opposite effects on the price p and the market liquidity L of the stocks of two such firm types.

Proposition 25. *Consider a (Δ_f, Δ_r)-regulation at instant t.*

1. *For a stock with $\pi^f_{1,t}/\pi^f_{0,t} = \pi^r_{1,t}/\pi^r_{0,t}$, we have $p_{t+} = p_t$ and $L_{t+} = L_t$;*

2. *For a stock with $\pi^f_{1,t}/\pi^f_{0,t} > \pi^r_{1,t}/\pi^r_{0,t}$,*

(a) $p_{t+} < p_t$ *if and only if* $\Delta_f > \Delta_r$;

(b) *If* $\min\{p_{t+}, p_t\} \geq \frac{1}{2}$, *we have* $L_{t+} > L_t$ *if and only if* $\Delta_f < \Delta_r$;

(c) *If* $\max\{p_{t+}, p_t\} \leq \frac{1}{2}$, *we have* $L_{t+} > L_t$ *if and only if* $\Delta_f > \Delta_r$;

3. *For a stock with* $\pi^f_{1,t}/\pi^f_{0,t} < \pi^r_{1,t}/\pi^r_{0,t}$,

(a) $p_{t+} > p_t$ *if and only if* $\Delta_f > \Delta_r$;

(b) *If* $\min\{p_{t+}, p_t\} \geq \frac{1}{2}$, *we have* $L_{t+} > L_t$ *if and only if* $\Delta_f > \Delta_r$;

(c) *If* $\max\{p_{t+}, p_t\} \leq \frac{1}{2}$, *we have* $L_{t+} > L_t$ *if and only if* $\Delta_f < \Delta_r$.

Recall that for each channel, a regulation equally affects its probability of delivering positive news and negative news. If both channels are equally likely to deliver positive news about a firm (i.e., $\pi^f_{1,t}/\pi^f_{0,t} = \pi^r_{1,t}/\pi^r_{0,t}$), then regardless of its magnitude, a regulation does not affect π_1/π_0, the relative likelihood of the inside information being positive (i.e., of the insider being of type-1), and thus does not affect the price and the liquidity of the stock. This interprets the first statement of Proposition 25.

Now we provide the interpretation for Statements 2a and 2b, and that for all the other statements is analogous. Suppose channel f has a relatively higher probability of delivering positive news (i.e., if $\pi^f_{1,t}/\pi^f_{0,t} > \pi^r_{1,t}/\pi^r_{0,t}$). Concerning Statement 2a, if a regulation affects channel f more (i.e., if $\Delta_f > \Delta_r$), it reduces the relative likelihood of the inside information being positive. Thus, by Proposition 23, the asset price decreases after the regulation. The contrapositive of the "only if" argument can be understood analogously.

Concerning Statement 2b, recall from Proposition 24 that liquidity is lowest when $\pi_1/\pi_0 = 1$; i.e., when the asset price is $\frac{1}{2}$, and is enhanced when π_1/π_0 moves further away from 1. Thus, provided that the stock price before and after the regulation is always above $\frac{1}{2}$ (i.e., $\pi_1/\pi_0 > 1$), if channel f has a higher probability of delivering positive news before the regulation, then a regulation that affects it less would further increase π_1/π_0 and thus improve the liquidity of the stock. Again, the contrapositive of the "only if" argument can be understood analogously. Note that if the asset price moves across $\frac{1}{2}$ after the regulation, the direction of the change in liquidity is ambiguous. However, as long as the regulation does not drastically change asset prices, this circumstance is rare.

Although, by definition, a regulation lowers the insider's existence probability $\pi_1 + \pi_0$, Proposition 25 shows that firms with different attributes may experience different liquidity changes after a regulation. Furthermore, a firm's attributes may also correlate with the exchange on which it gets listed, and thus the liquidity of different exchanges may also change in opposite directions. For instance, value firms usually prefer to be listed on NYSE, while growth firms prefer to be listed on NASDAQ. This might explain why liquidity in NYSE and NASDAQ changed in different directions after Regulation Fair Disclosure. Moreover, a regulation always benefits one type of insider. To see this, suppose that the stock price increases after the regulation. This reflects that the market maker correctly anticipates that the type-0 insider is less likely to exist. Then the type-0 insider enjoys a strategic advantage and a higher payoff if she does exist. This mechanism differs from the chilling effect, which hypothesizes that regulations enhance the benefit from inside information by

exogenously making insiders' more reluctant to disclose it.

4.5 Competition Among Market Makers

So far, we have assumed that the market maker faces no competition, and can set the asset price at any level.[17] While many market makers in reality do face competitions, this assumption is innocuous when the insider definitely exists. Indeed, in the special case of $\pi_u = 0$, Anderson and Smith (2013) show that monopolistic and competitive market making yield the same equilibrium outcomes.[18] Does this conclusion still hold if there is uncertainty about the existence of the insider? We show in this section that competitive market making results in market breakdown whenever $\pi_u > 0$. This provides an alternative justification for the designation of monopoly power to market makers in reality.[19]

Specifically, we consider the standard setting in the literature, in which the market maker faces perfect competition and has to set the asset price at the asset's expected value. The logic requires two steps. We first discuss competitive market making with only one type of insider and show that no well-behaved equilibrium exists. Then, since the solution to the one-type setup essentially serves as the boundary condition for the two-type setup, we conclude that no well-behaved equilibrium exists under the competitive-market-making assumption.

[17]For example, think of a NYSE designated market maker (DMM) of a specific stock. Since a DMM enjoys higher rebates and faces lower transaction fees than other market makers of the same stock, it is reasonable to model him/her as a monopolist.

[18]We discuss its intuition in Section 4.5.2.

[19]See footnote 7 for empirical evidence.

4.5.1 One-Dimensional State Space Extension

We first assume that when the insider exists, the inside information can only be $\upsilon = 1$. The state space under this assumption corresponds to the vertical axis in Figure 4.1. Under this assumption, the state space is one dimensional and we can use π_1 as the state variable. To simplify notation, we drop the subscript and use V to represent the type-1 insider's value function. The equilibrium definition is the same except that the market maker, facing perfect competition, has to set the asset price to its expected value,

$$p(\pi_1) = \pi_1 + \hat{p} \cdot (1 - \pi_1) \ .$$

It is straightforward to see that the insider earns zero profit when the market maker definitely knows that she exists, since the price will remain at $p(1) = 1$ forever. On the other hand, when the market maker does not believe that the insider definitely exist, the price will remain $p(0) = \hat{p}$ forever, and the insider, if she exists, would enjoy infinite payoff. Thus, the two boundary conditions for the insider's value function are $V(0) = \infty$ and $V(1) = 0$. Proposition 26 states that there is no Markov equilibrium compatible with competitive market making.

Proposition 26. *No Markov equilibrium exists with boundary conditions* $V(0) = \infty$, $V(1) = 0$.

The intuition can be seen from a comparison between this setup and that with a monopolistic market maker. If the market maker can set any price, his dominant strategy is to set the asset price to 1, regardless of the insider's trading rate. Under

this pricing scheme, the insider gains zero from trading. On the other hand, if the market maker faces perfect competition, he is no longer able to set the price to 1 and is thus unable to protect himself from losing money to the potentially existing insider. In turn, the insider, if she exists, obtains a strategic advantage in addition to the informational advantage, so that she can achieve an infinite expected payoff. As a result, such an equilibrium does not exist.

A similar non-existence result holds for the symmetrical one-dimensional case in which the inside information can only be $\upsilon = 0$ when the insider exists. Hence, we also obtain the boundary condition on the horizontal axis in Figure 4.1 for the two-dimensional problem.

4.5.2 Two-Dimensional State Space Extension

We now consider the extension that when the insider exists, the inside information can be either $\upsilon = 0$ or $\upsilon = 1$. The solution to the one-dimensional setup in competitive market making yields the boundary conditions for the two-dimensional setup. According to Proposition 26, the type-1 insider can achieve an infinite expected payoff on the vertical and horizontal boundaries of the state space Π except for the point $(\pi_0, \pi_1) = (0, 1)$, where her payoff is zero (because the market maker would set the asset price to 1). This results in a discontinuity of the expected payoff for the type-1 insider on the boundary of Π, and thus precludes the existence of a continuous value function for the type-1 insider over the interior of Π (denoted by Π°). A similar argument also precludes the existence of a continuous value function for the type-0 insider over Π°. Therefore, these boundary conditions do not admit a "regular"

equilibrium.

Proposition 27. *There is no Markov equilibrium in which the insider's value functions are continuous in* Π°.

Intuitively, note first that while the insider's trading rate θ is unbounded, by definition, she should not be able to make infinite profit in equilibrium. Proposition 23 shows that her profit is bounded only if the market maker prices the asset at its expected value conditional on the existence of the insider, $\frac{\pi_1}{\pi_0+\pi_1}$. However, if the market maker faces perfect competition, he has to price the asset at its unconditional expected value, $\pi_1 \cdot 1 + \pi_0 \cdot 0 + (1 - \pi_1 - \pi_0) \cdot \hat{p}$.

When the insider definitely exists; i.e., when $\pi_1 + \pi_0 = 1$, these two pricing strategies coincide. This is due to the facts that 1) the Markovian properties of equilibria imply that the insider's HJB equations (4.9) and (4.10) are essentially invariant in the two setups, that 2) the inadmissibility of infinite profits for insiders imposes the same condition between equilibrium prices and the insider's trading strategies, and that 3) Back and Baruch (2004) establish that the competitive price π_1 and the insider's corresponding trading rate given by Proposition 23 constitute an equilibrium.

To understand facts 1) and 2), observe first that given the values of p and λ_v, whether they stem from a monopolistic or competitive market maker, equations (4.9) and (4.10) are invariant. Due to her risk neutrality, the insider's payoffs depend linearly on her trading rate θ_v. While the insider's trading rate is unbounded, in equilibrium, she should not be able to make infinite profit. It follows that the terms involving θ_v in (4.9) and (4.10), which are invariant to the monopoly power of the

market maker, must sum up to zero. This interprets fact 2). Note that the price p only shows up in the respective first terms in (4.9) and (4.10) corresponding to the insider's instantaneous payoffs. This is because, in a Markov equilibrium, the price affects the evolution of state variables only indirectly through the insider's trading rate θ_υ, which also pins down λ_υ. Therefore, the insider's HJB equations are only left with terms that are identical in the two setups and are independent of prices. This interprets fact 1).

When there is uncertainty about the existence of the insider; i.e., when $\pi_1 + \pi_0 < 1$, facts 1) and 2) still hold. However, the pricing strategy of a competitive market maker, $\pi_1 \cdot 1 + \pi_0 \cdot 0 + (1 - \pi_1 - \pi_0) \cdot \hat{p}$, no longer coincides with that of a monopolistic market maker, $\frac{\pi_1}{\pi_0 + \pi_1}$, that bounds the insider's profit. Since the market maker can no longer prevent the insider from making infinite profit, an equilibrium fails to exist, indicating market breakdown.

4.6 Conclusion

This paper contributes to the market microstructure literature by considering the uncertainty about existence of the insider in an otherwise standard continuous-time insider trading model. We characterize the essentially unique equilibrium, which has two key features. First, the market maker never updates the existence probability of the insider over time. Second, the ratio of existence probabilities between type-1 and type-0 insiders governs the equilibrium strategies and payoffs. We then discuss two applications of this model. The first application is the impact of stock market

regulations on price and liquidity. A key insight is that liquidity in this model is governed by the relative likelihood of two types of insider, rather than by the absolute level of the insider's existence probability. Thus, although a regulation always reduces the latter, it may have opposite effects on the liquidity of different firms, depending on the change in the former. In the second application, we show that, with the presence of uncertainty about the existence of the insider, competitive market-making leads to market breakdown. This offers an alternative rationale for the monopoly power of market makers in reality.

Chapter 5

Dynamic Contracting with Flexible Monitoring[1]

5.1 Introduction

Bureaucratic systems are ubiquitous, forming the skeleton of governments, large business firms and NGOs. A key issue for an efficient bureaucratic system is how to incentivize its officers to satisfactorily perform their job duties. In practice, both monitoring and compensation schemes play significant roles in providing such incentives. The existing literature focuses its analysis mainly on the latter, while keeping the former exogenously fixed.[2] This paper incorporates flexibility in designing the monitoring scheme into an otherwise standard dynamic contracting framework. It shows that the joint design of the two schemes engenders non-trivial interaction that results in novel implications in incentive provision.

[1]This chapter is based off of a paper of the same name co-authored with Dai Liang and Ming Yang (Dai et al. (2020a)).

[2]That is, it either assumes a single exogenous performance indicator, or focuses on how much monitoring capacity should be devoted to a given performance indicator.

Specifically, we allow the principal ("she"[3], designer of the schemes) to have the flexilibity in allocating her limited monitoring capacity to seek different types of evidence about whether the agent ("he", a representative officer) is taking her desired action. In reality, the monitoring capacity can be understood as the designated budget for hiring a quality control team, installing call recorders or surveillance cameras, etc. Subject to the monitoring capacity constraint, the principal can seek confirmatory and/or contradictory evidence of the agent's effort. The more capacity allocated to confirmatory evidence (contradictory evidence), the more likely such evidence arrives if the agent indeed works (shirks) and hence the more effective is the reward (punishment) in incentivizing the agent, and this in turn makes it more worthwhile to seek such evidence in the first place. Such interaction between the monitoring and compensation schemes calls for their joint design.

To fix ideas, consider a continuous-time setup, in which the principal has a project that requires the agent's operation. The agent is less patient than the principal and can work or shirk at each instant. From the perspectives of both the principal and social welfare, it is optimal for the agent to work, but the agent enjoys a private benefit from shirking. To incentivize the agent, at each instant, the principal chooses a combination of "carrot-based search" ("C-search" hereafter) and "stick-based search" ("S-search" hereafter). That is, she can allocate her fixed amount of monitoring capacity to seek two types of evidence, and can determine how much to reward or punish the agent when the evidence arrives. C-evidence confirms the agent's effort since it emerges only if the agent has worked, while S-evidence refutes the agent's

[3]We do not intentionally associate the players with particular genders.

effort since it emerges only if the agent has shirked. The principal can also terminate the project at any time, which is socially inefficient.

Our setup accommodates the discussion of two issues novel in the literature. First, we identify a key tradeoff between C-search and S-search as a means to incentivize the agent. This determines the principal's optimal allocation of her limited monitoring capacity as a function of the agent's continuation value. On one hand, C-search generates greater variation than S-search in the agent's continuation value, and is thus less advantageous to the principal, who is effectively risk averse in the relevant range of the agent's continuation value. This is because, given that the agent indeed works, no S-evidence exists, and thus no adjustment to the agent's continuation value is required; while C-evidence does emerge in equilibrium, which necessarily involves a reward upon its receipt ("carrots" hereafter) and the downward adjustment of the agent's continuation value in the absence of C-evidence. On the other hand, for S-search alone to be an effective incentive, a sufficiently high continuation value is required as the agent's stake in the project, whereas the effectiveness of C-search does not depend on the agent's continuation value. Moreover, even if S-search can work alone, a high continuation value for the agent must be maintained, which involves high interest expenditure for the principal, making S-search less advantageous than C-search. This tradeoff between C-search and S-search, together with the incentive versus interest tradeoff, shapes the optimal incentive scheme.

Consequently, when the agent's continuation value is low, the principal allocates all her monitoring capacity to C-search. Instead of paying the agent immediately upon receiving C-evidence, the principal adds the whole reward to the agent's con-

tinuation value to build a buffer against inefficient termination and to make S-search effective in the future. In addition, since the arrival rate of C-evidence is set to its maximum (as C-search attracts all the monitoring capacity), carrots should be just enough to deter shirking.

When the agent's continuation value has reached a level sufficient for S-search to be effective, but is not enough for S-search alone to deter shirking, the optimal incentive scheme features a "phase change." That is, instead of the carrot-only mode, the principal now relies mainly on S-search, and sets the penalty for observing S-evidence ("sticks" hereafter) to its maximum: confiscation of the whole stake promised to the agent, resulting in termination of the project. C-search is still used to make up for the S-search, but carrots are larger to minimize reliance on C-search. Carrots decrease as the agent's continuation value increases. Moreover, the standard incentive versus interest tradeoff determines a payout boundary. Payments to the agent are incurred only when the agent's continuation value grows beyond that boundary.

The second issue novel to the literature concerns the option of perpetuating the agent's effort. Specifically, the flexibility of combining C-search with S-search offers the principal the option of first building up the agent's stake in the game (i.e., his continuation value) with C-search, and then perpetuating his effort mainly with S-search, which avoids inefficient termination. But is this option optimal? We show that the answer is yes if and only if the latent benefit from the agent's effort and the principal's flexibility in allocating monitoring capacity are both sufficiently large. Moreover, when perpetuation of the agent's effort is optimal, the value function is convex in the vicinity of the (absorbing) payout boundary when public randomiza-

tion is not allowed. This is due to a new economic force in addition to the standard incentive versus interest tradeoff. That is, the higher the agent's continuation value, not only is it the less likely to reach the (inefficient) termination boundary as in existing models, but it is also more likely to reach the absorbing payout boundary, where the project becomes completely immune to inefficient termination. The latter fact makes the marginal benefit of accumulating the agent's continuation value increasing instead of decreasing in the continuation value in the vicinity of the payout boundary.

Our model also yields plausible predictions. If we interpret the continuation value as seniority in the bureaucratic system, then junior officers are incentivized mainly based on evidence that confirms their contribution, while senior officers are incentivized mainly based on evidence that refutes their contribution. Second, concerning the compensation scheme, the reward for each piece of evidence that confirms a contribution varies little among junior officers, but decreases with seniority for senior officers, and features an upward jump when a junior officer becomes senior. The penalty for each piece of evidence that contradicts a contribution increases with seniority for both junior and senior officers. Third, except for those hired permanently, all officers are more prone to unemployment in the absence of evidence that confirms their contribution, and more so if they are less senior. Lastly, permanent positions are offered if and only if both the flexibility in adjusting monitoring schemes and the potential synergy created by officers are sufficiently large.

5.1.1 Literature Review

Our work is related mainly to the continuous-time dynamic contracting literature, pioneered by DeMarzo and Sannikov (2006), Biais et al. (2007) and Sannikov (2008). Both DeMarzo and Sannikov (2006) and Biais et al. (2007) study continuous-time variants of the discrete-time dynamic security design model in DeMarzo and Fishman (2007). DeMarzo and Sannikov (2006) directly apply the martingale representation technique developed in Sannikov (2008) in a continuous-time setup, while Biais et al. (2007) is based on the continuous-time limit of a discrete-time model. Early work on dynamic moral hazard models also includes Biais et al. (2010). Like our model, Biais et al. (2010) use a Poisson process instead of Brownian motions to model discrete losses in continuous time, whose arrival rate depends only on the agent's hidden action. Myerson (2015) considers a similar problem under a political economics framework where a political leader uses randomized punishment to motivate governors to work. In contrast to the discrete losses in Biais et al. (2010), Sun and Tian (2017) use Poisson processes to model arrivals of discrete revenue. Similarly, He (2012) considers a risk-averse agent who can save privately and whose hidden effort affects the arrival rate of discrete revenue. In those models, monitoring is exogenous. In other words, the output processes, which are functions of hidden actions and other random factors, are exogenously assumed, and play dual roles as both direct determinants of physical payoff and bases for monitoring and contracting. The essence of our model is to separate these two roles in order to study the interaction between the design of monitoring schemes and that of contracts.

Recent work also endogenizes the monitoring scheme in dynamic moral hazard

models. On top of the framework of DeMarzo and Sannikov (2006), Piskorski and Westerfield (2016) allow the principal to monitor the agent at a cost that increases with her monitoring intensity. Based on a framework similar to that of Biais et al. (2010), Chen et al. (2020) consider the timing decision of monitoring, where monitoring is modeled as paying a fixed cost for a credible guarantee of the agent taking the desired action. Varas et al. (2020) consider a problem where monitoring serves as an incentive device and also provides information to the principal. In Orlov (2018), the principal can change her monitoring intensity. While these papers explore how much monitoring capacity should be devoted to a given monitoring technology and its optimal timing, our focus is on the principal's optimal allocation of limited capacity to different aspects of monitoring, as the basis for both her monitoring activities and the design of her incentive scheme.

In terms of results, the payout boundary in our model is absorbing when it is optimal for the principal to perpetuate the agent's effort, while the payout boundaries in the aforementioned models are reflective. One exception is that in Sannikov (2008), where the agent is risk averse and the mechanism is completely different. There, while the contractual relation is perpetuated, the agent no longer exerts effort; i.e., he retires once his continuation value reaches the payout boundary. The high continuation value there implies too low a marginal utility of consumption, making it too expensive to incentivize the agent.

In an essentially static setup, Li and Yang (2019) and Georgiadis and Szentes (2020) also study the impact of the principal's flexibility on the design of her monitoring scheme. Based instead on a dynamic setup, we are able to explore when it is

optimal for the principal to perpetuate the agent's effort. In addition, our notion of flexibility is different from that in **?**. Their information source is a single exogenous (conditional on the agent's effort) linear diffusion process, and the flexibility that they consider refers to the principal's freedom to stop observing that process earlier if existing observations are sufficient to prove the agent's deviation from the desired action. Instead, the notion of flexibility in our paper refers to the principal's freedom to allocate different levels of monitoring capacity to various processes (interpreted as different performance indicators) contingent on the whole history summarized by the agent's continuation value.

Our work is also related to the literature on problems of dynamic attention allocation. Smolin (2017) studies a problem where the principal designs an evaluation policy for both principal and agent to learn about the agent's type. Instead, we consider a moral hazard problem, in which the agent takes hidden actions and his compensation scheme is endogenous. Nikandrova and Pancs (2018) analyze a dynamic problem in which an investor decides how to allocate her limited attention between seeking confirmatory evidence of the profitability of one project and seeking that of another. Also in a dynamic setting, Che and Mierendorff (2019) study an individual's decision among immediate action, confirmatory learning (i.e., seeking evidence to confirm the state she finds relatively more likely) and contradictory learning (i.e., seeking evidence to confirm the state she finds relatively less likely), before taking actions that affect his state-contingent payoff. Mayskaya (2020) generalizes Che and Mierendorff (2019) by considering a decision problem where the agent needs to pick one of two alternatives and can split his attention to learn information

about either. Also similar to Che and Mierendorff (2019), Kuvalekar and Ravi (2019) consider how a principal should incentivize an agent, who is to allocate limited attention between seeking evidence that confirms and evidence that refutes a project's quality. While the monitoring capacity allocation between C-search and S-search in our model is similar to the learning problems in these papers, the problem we study is fundamentally different. While in their models, the fact to be learned is exogenous , our model features strategic interaction with moral hazard, in which the fact to be learned (i.e., whether the agent is shirking) is endogenous to the choice of monitoring technologies, and in turn, to the principal's design of an incentive scheme.

5.2 The Model

5.2.1 Setup

Time is continuous and infinite. There is a principal ("she", designer of a bureaucratic system) and an agent ("he", a representative officer in the system). Both are risk neutral. The principal has a discount rate $r > 0$ and unlimited access to capital. The agent has a discount rate $\rho > r$ and is protected by limited liability; i.e., his cumulative payment from the principal must be non-negative and non-decreasing over time. The principal owns a project that requires the agent's operation, which involves an action $a_t \in [0, 1]$ taken by the agent. The action can be understood as the level of shirking. If action a_t is taken at instant t, in period $[t, t + dt]$, the agent enjoys a private benefit of $\lambda \cdot a_t dt$, while the principal's benefit is $z \cdot (1 - a_t)\, dt > 0$.

The principal can terminate the project at any time, and the project then generates a payoff of zero for both players.

Here, we interpret z as the latent progress of a project or the reputation of an entity that is lost without the agent's due diligence and is not discernible immediately.[4] Therefore, contracts cannot be made contingent on whether z is accrued. We interpret z this way for two reasons. First, it captures the reality, mentioned in the Introduction, that the agent's hidden actions are often not reflected in existing indicators, such as current output, sales or stock prices. This is because, the outcome of such actions may be realized only in the long run. For example, the daily practice of officers in charge of disease prevention can hardly be evaluated until an epidemic arrives. A manager focusing on the long-term development of his firm should not be over-responsive to the firm's current sales, output or stock prices. Second, it separates the role of output as a component of physical payoff from that as a given performance indicator; the latter having been well studied. This allows us to focus on the principal's active monitoring of the agent's action. For ease of presentation, we hereafter refer to z as the "synergy" (between principal and agent).

To model the principal's capacity-allocation decision, we assume that at each instant the principal can choose how to allocate her μ units of monitoring capacity between "carrot-based search" ("C-search") and "stick-based search" ("S-search"); i.e., to seek one of two types of evidence as the basis for reward and penalty. The receipt of C-evidence confirms the agent's effort, while the receipt of S-evidence

[4]Alternatively, one can interpret the principal and agent (instead with discount rates 0 and $\rho - r$, respectively) as playing a repeated game that ends exogenously with arrival rate r, when the principal receives her whole payoff from the game. Accordingly, z can be understood as the instantaneous contribution to that payoff.

contradicts it. Specifically, if the principal allocates a fraction $\alpha_t \in [0, \bar{\alpha}]$ of her μ units of monitoring capacity to seeking S-evidence and the remaining $1 - \alpha_t$ to C-evidence, she receives S-evidence at the arrival rate $\mu \cdot \alpha_t \cdot a_t$, and C-evidence at the arrival rate $\mu \cdot (1 - \alpha_t) \cdot (1 - a_t)$. Hence, the agent's chance of being caught shirking is proportional to a_t, the level of shirking, and $\mu \cdot \alpha_t$, the capacity allocated to monitoring shirking. Intuitively, if the agent does not shirk, no evidence of shirking exists, and the principal cannot find S-evidence no matter how much capacity is allocated to seeking it; if the principal allocates no capacity to monitor shirking, she receives no S-evidence regardless of the agent's level of shirking. The arrival rate of C-evidence can be interpreted similarly. More specifically, the cumulative number of arrivals of S-evidence, Y_1, and that of C-evidence, Y_0, satisfy

$$dY_{1,t} = \begin{cases} 1, & \text{with probability } \mu \alpha_t a_t dt \\ 0, & \text{otherwise} \end{cases},$$

and

$$dY_{0,t} = \begin{cases} 1, & \text{with probability } \mu \left(1 - \alpha_t\right) \left(1 - a_t\right) dt \\ 0, & \text{otherwise} \end{cases},$$

respectively. To save the notation, we write $Y = (Y_0, Y_1)$.

It is worth noting that upper bound $\bar{\alpha}$ measures the flexibility of the principal in allocating her capacity across C-search and S-search. By definition, as a fraction, $\bar{\alpha} \leq 1$. To highlight the role of this flexibility, we assume $\bar{\alpha}$ to be close to 1. Formally,

Assumption 1. $\bar{\alpha} \geq 1 - \frac{\rho}{\mu}$.

In addition, we assume that $r < \rho < \mu$; i.e., the principal is more patient than

the agent,[5] and that the principal has enough capacity to monitor the agent. As standard in the dynamic contracting literature, we assume that $z > \lambda > 0$; i.e., z is large enough so that shirking (action 1) is inefficient even taking into account the agent's private benefit. This assumption ensures that it is optimal for the principal to always implement $a_t = 0$,[6] and allows us to focus on the interaction between the monitoring scheme and the agent's compensation scheme.

A contract X specifies the recommended action a taken by the agent, the monitoring scheme α,[7] the cumulative payment I to the agent and the time of termination τ as functions of the history of past evidence. As mentioned before, without loss of generality, we focus on contracts that implement $a_t = 0$ for all t, so that we suppress a and write $X = (\alpha, I, \tau)$.

[5] As in DeMarzo and Sannikov (2006), this assumption is made for two reasons. First, it captures the fact that a bureaucratic system usually has a greater risk-bearing capacity than an individual officer. Second, it rules out the possibility that the principal indefinitely postpones payments to the agent, which is neither interesting nor realistic.

[6] Formally established in Online Appendix. This result differs from a setup featuring Brownian motions (e.g., Proposition 8 in DeMarzo and Sannikov (2006)), where it is optimal to induce the agent's effort only if the surplus generated is significantly greater than the agent's private benefit from shirking. Implementation of effort requires the agent to be exposed to the adverse effect of the quadratic variation of his continuation value. Such exposure has a first-order impact if his continuation value follows a Brownian motion. In our setup, the agent's continuation value follows a (generalized) Poisson process, and such exposure has no first-order impact. This is evident from the fact that V'' does not enter our HJB equation (5.9).

[7] The capacity μ in our model should be understood generically as resources available to the principal for monitoring the agent. In reality, this corresponds to the total budget for hiring a quality control team, installing call recorders or surveillance cameras, etc. By including the monitoring scheme α (i.e., the allocation of capacity) in the contract, we are studying the benchmark in which evaluation of the agent's performance changes focus as the contractual relationship develops, and this is explicitly stated at the outset and strictly implemented. This benchmark is realistic, especially for firms, organizations or bureaucratic systems that specify the details of their routine monitoring of employees in different positions with different seniority in contracts, charters or codes of conduct. Situations where the principal cannot commit to a monitoring scheme are also realistic in other circumstances, but are beyond the scope of this paper.

Given the contract X and an action process a, the expected discounted utility of the agent is

$$\mathbb{E}^a\left[\int_0^\tau e^{-\rho t}\left(dI_t+\lambda a_t dt\right)\right],$$

and that of the principal is

$$\mathbb{E}^a\left[\int_0^\tau e^{-rt}\left(z\left(1-a_t\right)dt-dI_t\right)\right]. \tag{5.1}$$

For notational convenience, we hereafter suppress all time subscripts when no confusion can be caused.

While contracts involving public randomization are of theoretical interest, they are typically not practical. Therefore, we postpone the discussion of public randomization to Section 5.5, and consider only deterministic contracts for the rest of this paper unless otherwise mentioned.

5.2.2 Incentive Compatibility and Limited Liability

To characterize the incentive compatibility condition, we employ martingale techniques similar to those introduced by Sannikov (2008). When choosing his action at time t, the agent considers how it will affect his continuation value, defined as

$$w_t\left(X,a\right)=\mathbb{E}^a\left[\int_t^\tau e^{-\rho u}\left(dI_u+\lambda a_u du\right)\Big|\,\mathcal{F}_t\right]1_{\{t<\tau\}},$$

where $\{\mathcal{F}_t\}$ is the filtration generated by Y. Martingale representation theorem yields the following lemma.

Lemma 5. *For any contract X that implements $a_t = 0$ for all $t \leq \tau$, there exist predictable processes (β_0, β_1) such that w_t evolves before termination $(t \leq \tau)$ as*

$$dw_t = \rho w_t dt - dI_t + \beta_{0,t} \left[dY_{0,t} - \mu \left(1 - \alpha_t \right) dt \right] - \beta_{1,t} dY_{1,t} \ . \tag{5.2}$$

The contract is incentive compatible if and only if

$$\mu \alpha_t \beta_{1,t} + \mu (1 - \alpha_t) \beta_{0,t} \geq \lambda \ . \tag{IC}$$

And the contract satisfies the limited liability constraint of the agent if and only if

$$\beta_{1,t} \leq w_t \tag{5.3}$$

and

$$\beta_{0,t} + w_t \geq 0 \ . \tag{5.4}$$

Proofs of this lemma and of all the other lemmas and propositions are relegated to the Appendix unless otherwise specified. Intuitively, β_0 refers to the agent's reward upon receipt of C-evidence, and β_1 refers to his punishment upon receipt of S-evidence. Hereafter, we refer to β_0 as "carrots," and β_1 as "sticks." (IC) highlights our model's key feature. Its left-hand side consists of the instruments, C-search and S-search, that the principal uses to incentivize the agent, which together with the associated carrots and sticks, must sum to at least λ, the agent's private benefit from shirking. The principal can choose not only the allocation of her monitoring capacity α, but also β_0 and β_1, the carrots and sticks.

Two limited liability constraints in Lemma 5 restrict the magnitudes of reward and punishment. (5.3) requires that sticks should be no more than the whole stake promised to the agent. (5.4) says that carrots plus the stake already promised to the agent must be non-negative, which will be shown slack.

5.3 Basic Properties of the Optimal Contract

This section provides a heuristic derivation of some basic properties of the optimal contract. Theorem 7 at the end of this section verifies that this contract is indeed optimal.

Let $B(w)$ denote the principal's value function. We have the Hamilton–Jacobi–Bellman (HJB) equation in the continuation region ($t < \tau$)

$$rB\left(w\right) = \max_{\alpha,\beta_0,\beta_1} z + \left(1-\alpha\right)\mu\left[B\left(w+\beta_0\right) - B\left(w\right)\right] + \left[\rho w - \beta_0\mu\left(1-\alpha\right)\right]B^{'}\left(w\right) \ , \tag{5.5}$$

subject to

$$\mu\alpha\beta_1 + \mu(1-\alpha)\beta_0 \geq \lambda \ ; \tag{IC}$$

$$\beta_1 \leq w \ ; \tag{5.6}$$

$$\beta_0 + w \geq 0 \ ; \tag{5.7}$$

and

$$\alpha \in [0, \bar{\alpha}] \ . \tag{5.8}$$

The left-hand side of (5.5) is the principal's expected flow of value. The first term on the right-hand side, z, is the flow of synergy. The second term is due to the carrots β_0 given to the agent if C-evidence is obtained, which happens with probability $(1-\alpha)\mu dt$ conditional on $a = 0$ being implemented from t to $t + dt$. The third term arises from the drift of w, where ρw is the rate at which interest accrues, and $-\beta_0\mu(1-\alpha)$ is the flip side of carrots due to promise keeping: if there is no C-evidence, the principal reduces the agent's continuation value at this rate to balance against carrots, so that the continuation value w_t net of a drift $\rho w_t dt$ is a martingale, and thus the contract does deliver w_t in expectation to the agent.

Note that no term in (5.5) corresponds to sticks (i.e., no term containing β_1), because S-evidence is never obtained if the agent follows the contract and takes $a = 0$ at each instant. In this sense, sticks serve only as an off-equilibrium threat. Therefore, the limited liability constraint (5.6) must be binding: If S-evidence were obtained, the principal would maximize the penalty by terminating the project and confiscating the agent's whole stake w.

Notationally, superscript * hereafter denotes items in the optimal contract.

Property 1. $\beta_1^*(w) = w$.

Instead of $B(w)$, it is equivalent but more convenient to continue our analysis based on $V(w) = B(w) + w$, the sum of the principal's value function and the agent's continuation value, or their joint surplus. (5.5) then becomes

$$rV(w) = \max_{\alpha,\beta_0} z + [\rho w - \beta_0\mu(1-\alpha)]V'(w) + (1-\alpha)\mu[V(w+\beta_0) - V(w)] - (\rho - r)w \ . \tag{5.9}$$

Next, since $r < \rho$, we guess and later verify that there is a payout boundary $\bar{w}$ as standard in existing dynamic contracting models, e.g., DeMarzo and Sannikov (2006) and Biais et al. (2010). If $w > \bar{w}$, the principal will simply pay $dI = w - \bar{w}$ immediately and reduce the continuation value to $\bar{w}$. Otherwise, the principal will use backloading; i.e., wait for the agent's continuation value w to increase instead of paying him immediately (i.e., $dI = 0$). By construction, $V(\bar{w} + \beta_0) = V(\bar{w})$, so that when $w = \bar{w}$, the third term on the right-hand side of (5.9) equals zero, and $V'(\bar{w}) = 0$ if it exists. If $V'(\bar{w})$ does not exist; i.e., the left and right derivatives at $\bar{w}$ are not equal, (5.9) is not defined at $w = \bar{w}$, which means that the coefficient in front of $V'(w)$ is zero at $\bar{w}$. Notice that this coefficient is the drift of the continuation value. Hence, when $V'(\bar{w})$ does not exist, $\bar{w}$ is an absorbing payout boundary. As a result, regardless of whether the payout boundary $\bar{w}$ is absorbing or not, the second term in (5.9) must also equal zero when $w = \bar{w}$, so that

$$V(\bar{w}) = \frac{z}{r} - (\rho - r)\frac{\bar{w}}{r} \tag{5.10}$$

and

$$B(\bar{w}) = \frac{z}{r} - \frac{\rho}{r}\bar{w} \ . \tag{5.11}$$

Moreover, we must have $\bar{w} \leq \frac{\lambda}{\rho+\mu\bar{\alpha}}$. If not, then at any continuation value $w \in \left(\frac{\lambda}{\rho+\mu\bar{\alpha}}, \bar{w}\right)$, the principal could always incentivize the agent with the following contract: paying out $w - \frac{\lambda}{\rho+\mu\bar{\alpha}}$ immediately to reduce the agent's continuation value to $\frac{\lambda}{\rho+\mu\bar{\alpha}}$; setting $\alpha = \bar{\alpha}$, $\beta_1 = \frac{\lambda}{\rho+\mu\bar{\alpha}}$ and $\beta_0 = \frac{\rho\lambda}{\mu(1-\bar{\alpha})(\rho+\mu\bar{\alpha})}$, so that (IC) is binding, and that $\beta_0\mu(1-\bar{\alpha}) = \rho\frac{\lambda}{\rho+\mu\bar{\alpha}}$; i.e., the drift of the agent's continuation value is

zero, and thus $w = \frac{\lambda}{\rho+\mu\bar{\alpha}}$ is an absorbing state.[8] The principal's payoff from this new contract is

$$\frac{z}{r} - \frac{\rho}{r} \cdot \frac{\lambda}{\rho+\mu\bar{\alpha}} - (\bar{w} - \frac{\lambda}{\rho+\mu\bar{\alpha}}) > \frac{z}{r} - \frac{\rho}{r}\bar{w} = B(\bar{w}) \ ,$$

where the inequality follows $\bar{w} > \frac{\lambda}{\rho+\mu\bar{\alpha}}$, contradicting the optimality of $B(\bar{w})$. As a standard result in this literature, the optimality of B implies $B'(w) > -1$ for $w < \bar{w}$ and $B'(w) = -1$ for $w > \bar{w}$. Then by definition, $V'(w) > 0$ for $w < \bar{w}$ and $V'(w) = 0$ for $w > \bar{w}$. We summarize these results in the following property.

Property 2. *There exists a* $\bar{w} \in (0, \frac{\lambda}{\rho+\mu\bar{\alpha}}]$ *such that i)* $dI^* = (w - \bar{w})^+$ *; ii)* V *is increasing in* $[0, \bar{w}]$*; iii) if* $w \geq \bar{w}$*,*

$$V(w) = z/r - (\rho - r)\bar{w}/r \ ; \tag{5.12}$$

and iv) either $V'(\bar{w}) = 0$*, or* $\rho\bar{w} - \mu(1 - \alpha^*(\bar{w}))\beta_0^*(\bar{w}) = 0$*, i.e., the drift at* $w = \bar{w}$ *is* 0*.*

Together with Assumption 1 and Property 1, we have $\beta_1^*(w) = w < \bar{w} \leq \frac{\lambda}{\rho+\mu\bar{\alpha}} < \lambda/\mu$ for $w < \bar{w}$. Hence, by (IC), sticks alone are not sufficient to incentivize the agent to work. Moreover, (IC) and Property 1 imply that $w\alpha^* + \beta_0^*(1 - \alpha^*) \geq \lambda/\mu$, thus $\beta_0^*(w) \geq \lambda/\mu \geq \frac{\lambda}{\rho+\mu\bar{\alpha}} \geq \bar{w}$ for $w < \bar{w}$. This, together with Property 2, implies

Property 3. $w + \beta_0^* \geq \bar{w}$ *for all* $w < \bar{w}$*.*

That is, a single piece of C-evidence suffices to make the continuation value w jump to the payout region $[\bar{w}, +\infty)$, so that $V(w+\beta_0^*) = V(\bar{w})$; i.e., β_0^*, carrots, raises

[8] More precisely, under this contract, once $w = \frac{\lambda}{\rho+\mu\bar{\alpha}}$, the continuation value never drifts away and the agent receives discrete payments of $\beta_0 = \frac{\rho\lambda}{\mu(1-\bar{\alpha})(\rho+\mu\bar{\alpha})}$ at the arrival rate $\mu(1-\bar{\alpha})$ forever.

their joint surplus only from $V(w)$ to $V(\bar{w})$, and the remaining reward, $\beta_0^* - (\bar{w} - w)$, is an immediate transfer from the principal to the agent and has no impact on their joint surplus. Also, the limited liability constraint (5.7) slacks as conjectured.

Property 3 simplifies our derivation of the optimal contract, given that the value function V may not always be concave.[9] To see this, note that according to Property 3, (5.9) becomes

$$rV(w) = \max_{\beta_0, \alpha} z + [\rho w - \beta_0 \mu(1-\alpha)] V'(w) + (1-\alpha)\mu[V(\bar{w}) - V(w)] - (\rho - r)w \ , \tag{5.13}$$

whose right-hand side is always decreasing in β_0. This has two important implications. First, it indicates the advantage of using S-search rather than C-search, regardless of the concavity of V. In equilibrium, S-evidence is never obtained, and thus S-search incentivizes the agent without causing variation in his continuation value w. But if C-search is used (i.e., $\alpha < 1$), C-evidence is obtained in equilibrium and generates variation in w. Property 3 implies that effectively, the upward jump in w upon the receipt of C-evidence is always $\bar{w} - w$ (after the bonus payment), which is independent of α and β_0. But the magnitude of the downward drift of w in the absence of C-evidence, $\beta_0\mu(1-\alpha)$, is increasing in both the capacity allocated to C-search, $\mu(1-\alpha)$, and the associated carrots, β_0. Therefore, the more the principal resorts to C-search, the more adverse variation in w is generated, making it detrimental relative to sticks.

Second, the fact that the right-hand side of (5.13) is decreasing in β_0 implies a

[9]This possibility is discussed in Subsection 5.4.2. We also concavify the value function via public randomization in Section 5.5.

binding (IC) in the no-payment region $[0, \bar{w}]$, i.e.;

Property 4. $\mu\left[\alpha^* w + (1-\alpha^*)\beta_0^*\right] = \lambda$.

The incentive compatibility constraint (IC) plays a central role in our model. Property 4 establishes that the combination of C-search and S-search should be just enough to overcome the agent's private benefit from shirking.

Note that the principal still has two degrees of freedom to adjust the sensitivities of the agent's continuation value to evidence reflecting his actions. As mentioned in the literature review, this contrasts with the counterpart in models without choice among multiple performance indicators; e.g., in Sannikov (2008) and Biais et al. (2010), where there is no such degree of freedom.

Now we are ready to derive the central piece of the model — the optimal allocation of monitoring capacity, α, and the optimal carrots, β_0, in the no-payment region $[0, \bar{w}]$. Given Properties 1 and 4, we obtain

$$\beta_0^* = \frac{\lambda - \mu\alpha^* w}{\mu\left(1-\alpha^*\right)}. \tag{5.14}$$

(5.14) highlights the substitution between the capacity allocated to C-search, $1-\alpha$, and the associated carrots, β_0, which is peculiar to our setup with flexibility in monitoring design. The more capacity allocated to C-search, the higher is the probability of obtaining C-evidence that confirms the agent's effort, and thus less reward is needed to incentivize the agent. Conversely, higher carrots provide a stronger incentive for the agent, and thus reduce the principal's reliance on obtaining C-evidence, enabling her to use S-search.

Note that while the control is well-behaved as long as $\alpha^* < 1$, we have $\beta_0^* \to \infty$ as $\alpha^* \to 1$. In that limit, while the bonus upon arrival of C-evidence, β_0^*, approaches infinity, the arrival rate of such evidence, $\mu\,(1-\alpha^*)$, approaches zero. But by (5.14), the expected bonus payment, $\mu\,(1-\alpha^*)\,\beta_0^*$, approaches $\lambda - \mu w$, which is well-defined. Thus, this singularity is inconsequential for our qualitative results regarding the shape of the value function and the payout boundary. To avoid distractions to our main economic insight, we assume that $\bar{\alpha} < 1$ in addition to Assumption 1 to preclude this singularity. For interested readers, we can provide the analysis for the case of $\bar{\alpha} = 1$ upon request.

In the no-payment region, we have $dI = 0$ by definition and $V(w + \beta_0) = V(\bar{w})$ from Property 2. Plugging (5.14) into (5.13), the HJB equation (5.9) becomes

$$rV(w) = \max_{\alpha \in [0,\bar{\alpha}]} z - (\rho - r)w + (1-\alpha)\mu[V(\bar{w}) - V(w)] + (\rho w - \lambda + \mu\alpha w)V'(w). \quad (5.15)$$

Notice that α affects the right-hand side of (5.15) through the last two terms. As explained before, the third term reflects its impact through carrots; i.e., raising α reduces the arrival rate of C-evidence and that of the contingent increment $V(\bar{w}) - V(w)$ in their joint surplus. This in turn reduces the expected instantaneous joint surplus $(1-\alpha)\mu[V(\bar{w}) - V(w)]$. The impact is linear in α, and the marginal impact is $-\mu[V(\bar{w}) - V(w)]$, whose absolute value decreases monotonically with w.

The last term on the right-hand side of (5.15) reflects the impact of α through the flip side of carrots; i.e., a lower arrival rate of C-evidence also reduces the downward

drift of the agent's continuation value w due to promise keeping.[10] This increases the expected instantaneous joint surplus $(\rho w - \lambda + \mu\alpha w)V'(w)$. This effect is also linear in α, with a marginal impact $\mu w V'(w)$, which could be non-monotonic in w. Since the total impact of α is linear, with marginal impact

$$\mu\left[wV'(w) + V(w) - V(\bar{w})\right], \tag{5.16}$$

we obtain the following property.

Property 5. *If* $wV'(w) + V(w) < V(\bar{w})$, *then* $\alpha^* = 0$ *and* $\beta_0^* = \lambda/\mu$;

If $wV'(w) + V(w) = V(\bar{w})$, *then* $\alpha^* \in [0, \bar{\alpha}]$ *and* $\beta_0^* = \frac{\lambda - \mu\alpha^* w}{\mu(1-\alpha^*)}$;

If $wV'(w) + V(w) > V(\bar{w})$, *then* $\alpha^* = \bar{\alpha}$ *and* $\beta_0^* = \frac{\lambda - \mu\bar{\alpha}w}{\mu(1-\bar{\alpha})}$.

The following theorem verifies that our derived contract is indeed optimal.

Theorem 7. *The solution* V *to HJB equation* (5.9) *is principal and agent's joint surplus under the optimal contract. Moreover, the optimal contract is characterized by Property 5.*

5.4 The Role of Flexible Monitoring

This section highlights the critical role of flexible monitoring, which is central to this paper. Section 5.4.1 shows that such flexibility is indeed utilized by and thus valuable to the principal. Section 5.4.2 articulates that such flexibility allows a

[10]Note that $\rho w - \lambda + \mu\alpha w \leq 0$ since $\bar{w} \leq \frac{\lambda}{\rho + \mu\bar{\alpha}}$. Raising α thus reduces the magnitude of the downward drift.

long-term contractual relationship that perpetuates the agent's effort with positive probability when the synergy, z, is sufficiently large, and that the value function is convex in the vicinity of the payout boundary $\bar{w}$ if and only if such perpetuation is optimal. Section 5.4.3 summarizes these results with a graphic illustration using the narrative of career path and provides a few empirically plausible predictions.

5.4.1 Flexibility in Monitoring is Utilized

Property 5 establishes that other than in knife-edge cases, the optimal monitoring capacity allocated to S-search, α^*, is either 0 or $\bar{\alpha}$.[11] This subsection further establishes that an optimal contract necessarily involves both possibilities. Specifically, Proposition 28 establishes that $\alpha^*(w) = 0$ when the agent's continuation value w is close to 0, and $\alpha^*(w) = \bar{\alpha}$ when w is close to the payout boundary $\bar{w}$. This indicates that flexibility in allocating monitoring capacity between C-search and S-search allows the principal to incentivize the agent differently at different stages of his career, and is thus valuable to the principal.

Proposition 28. *There exists a $\hat{w}_0 \in (0, \bar{w})$ and a $\hat{w}_{\bar{\alpha}} \in [\hat{w}_0, \bar{w})$, such that $\alpha^*(w) = 0$ and $\beta_0^*(w) = \lambda/\mu$ for $w \in (0, \hat{w}_0)$, and that $\alpha^*(w) = \bar{\alpha}$ and $\beta_0^*(w) = \frac{\lambda - \mu\bar{\alpha}w}{\mu(1-\bar{\alpha})}$ for $w \in (\hat{w}_{\bar{\alpha}}, \bar{w}]$.*

From Property 5, the optimal contract involves only $\alpha = 0$ and $\alpha = \bar{\alpha}$ except for the knife-edge case featuring indifference. From (5.15) we know that for each

[11] Lemma 13 in the Appendix shows that no interval of continuation values exists, such that the principal is indifferent between 0 and $\bar{\alpha}$. Thus the knife-edge cases are non-generic.

$w \in (0, \bar{w})$, either $\alpha = 0$ and

$$rV(w) = z + [\rho w - \lambda]V^{'}(w) + \mu[V(\bar{w}) - V(w)] - (\rho - r)w \ , \tag{5.17}$$

or $\alpha = \bar{\alpha}$ and

$$rV(w) = z + (1 - \bar{\alpha})\mu[V(\bar{w}) - V(w)] + [\rho w - \lambda + \mu\bar{\alpha}w]V^{'}(w) - (\rho - r)w \ . \tag{5.18}$$

Both equations can be solved in closed form (see Appendix). It can be verified that $V'(0)$ is finite. This implies $0 \cdot V^{'}(0) + V(0) = 0 < V(\bar{w})$, and by continuity, there is a neighborhood of $w = 0$ such that $wV^{'}(w) + V(w) < V(\bar{w})$. Thus, by Property 5, the principal relies completely on C-search when the agent's continuation value w is low. The statement for $(\hat{w}_{\bar{\alpha}}, \bar{w})$ can be similarly proved with closed-form solutions.

Intuitively, when the agent's continuation value w is low, the principal should not rely on S-search, because the agent has little to lose even if he is known to have shirked. Relying on C-search also maximizes the chance of obtaining C-evidence. This helps the principal quickly raise the agent's "skin in the game," which makes S-search (which is costless to the principal) more effective in the future, and pushes the project away from termination (which is socially inefficient). When the agent's continuation value w is higher, the principal can impose a large penalty for S-evidence. Since such a penalty is just an off-equilibrium threat, making S-search less costly than C-search, the principal should rely on S-search as much as possible.

The flexibility of combining C-search and S-search allows the principal to exploit their respective advantages. On one hand, C-search generates greater variation than

S-search in the agent's continuation value, and is thus less advantageous to the principal. Given that the agent does work, no S-evidence would arrive, and thus, no adjustment of the agent's continuation value would be required. However, in equilibrium, C-evidence would be obtained, which would necessarily involve a reward and the downward adjustment of the agent's continuation value in the absence of C-evidence. On the other hand, a sufficiently high continuation value is required as the agent's skin in the game for sticks alone to be an effective incentive, whereas the effectiveness of C-search does not depend on the agent's continuation value. Moreover, even if S-search could work alone, a high continuation value for the agent has to be maintained, which involves interest expenditure for the principal, making S-search less advantageous than C-search. This tradeoff between C-search and S-search induces the principal to rely only on C-search when w is low, and on S-search, as much as possible, when w is high.

Concerning carrots, β_0, recall that the right-hand side of (5.13) is decreasing in β_0, since an increase in β_0 makes the drift of the agent's continuation value, $\rho w - \beta_0 \mu(1-\alpha)$, more negative due to promise keeping, and thus makes the project more prone to termination. Hence, given the optimal capacity allocation α^*, β_0^* should be set as low as possible — such that (IC) is binding. Thus, for agents facing $\alpha^* = 0$, including those with $w \in (0, \hat{w}_0)$, we have $\beta_0^*(w) = \lambda/\mu$, and the resulting drift of w is $\rho w - \lambda < 0$. For agents facing $\alpha^* = \bar{\alpha}$, including those with $w \in (\hat{w}_{\bar{\alpha}}, \bar{w}]$, we have $\beta_0^*(w) = \frac{\lambda - \mu\bar{\alpha}w}{\mu(1-\bar{\alpha})}$, and the resulting drift of w is $\rho w - \lambda + \mu\bar{\alpha}w \leq 0$.[12]

Note first that $\beta_0^*(w)$ is constant in the region of $\alpha^*(w) = 0$, but is decreasing

[12]This is because $w \leq \bar{w} \leq \frac{\lambda}{\rho + \mu\bar{\alpha}}$.

in the region of $\alpha^* = \bar{\alpha}$. This is because in the latter case, sticks increase with w, partially substituting carrots that are required by (IC). Second, $\beta_0^*(w)$ features an upward jump when α^* switches from 0 to $\bar{\alpha}$. To see this, notice the fact that any switching point $w < \frac{\lambda}{\rho+\mu\bar{\alpha}} \leq \frac{\lambda}{\mu}$ implies that the size of the jump is $\frac{\lambda-\mu\bar{\alpha}w}{\mu(1-\bar{\alpha})} - \frac{\lambda}{\mu} > \frac{\lambda-\mu\bar{\alpha}\cdot\frac{\lambda}{\mu}}{\mu(1-\bar{\alpha})} - \frac{\lambda}{\mu} = 0$. Third, the drift of w increases (i.e., becomes less negative) with w, due to the interest accrued (i.e., due to the term ρw) and the increasing reliance on S-search in lieu of C-search (i.e., due to the term $\mu\bar{\alpha}w$). Lastly, the drift of w is negative, which moves w towards 0, the termination boundary, unless w reaches the payout boundary $\bar{w}$ and $\bar{w} = \frac{\lambda}{\rho+\mu\bar{\alpha}}$, where the drift is zero; i.e., the project and the agent's effort are perpetuated. Section 5.4.2 characterizes when such perpetuation is optimal.

5.4.2 Possibility of Perpetuating the Agent's Effort

This subsection discusses whether the optimal contract involves the perpetuation of the agent's effort with positive probability. Mathematically, this refers to whether the payout boundary $\bar{w}$ is an absorbing state. We show that this is related to the (local) convexity of the value function, which is in turn determined by the flexibility of the principal's capacity allocation as captured by $\bar{\alpha}$, and by the magnitude of the synergy z to that of the agent's private benefit from shirking, λ. Specifically, 1) $\bar{w}$ is absorbing if and only if $\bar{w} = \frac{\lambda}{\rho+\mu\bar{\alpha}}$; 2) $\bar{w}$ is absorbing if and only if the value function V is not universally concave[13]. More precisely, $\bar{w}$ is absorbing if and only if

[13]Recall from Section 5.2 that we discuss public randomization in Section 5.5 and preclude it in the rest of the paper unless otherwise mentioned.

V is convex in $(\hat{w}_{\bar{\alpha}}, \bar{w})$ given by Proposition 28; and 3) $\bar{w}$ is absorbing if and only if $\bar{\alpha} > \frac{r-\rho+\mu}{2\mu}$ and z is sufficiently large.

Again, the role of flexibility in capacity allocation is worth highlighting. We show that without such flexibility, perpetuation of the agent's effort is not optimal.

Recall from Property 2 that $\bar{w} \leq \frac{\lambda}{\rho+\mu\bar{\alpha}}$. We have in addition the following lemma.

Lemma 6. *$\bar{w}$ is absorbing if and only if $\bar{w} = \frac{\lambda}{\rho+\mu\bar{\alpha}}$.*

Proof. First consider the "if" statement. If $\bar{w} = \frac{\lambda}{\rho+\mu\bar{\alpha}}$, we show that the following strategy is feasible and optimal, and makes $\bar{w}$ absorbing: $\alpha = \bar{\alpha}$, $\beta_0 = \frac{\rho\lambda}{\mu(1-\bar{\alpha})(\rho+\mu\bar{\alpha})}$ and $\beta_1 = \bar{w} = \frac{\lambda}{\rho+\mu\bar{\alpha}}$. Feasibility results from the binding (IC) constraint. To see why $\bar{w}$ is absorbing, note that when $w = \bar{w}$, the positive component of the drift of the agent's continuation value due to accrued interest is $\rho\bar{w}dt = \frac{\rho\lambda}{\rho+\mu\bar{\alpha}}dt$, and the negative component as the flip side of carrots is $\mu(1-\bar{\alpha})\beta_0 dt$, which also equals $\frac{\rho\lambda}{\rho+\mu\bar{\alpha}}dt$, so that w remains constant when there is no C-evidence, and when it is obtained, the whole reward β_0 is paid out immediately so that w remains at $\frac{\lambda}{\rho+\mu\bar{\alpha}}$.

To see the optimality of this strategy, observe that the principal's expected payoff at $w = \frac{\lambda}{\rho+\mu\bar{\alpha}}$ is $\mathbb{E}(\int_0^{+\infty} ze^{-rt}dt - \beta_0 \int_0^{+\infty} e^{-rt}dY_{0,t})$. Since $Y_{0,t} - \mu(1-\bar{\alpha})t$ is a martingale,

$$\mathbb{E}(\int_0^{+\infty} ze^{-rt}dt - \beta_0 \int_0^{+\infty} e^{-rt}dY_{0,t}) = \frac{z}{r} - \frac{\beta_0\mu(1-\bar{\alpha})}{r} = \frac{z}{r} - \frac{\rho}{r}\cdot\frac{\lambda}{\rho+\mu\bar{\alpha}} \ .$$

Thus, the expected joint surplus is

$$\frac{z}{r} - \frac{\rho}{r}\cdot\frac{\lambda}{\rho+\mu\bar{\alpha}} + \bar{w} = \frac{z}{r} - \frac{\rho-r}{r}\cdot\bar{w} \ .$$

From (5.10), this strategy achieves the optimal joint surplus at the payout boundary.

Now consider the "only if" statement. From Property 2, it suffices to show that any $\bar{w} < \frac{\lambda}{\rho+\mu\bar{\alpha}}$ cannot be absorbing. Any contract respecting (IC) satisfies

$$\beta_0\mu(1-\bar{\alpha}) \geq \lambda - \bar{w}\mu\bar{\alpha} > \rho \cdot \frac{\lambda}{\rho+\mu\bar{\alpha}} > \rho\bar{w} \ .$$

Thus, when there is no C-evidence, the agent's continuation value always has a downward drift term $\rho\bar{w} - \beta_0\mu(1-\bar{\alpha}) < 0$. This implies that the payout boundary $\bar{w} < \frac{\lambda}{\rho+\mu\bar{\alpha}}$ is reflective. □

Notice the role of flexibility in capacity allocation here. If $\bar{w} = \frac{\lambda}{\rho+\mu\bar{\alpha}}$, the way for the principal to perpetuate the agent's effort is to set $\alpha = \bar{\alpha}$; i.e., to rely on S-search as much as possible. But if $w_0 < \hat{w}_0$, such a monitoring scheme is not viable since the agent has too little to lose if caught shirking. To avoid inefficient termination, the principal must first rely on C-search to build up the agent's skin in the game, and then switch to stick-dominant mode when the continuation value is high enough. This approach is impossible without flexibility in capacity allocation.

Proposition 29 further establishes the connection between the possibility of perpetuating the agent's effort, the (local) convexity of the value function, and the conditions on exogenous parameters.

Proposition 29. *Let $\hat{w}_0$ and $\hat{w}_{\bar{\alpha}}$ be given by Proposition 28. Then, $\bar{w} = \frac{\lambda}{\rho+\mu\bar{\alpha}}$ (i.e., absorbing) if and only if V is convex in $(\hat{w}_{\bar{\alpha}}, \bar{w})$, which holds if and only if $\bar{\alpha} > \frac{r-\rho+\mu}{2\mu}$ and z is sufficiently large. Moreover, when $\bar{w}$ is absorbing, $\hat{w}_0 = \hat{w}_{\bar{\alpha}}$.*

That is, the agent's effort can be perpetuated if and only if both the principal enjoys sufficient flexibility in capacity allocation and the synergy of the contractual relationship is large enough. In other words, no matter how large the synergy, the contractual relationship would terminate in probability one if the principal does not have sufficient flexibility of capacity allocation. This again stresses the importance of such flexibility in shaping the optimal contract.

Proposition 29 also establishes the equivalence relation between $\bar{w}$ being absorbing and the local convexity of the value function V. Moreover, if the agent's effort could be perpetuated, there is only one switching point $\hat{w}_{\bar{\alpha}} \in (0, \bar{w})$ for capacity allocation; i.e., the optimal capacity allocation is $\alpha^* = 0$ in $(0, \hat{w}_{\bar{\alpha}})$, and $\alpha^* = \bar{\alpha}$ in $(\hat{w}_{\bar{\alpha}}, \bar{w})$.

We explain the role of flexibility in capacity allocation (i.e., in combining C-search with S-search) in three steps. First, we argue that it is possible for the optimal contract to have an absorbing payout boundary while satisfying (IC) for all w only if the principal is able to combine C-search with S-search (i.e., to have $\alpha \in (0, 1)$). To see this, first consider the extreme situation where only C-search is viable (i.e., α is fixed to 0) as in Sun and Tian (2017). From Property 4, $\beta_0^*(w) = \lambda/\mu$ for all $w \leq \bar{w}$, and the value function V satisfies (5.17). It is straightforward from Sun and Tian (2017) that V must be concave, so that $\bar{w}$ is reflective. This reflects the standard incentive versus interest tradeoff in the literature. That is, an increase in w pushes the continuation value away from the termination boundary 0, whose marginal benefit decreases with w, but whose marginal cost, due to an increase in accrued interest, is constant. Such a tradeoff is also featured in Biais et al. (2010) and DeMarzo and Sannikov (2006). In this case, the agent receives a lumpy bonus of

$\beta_0 - (\bar{w} - w)$ and a jump of $\bar{w} - w$ in his continuation value upon the receipt of each piece of C-evidence, but his continuation value will still drift downward from $\bar{w}$ until the next piece of C-evidence is obtained, and will eventually reach zero in probability one, resulting in termination of the project. In the opposite extreme situation, where only S-search is viable (i.e., α is fixed to 1), if $w < \lambda/\mu$, it is impossible to satisfy (IC), and thus no contract implements the desired action; otherwise, it is optimal to choose an absorbing payout boundary $\bar{w} = \lambda/\mu$ by making a flow payment $\rho\bar{w}$ forever. Therefore, it is the feasibility of combining C-search with S-search that allows the combination of the incentive compatibility of effort for all w from C-search and the absorbing payout boundary from S-search in the optimal contract.

Second, we argue that flexibility in capacity allocation allows the principal to optimally decide whether to make the payout boundary $\bar{w}$ absorbing. Doing so perpetuates the agent's effort when his continuation value w is high. But it also leads to the accrual of higher interest from maintaining the continuation value at $\bar{w}$ and a lower arrival rate of C-evidence when w is low. To see this, suppose α is fixed to $\bar{\alpha}$, so that the value function V satisfies (5.18) in $(0, \bar{w}]$. Its closed-form solution implies that V is convex and thus that the payout boundary $\bar{w}$ is absorbing if and only if $\bar{\alpha} > \frac{r-\rho+\mu}{2\mu}$ and z is sufficiently large. In other words, although the option of an absorbing payout boundary is readily available even if α is fixed to $\bar{\alpha}$, such an option is optimal for the principal if and only if she can rely on S-search and the synergy is large enough. This is because, while the absorbing payout boundary makes it possible to permanently avoid inefficient termination, it also requires maintaining a high constant continuation value of $\bar{w}$ as potential sticks, resulting in a high constant

flow of interest and a reduction of surplus due to the difference in discount rates between principal and agent. Therefore, it is advisable to make the payout boundary absorbing if and only if the continuation value $\bar{w}$ can be sufficiently utilized (i.e., $\bar{\alpha}$ is large) and perpetuating the contractual relation is sufficiently beneficial (i.e., z is large). Moreover, once the principal has flexibility in adjusting α, she has another option to avoid termination, which is to set $\alpha = 0$ to maximize the arrival rate of C-evidence and upward jumps of w for low values of w (see Proposition 28). This makes an absorbing payout boundary even less attractive. Therefore, for values of z not high enough, although it is optimal to make $\bar{w}$ absorbing if α is fixed to $\bar{\alpha}$, it is no longer optimal if α can be flexibly adjusted in $[0, \bar{\alpha}]$.

Lastly, notice the novel feature of our model that the value function V is convex in the vicinity of the payout boundary $\bar{w}$ when $\bar{w}$ is absorbing. While the concavity of V in $(0, \hat{w}_{\bar{\alpha}})$ still reflects the standard incentive versus interest tradeoff, new economic forces come into play in $(\hat{w}_{\bar{\alpha}}, \bar{w})$, where $\alpha^* = \bar{\alpha}$. There, the reliance on S-search reduces the downward drift of the continuation value, $\mu(1-\alpha)\beta_0$, that balances carrots. This raises the marginal benefit of increasing w without affecting the marginal cost, and thus makes V less concave. Moreover, a fundamental change occurs when $\bar{w}$ becomes absorbing. In that case, the marginal benefit of increasing w results not only from the fact that w is further away from the inefficient absorbing state 0, but also from the fact that w is closer to the efficient absorbing state $\bar{w}$. The latter fact, together with the constant marginal cost due to accrued interest, makes the marginal benefit increasing instead of decreasing in w and thus the value function V convex in $(\hat{w}_{\bar{\alpha}}, \bar{w})$.

5.4.3 A Career Path Narrative

Using the narrative of career path, this subsection employs a graphic illustration to summarize the role of flexibility in monitoring practice, which is the core of this paper, and provides some empirically plausible predictions.

Proposition 28 establishes that the optimal monitoring and compensation schemes for junior officers (i.e., agents with continuation value $w \in (0, \hat{w}_0)$) differ qualitatively from those for senior officers (agents with continuation value $w \in (\hat{w}_{\bar{\alpha}}, \bar{w})$). Concerning monitoring schemes, incentives for junior officers are in carrot-only mode (i.e., $\alpha = 0$), since they need to accumulate a cushion against unemployment (i.e., termination) and have little to lose even if caught shirking. Senior officers are instead incentivized in stick-dominant mode (i.e., $\alpha = \bar{\alpha}$), since they have enough skin in the game, and sticks are off-equilibrium penalties, which are less costly than on-equilibrium carrots. Thus, our model predicts that

Prediction 1. *Incentives for junior officers are based mainly on confirmatory evidence of their contribution, while incentives for senior officers are based mainly on contradictory evidence of their contribution, but are compensated by higher rewards upon the receipt of confirmatory evidence.*

Concerning compensation schemes, as an off-equilibrium threat, sticks are always the whole continuation value w and thus increase with seniority. For junior officers, carrots are set to the minimum level required to induce effort, which is constantly λ/μ, while carrots for senior officers, $\frac{\lambda - \mu\bar{\alpha}w}{\mu(1-\bar{\alpha})}$, decrease with seniority, since they are replaced by sticks at higher continuation values. The promised stakes of super-senior

officers; i.e., agents with continuation value $w > \bar{w}$, are so large that a payment $w - \bar{w}$ to reduce accrued interest is urgent enough to dominate their incentive problems. Hence,

Prediction 2. *The reward for each piece of evidence confirming a contribution varies little among junior officers, but decreases with seniority for senior officers. The penalty for each piece of evidence contradicting a contribution increases with seniority for both junior and senior officers.*

Concerning the possibility of being fired (i.e., termination), in the absence of C-evidence, the drift of the continuation value of junior officers is $\rho w - \lambda$, and that of senior officers is $\rho w - \lambda + \mu\bar{\alpha}w$. Both are negative unless $w = \bar{w} = \frac{\lambda}{\rho+\mu\bar{\alpha}}$; i.e., unless the (senior) officer is made permanent. They become less negative as the continuation value w increases, for two reasons. First, larger stakes in the game carry a larger interest income. Second, larger stakes also allow for larger sticks and less reliance on C-search, and thus less downward drift in the continuation value to balance in-equilibrium carrots. Therefore,

Prediction 3. *Except for those hired permanently, in the absence of evidence confirming their contribution, an officer becomes more prone to unemployment; more so if the officer is less senior.*

When is it possible to hire officers permanently? Proposition 29 shows that it is the case if and only if both the flexibility in adjusting monitoring schemes and the potential synergy created by officers are sufficiently large.

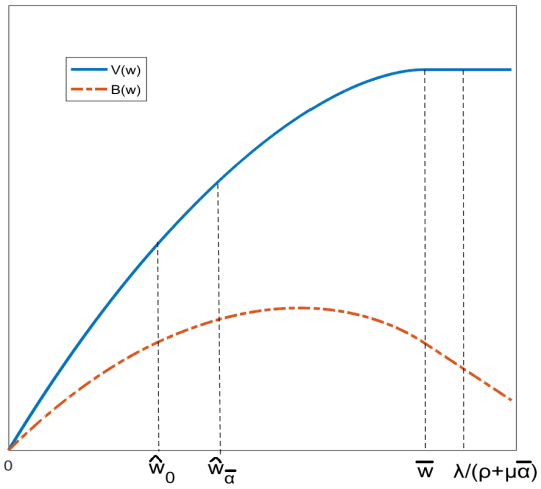


Figure 5.1: Reflective Payout Boundary $\bar{w}$

First, consider the case $\bar{\alpha} \leq \frac{r-\rho+\mu}{2\mu}$ illustrated in Figure 5.1; i.e., flexibility in monitoring is not large enough. The solid blue line corresponds to the value function V (in terms of the joint surplus), and the dash-dotted red line corresponds to the principal's value function $B(w) = V(w) - w$. The value function V is strictly concave in $(0, \hat{w}_0)$, reflecting the standard incentive versus interest tradeoff. V is also strictly concave in $(\hat{w}_{\bar{\alpha}}, \bar{w})$, where the fact that carrots decrease with the agent's stake in the game makes V less concave. However, since $\bar{\alpha}$ is low, the principal does not have enough flexibility to rely on S-search to the extent that she wants, so that $\bar{w}$ is reflective and is thus determined by $V'(\bar{w}) = 0$. That is, senior officers who receive carrots still face a downward drift in their promised stakes and thus the risk of being fired. This is the case no matter how large the synergy z is.

Now fix an $\bar{\alpha} > \frac{r-\rho+\mu}{2\mu}$, so that the principal does have enough flexibility in adjusting the monitoring scheme. If z is small, the synergy is too low to justify

perpetuation of the agent's effort, so the optimal contract is qualitatively similar to that of the case $\bar{\alpha} \leq \frac{r-\rho+\mu}{2\mu}$. Once z is large enough, the optimal contract changes fundamentally, as shown in Figure 5.2. — Now the principal has both the flexibility and the desire to perpetuate the agent's effort, so now the payout boundary $\bar{w}$ becomes absorbing. That is, the agent is "tenured" once his effort is confirmed by the receipt of C-evidence. In addition, the possibility of completely avoiding termination creates a new marginal benefit of increasing continuation value, and thus makes the value function V convex in $(\hat{w}_{\bar{\alpha}}, \bar{w})$. Moreover, $\hat{w}_0 = \hat{w}_{\bar{\alpha}}$, so that α, the capacity allocated to S-search, only switches once as the agent's continuation value rises from 0 to $\bar{w}$. Thus, we have

Prediction 4. *Permanent positions are offered if and only if both the flexibility in adjusting monitoring schemes and the potential synergy created by officers are sufficiently large.*

5.5 Public Randomization

So far, we have been focusing on deterministic contracts, on the basis that random contracts are of little practical relevance in reality. This is also theoretically without loss of generality if the resulting value function is globally concave as in the case illustrated in Figure 5.1 and as in most models in the literature. However, as established in Proposition 29, our value function is convex in the vicinity of the payout boundary $\bar{w}$ if it is absorbing (Figure 5.2). For this situation, this section discusses

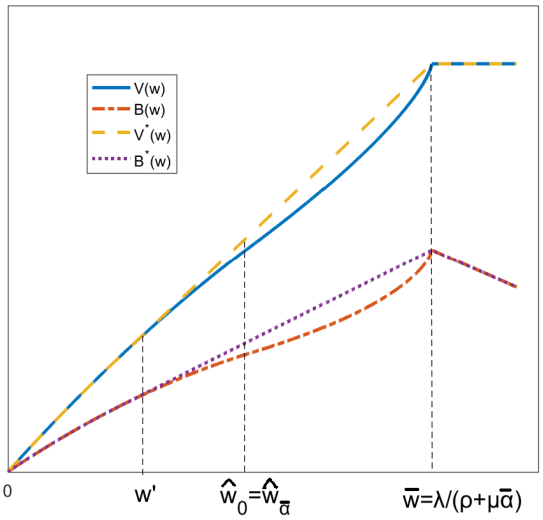


Figure 5.2: Absorbing Payout Boundary $\bar{w}$

the extension in which public randomization of the following form is allowed. At time 0, in addition to starting the contractual relationship with a deterministic continuation value w_0, the principal can choose a mean-preserving spread of w_0 as the basis for random contracts, but no further randomization is allowed for $t > 0$. Since $B = V - w$, and the linear term has no effect on the concavification operation, we can work with the joint surplus function V without loss of generality.

Proposition 30. *With public randomization, the principal's value function is* $B^{rdm} = V^{rdm} - w$*, where* V^{rdm} *is the concavification of* V*.*

Proof. Proposition 29 establishes that when V is not globally concave, we must have $\bar{w} = \frac{\lambda}{\rho+\mu\bar{\alpha}}$, and $V(\bar{w})$ is uniquely determined by Property 2. In addition, V is concave in $(0, \hat{w})$ and convex in $(\hat{w}, \bar{w})$, where $\hat{w} \equiv \hat{w}_0 = \hat{w}_{\bar{\alpha}}$. Therefore, the concavification of V must be over $\bar{w}$ and some $w' \in (0, \hat{w})$ as shown with the yellow broken line in

Figure 5.2.[14]

We check that the values of non-randomized states are not changed. First, $V(\bar{w})$ does not change because $\bar{w} = \frac{\lambda}{\rho+\mu\bar{\alpha}}$ is absorbing and its value does not depend on the values of other states. For $w \in (0, w')$, notice that the continuation value may only drift downward or jump upward over $\bar{w}$. Since $V(\bar{w})$ remains the same and $V^{rdm} = V$ for $w \in (0, w')$, the values of these states satisfy the same HJB equation and thus remain the same. □

5.6 Conclusion

This paper studies the joint design of monitoring and compensation schemes for officers in bureaucratic systems in a continuous-time moral hazard model. In the model, the principal (i.e., the designer of the scheme) can flexibly combine C-search with S-search to incentivize the agent (i.e., a representative officer). That is, the principal can flexibly allocate her limited monitoring capacity between confirmatory and contradictory evidence concerning the agent's effort as the basis for reward or punishment. We find that such flexibility generates rich dynamics, which differ qualitatively from the situation where only one of the two methods is feasible. When the agent has little skin in the game, the principal resorts only to C-search; when the agent has sufficient skin in the game, the principal instead assigns the highest possible weight to S-search. Moreover, only with such flexibility can the agent's effort be perpetuated with positive probability when the agent is less patient.

[14]The purple dotted line in Figure 5.2 illustrates the corresponding concavification B^* of the principal's value function B.

Chapter 6

Learning from Awareness[1]

6.1 Introduction

Awareness (unawareness) describes the situation where agents ignore relevant uncertainties or actions when making decisions. Dekel et al. (1998), the seminal paper on this topic in economics, establishes that the single state space model only allows for a degenerate awareness structure.[2] Since then, several multi-state space frameworks are built to allow for non-trivial awareness structures in an economics model. A natural follow-up of this research paradigm is to incorporate awareness into belief updating process to understand how an agent's awareness constrains his belief and how growing awareness affects belief updating.

The existing literature can be roughly classified into two categories based on their approaches addressing this problem. The first type of models, such as Karni

[1]This chapter is based off of my paper (Wang (2021b)).

[2]Earlier papers, such as Fagin and Halpern (1988) in computer science and Modica and Rustichini (1994), use the syntax approach to analyze awareness. The syntax modeling approach, standard in mathematical logic models, is seldom used in economic models.

and Vierø (2013) and Karni and Vierø (2017) consider the situation where an agent's unawareness leads him to mistakenly assign zero probability to possible events. An agent's awareness affects belief updating when an event assigned zero probability happens. Upon observing such an event, the agent will revise his belief by proportionally transferring probability masses from all events with nonzero probabilities to the observed event. Essentially, how awareness affects belief is embodied in the specification of the erroneous belief held by the agent.

The second type of models, such as Li (2009) and Heifetz et al. (2013), treats unawareness as an ignorance of relevant uncertainties. Loosely speaking, the objective state space is a vector space where each dimension represents a relevant aspect of the problem. Agents have subjective state spaces as subspaces of the objective state space due to unawareness.[3] An agent's belief is defined over his subjective state space and is correct in the sense that probabilities over events on his subjective state space are consistent with the marginal distribution of the distribution over the objective state space. Under this framework, growing awareness expands an agent's subjective state space and thus the agent's new belief will be defined on a space with higher dimension. On the other hand, it does not change the agent's marginal belief over his original subjective state space. In other words, growing awareness has no effects on belief updating if we only focus on events defined on the agent's original subjective state space.

We extend the subjective state space approach in this paper to allow for non-

[3]Strictly speaking, in some frameworks, the set of all state spaces form a complete lattice and a larger state space is "more expressive" in that it can describe a state in richer details. We only discuss product state spaces for the ease of illustration. Most insights from these frameworks can be obtained from a product state space structure.

trivial effects of awareness over belief updating. In this framework, an agent's subjective state space depends on the state realization (in the objective state space). The agent partially understands the mapping between state realizations and his subjective spaces and can form conjectures over his subjective state spaces in different state realizations. He then updates his belief by excluding states with conjectured subjective state spaces different from his subjective state space. If the agent is rational in the sense that he never excludes the true state, he can achieve a more precise probability estimation of the state realization.[4] We define this type of reasoning to be awareness inference since it captures how belief is changed by an agent's awareness.

A simple example related to the Dunning-Kruger effect can illustrate how awareness inference affects belief updating. The Dunning-Kruger effect, first documented in Kruger and Dunning (1999), describes the phenomenon that people who are incompetent often fail to recognize their incompetence. That is, poor performers tend to over-estimate their performances and the over-estimation is negatively correlated with the performance.[5] According to Dunning (2011), this effect is robust in many areas, including tests in logical reasoning and grammar skills, emotional intelligence, ability to discern funny jokes, etc. This example shows how awareness inference can rationalize this cognitive bias. Suppose students are asked to predict their performances in a test measured by percentile. A student's performance in the test is determined by his/her type $S_1 = \{L, M, H\}$. Type L represents 0% to 33% percentile; type M represents 33% to 66 % percentile; type H represents 66% to 100

[4]In this paper, the agent's rationality of not ruling out the true state is implied by axioms; or in other words, assumed.

[5]They attribute this effect to poor performers' ignorance bias, which is conceptually similar to awareness discussed in this paper.

% percentile. Two other aspects $S_2 = \{Y, N\}$ and $S_3 = \{Y, N\}$ are related to a student's type. For example, in a test of economics, S_2 may represent whether the student understands the concept of Nash equilibrium (Y) or not (N). S_3 may represent whether the student knows how to solve constrained maximization (Y) or not (N). A type who does not understand a certain piece of knowledge also fails to recognize how that piece of knowledge is related to the test. Type L understands neither of the two pieces of knowledge; type M only knows about either Nash equilibrium or constrained maximization; type H knows both pieces of knowledge.

In our framework, S_1, S_2 and S_3 are three relevant aspects of this problem and the objective state space is $S = S_1 \times S_2 \times S_3$, the Cartesian product of these aspects. Since a type H student is aware of all three aspects, his subjective state space is $S = S_1 \times S_2 \times S_3$. Other two types of students have smaller subjective state spaces. A type L student's subjective state space is S_1 while a type M student's subjective state space is either $S_1 \times S_2$ or $S_1 \times S_3$. If all students understand that their subjective state spaces are state-dependent, their reasoning process might be the following: A type L student's subjective state space provides him no additional information about his type. Thus, he takes the unconditional expectation and estimates his percentile to be 50%. A type M student has the following thought procedure: I understand that it is important to know Nash equilibrium (constrained maximization) to solve problems in this test. I would not be aware of this if I am a type L student or a type M student with no knowledge of Nash equilibrium (constrained maximization). Then he would takes average performance score for part of type M and type H to estimate his percentile to be higher than 66%. Following the same logic, a type

H student understands that since he knows both Nash equilibrium and constrained maximization, he must be a type H student and thus estimates himself to be at 83% percentile. Although all types of students have an equally weighted prior over three types, type M and type H students receive additional information from their state spaces and thus can achieve more accurate estimations of their performances. On the contrary, with no additional information, type L students make the least accurate estimations, which leads to the largest overestimation of their performances.

This example demonstrates two key elements for awareness to have non-trivial effects on belief updating. First, an agent's subjective state spaces need to be non-trivially state-dependent. In this example, a student's state space would provide no useful information in estimating performance if all types of students share the same subjective state space. Second, the agent needs to be sophisticated enough to understand this state dependency. If a type M student is naive in the sense that he thinks a type L student and a type M student share the same subjective state space, he cannot rule out the possibility that he is a type L student.

Some restrictions (over conjectures of subjective state spaces under different state realizations) are needed to achieve a reasonable prediction over how awareness affects belief updating. We adopt an axiomatic approach to impose these constraints. Axiom 2 requires that an agent cannot differentiate states that "look the same" in his subjective state space. Specifically, type L cannot differentiate the type M who is aware of aspect S_2 and the type M who is aware of aspect S_3 since both M types are mapped to the same subjective state for type L. Axiom 3 requires that an agent's predictions of his subjective state spaces in other states are subspaces of his current

subjective state space. In the previous example, a type M student cannot predict that a type H student's subjective state space is $S_1 \times S_2 \times S_3$ since he himself is not aware of the aspect S_3. Axiom 4 requires that, if an aspect is in an agent's current subjective state space and his subjective state space in another state realization, that agent must correctly conjecture that he is aware of that aspect in that state realization. This rules out the "correct conclusion from wrong inference" scenario. In such a scenario, a type M student may correctly predict that he is type M because he mistakenly thinks both type L and type H students' subjective state spaces are S_1.

Multiple conjectures (of subjective state spaces under different state realizations) are consistent with these axioms. By Axiom 3 and Axiom 4, the true state would never be excluded by the agent. In this sense, a conjecture is more accurate if it allows the agent to exclude more states. Under the least accurate conjecture (naive awareness structure), the agent thinks his subjective state space in any state is the same as his current subjective state space. Naturally, this leads to no belief updating. One the other hand, even with the most accurate conjecture, the belief updating induced by awareness cannot be perfect revealing under all state realizations. For instance, when a student is type M, his subjective state space is a subspace of the counterpart when he is type H. Thus, by Axiom 3 and Axiom 4, under any permissible conjecture, he must think that when he is type H, his subjective state space coincides his current subjective state space. In other words, no conjecture satisfying all three axioms can lead a type M student to conclude that he is not type H. We specify the most accurate conjecture (sophisticated awareness structure) and

show that the belief updating rule under this situation admits a simple form.

We further analyze how growing awareness interacts with belief updating. We consider this process as an agent receiving an awareness signal containing some aspects. This awareness signal expands the agent's subjective state spaces and change the agent's belief through two channels. First, if there is a Bayesian signal about a newly-learned aspect, then the awareness signal enables the agent to preform Bayesian updating. This channel, relying essentially on the existence of a Bayesian signal, exists in previous subjective state space frameworks. This paper indicates an additional channel of belief updating. With a larger subjective state space after receiving the awareness signal, the agent can form more detailed conjectures over agents' subjective state spaces and agents' conjectures in different state realizations. Specifically, he is able to better differentiate (objective) states when conjecturing. Thus, if the agent's conjectures are relatively accurate,[6] this enables him to exclude more states and form a more precise state estimation. This channel is novel and does not rely on the existence of a Bayesian signal. Moreover, states excluded from this channel not only depend on the signal itself but depend on the agent's subjective state space before receiving the awareness signal as well. Thus, the belief process induced by the awareness signal may not be a martingale. This is the major difference between belief updating from an awareness signal and belief updating from a Bayesian signal.

We apply this framework in a persuasion game example to examine the effectiveness of persuasion by sending awareness signals. A sender may either sending

[6]If the agent's conjecture is naive, growing awareness does not lead to belief updating.

awareness signals or performing a Bayesian experiment to change a receiver's belief. Since the belief updating induced by an awareness signal is state-dependent, the expectation of posteriors does not necessarily equal to the prior if the sender adopts the former persuasion approach. Thus, sending awareness signals can induce posteriors which cannot be induced by performing any Bayesian experiment. Under some belief structures, the former persuasion approach dominates the latter persuasion approach.

The subjective state space setting in this paper is similar to frameworks such as Li (2009), Heifetz et al. (2006) and Heifetz et al. (2013). Most papers with subjective state space aiming at defining an unawareness operator and generalizing the knowledge operator to fit the multiple state space setting. Instead, this paper's focus is on the interaction between awareness and belief updating. This paper is also closely related to Galanis (2015) and Galanis (2016). Galanis (2015) and Galanis (2016) make the observation that when an agent's awareness level is state-dependent, he may have "information process error" in Bayesian updating. Similar to these papers, in our framework agents have state-dependent subjective state spaces and may exclude states using awareness. Yet our paper's focus is different from these papers. Galanis (2015) and Galanis (2016) focus on defining and comparing value of information. That is, given the information structure induced by awareness, how agents make decisions differently from a standard framework after receiving Bayesian signals. Our paper adopts an axiomatic approach to analyze what are the possible information structures induced by agents' awareness structures. Moreover, the axiomatic approach used in this paper enables us to model the belief updating process

after an agent expands his subjective state space through learning new aspects.

Galperti (2019) also discusses how awareness plays a role in persuasion games. In Galperti (2019), similar to Karni and Vierø (2013), the receiver mistakenly assign zero probability to possible events. Once the receiver observes a probability zero event, he would adopt a new set of prior and reexamine the implications of all previous signals. Galperti (2019) focuses on the sender's trade off between taking advantage of the receiver's wrong belief and revealing the receiver's mistake to induce a drastic change in the receiver's belief. In this paper, persuasion is affected by awareness through a different channel. An agent never assigns zero probability to possible events. The sender takes advantage of the state dependency of awareness inference to induce a non-martingale belief process to achieve higher payoff.

This paper contributes to several strands of literatures. First, we introduce a framework to systematically describe the interaction between the awareness structure and growing awareness with belief updating. This framework allows for interactive higher order thinkings over awareness among agents. Moreover, this framework introduces a new form of signaling. In this framework, an expansion of subjective state spaces itself, even without any Bayesian signal, can lead to belief updating. This paper further enriches persuasion games by introducing the awareness persuasion channel. A sender may change the receiver's belief through sending awareness signals about relevant aspects. Since the belief process induced by awareness signals is not necessarily a martingale, in some cases, this enables the sender to achieve higher payoffs than performing a Bayesian experiment.

We proceed as follows. Section 6.2 describes the framework and awareness in-

ference principles. Section 6.3 discusses awareness signals and awareness inference upon receiving awareness signals. In section 6.4, we consider a persuasion game to illustrate the role of awareness signals in persuasion and the difference between awareness inference and Bayesian updating.

6.2 Baseline Framework

6.2.1 State Space and Awareness Structure

Let $I = \{1, .., N\}$ be the set of agents and $\mathcal{S} = \{S_j\}_{j=1}^d$ be the set of all relevant aspects. Each aspect S_j is a set contains finitely many possible realizations. For instance, if S_1 corresponds to weather conditions, we may have $S_1 = \{sunny, rainy\}$. The objective state space is a the Cartesian product of all aspects $S = \prod_{j=1}^d S_j$, which is a d-dimensional space. We endow a probability measure π on the objective state space.

Define a first order awareness operator $A^1 : I \times S \to 2^{\mathcal{S}}$. For each agent, this operator specifics the aspects he is aware of in each state.[7] An agent's subjective state space is the Cartesian product of the aspects he is aware of. Notice that aspects an agent is aware of in a certain state uniquely determine his subjective state space in that state, or vice versa. For example, if $A^1(i, s) = \{S_1, S_3, S_5\}$, agent i is aware of aspects S_1, S_3 and S_5 and his subjective state space is $S_1 \times S_3 \times S_5$ at state s.[8] An

[7]If a state s happens with probability zero, i.e., $\pi(s) = 0$, set $A^1(i, s) = \emptyset$ for all i. This assumption is only for the ease of interpretation, other specifications of $A^1(i, s)$ do not change the result qualitatively.

[8]The order of aspects is irrelevant. Thus we fix an increasing order of subscripts.

agent's initial belief is the marginal distribution of π over his subjective state space. In the standard state space model, every agent is aware of all aspects and thus his subjective state space in any state coincides the objective state space. In this case, π is the common prior.

In this framework, agents can form conjectures (and higher order conjectures) over awareness structures of themselves and other agents. We use awareness operators $A^n : (I \times S)^n \to 2^{\mathcal{S}}$, $n = 2, 3, ...$ to represent these conjectures. For example, for $i_1, i_2 \in I$ and $s_1, s_2 \in S$, $A^2((i_1, s_1), (i_2, s_2))$ specifies the conjecture formed by agent i_1 in state s_1 about the aspects agent i_2 is aware of, or equivalently, agent i_2's subjective state space in state s_2. Admittedly, s_2, as an element of the objective state space, may not be in i_1's subjective state space in state s_1. We resolve this issue by imposing an axiom to require that agent i_1 in state s_1 has the same conjecture over the aspects agent i_2 is aware of in all states sharing the same projection in i_1's subjective state space. A detailed discussion is deferred till next section.

Formally, for any integer $n > 1$, $A^n((i_1, s_1), ..., (i_n, s_n))$ contains the set of aspects that agent i_1 at state s_1 thinks that agent i_2 at state s_2 thinks that,... , thinks that agent i_n is aware of at state s_n. We call $\{A^n\}$, the set containing all operators A^n, the awareness structure. This set completely characterize all agents' subjective state spaces and their conjectures (and higher order conjectures) over all agents' subject subjective state spaces in all state realizations.

To simplify notation, let $A^n(i^n, s^n)$ to be an abbreviation of $A^n((i_1, s_1), ..., (i_n, s_n))$ where $i^n = (i_1, ..., i_n) \in I^n$ is a sequence of agents and $s^n = (s_1, ..., s_n) \in S^n$ is a

sequence of states.[9] Naturally, agent i_1's conjecture over agent i_2's conjecture over aspects that agent i_3 are aware of must be related to i_1's conjecture over the set of aspects that i_2 is aware of. Thus, sometimes we want to consider new sequences of agents and states formed by removing the last element of existing sequences. We use $i^n \setminus i_n$ to represent the sequence $(i_1, ..., i_{n-1})$ and $s^n \setminus s_n$ to represent the sequence $(s_1, ..., s_{n-1})$.

6.2.2 Inference from the Awareness Structure

In this section we discuss how an agent updates his initial belief (the marginal of π over his subjective state space) and forms conjectures over agents' beliefs in any state using $\{A^n\}$, the awareness structure. We refer to this process as awareness inference throughout this paper. First, we emphasize a consistency axiom that guarantees the validity of awareness inference. Discussions of other axioms over the awareness structure are postponed till the next section.

Axiom 1. *For any $s \in S, i \in I$, let $i^n = (\underbrace{i, ..., i}_{n}), s^n = (\underbrace{s, ..., s}_{n})$. Then $A^n(i^n, s^n) = A^1(i, s)$.*

To better understand this axiom, consider the situation where $n = 2$. For any agent i in any state s, his conjecture about aspects he is aware of in state s' is represented by $A^2((i, s), (i, s'))$. This axiom imposes a constraint over rationality that, when state s' is the current state s, agent i's conjecture $A^2((i, s), (i, s))$ is

[9]For the ease of notation, with no possibility of confusion, we sometimes may abuse the notation by defining i^n and s^n to be specific sequences satisfying certain conditions.

correct (in the sense that it coincides the set of aspects agent i is aware of in state s). Generally, this axiom indicates that each agent knows the set of aspects he is aware of in the current state and knows that he knows the set of aspects he is aware of in the current state, and so on ad infinitum. This axiom generates a straightforward criterion for awareness inference. Agent i at state s can ask himself the following questions: What is my conjecture about the set of aspects I would be aware of in state $s^{'}$? What is my conjecture about my conjecture in state $s^{'}$ about the set of aspects that I would be aware of in state $s^{'}$, and so on. If the answers to any of those questions are different from $A^1(i,s)$, then by Axiom 1, agent i in state s can exclude state $s^{'}$ from being the true state.[10] We name this thought process to be the first order awareness inference. It is the first order in the sense that it is about whether a state is realized or not. This inference principle can be formally summarized as follows:

Definition 9. ***Awareness Inference Principle (First Order)***

For all $s \in S$, agent i in state s excludes state $s^{'} \in S$ if there exists $n = 2, 3, ...$ such that

$$A^n(i^n, s^n) \neq A^1(i,s), \text{ where } i^n = (\underbrace{i, ..., i}_{n}), s^n = (s, \underbrace{s^{'}, ..., s^{'}}_{n-1}) .$$

Now consider the belief updating induced by the first order awareness inference. An agent's initial belief is the marginal distribution of π over his subjective state space. His belief is updated over the state realization by excluding states with awareness inference. We introduce some additional notations to formally represent

[10]Obviously, if Axiom 1 only holds for certain values of n, the inference criterion can be adjusted accordingly.

this belief updating process. Denote agent i's subjective state space in state s by $\prod A^1(i,s)$. Denote any subjective state in this space by $\kappa^{i,s}$. To consider an agent's subjective states, define a collection of projection operators P that projects the objective state space S onto its subspaces. Specifically, $P_{A^1(i,s)}$ projects the objective state space onto the subjective state space $\prod A^1(i,s)$. Moreover, for each $i \in I$, define a likelihood operator $L_i : S \to 2^S$. $L_i(s)$ is the set of all (objective) states that cannot be excluded by agent i in state s with the first order awareness inference. Denote agent i's updated belief in state s by $\pi_{A^1(i,s)}$.[11] For any $\kappa^{i,s} \in \prod A^1(i,s)$,

$$\pi_{A^1(i,s)}(\kappa^{i,s}) = \frac{\sum_{\hat{s} \in L_i(s), P_{A^1(i,s)}(\hat{s}) = \kappa^{i,s}} \pi(\hat{s})}{\sum_{\tilde{s} \in L_i(s)} \pi(\tilde{s})} .$$

The denominator is the sum over the probabilities of all states that are not excluded by agent i at state s. The nominator is the sum over all states that are not excluded and coincide to $\kappa^{i,s}$ when projected onto the subjective state space $\prod A^1(i,s)$. Note that the denominator is non-zero because by Axiom 1, $s \in L_i(s)$.

An agent can also perform inference over other agents' awareness inferences. The inference involving multiple agents is called the higher order awareness inference. For instance, agent i_1 in state s_1 may ask himself the following questions: What is my conjecture about agent i_2's conjecture in state s_2 about the set of aspects that agent i_2 is aware of in state $s_2^{'}$? If the conjecture is different from i_1's conjecture over the set of aspects that i_2 is aware of in state s_2, agent i_1 at state s_1 would deduce that agent i_2 would exclude state $s_2^{'}$ in state s_2. Formally, we can summarize the higher

[11] Note that this is different from the projection of π onto the agent's subjective state space. The reason is exactly that the agent can perform awareness inference.

order inference principles as follows:[12]

Definition 10. ***Awareness Inference Principle (Higher Order)***

For all $i_1, ..., i_l \in I$ *and* $s_1, ..., s_l, s_l^{'} \in S$*, let*

$$i_1^n = i_2^n = (i_1, ..., i_{l-1}, \underbrace{i_l, ..., i_l}_{n-l+1}) \ ,$$

$$s_1^n = (s_1, ..., s_{l-1}, \underbrace{s_l, ..., s_l}_{n-l+1}) \ ,$$

$$s_2^n = (s_1, ..., s_{l-1}, s_l, \underbrace{s_l^{'}, ..., s_l^{'}}_{n-l}) \ .$$

Agent i_1 *in state* s_1 *conjectures that agent* i_2 *in state* s_2 *conjectures that, ..., agent* i_l *in state* s_l *would exclude state* $s_l^{'}$ *if there exists* $n = l, l+1, ...$ *such that*

$$A^n(i_1^n, s_1^n) \neq A^n(i_2^n, s_2^n) \ .$$

The higher order inference principles can also induce predictions over other agents' beliefs in each state in a similar manner. Since the intuition is similar to the first order awareness inference, we omit the details.

[12]In this case, the consistency axiom alone cannot guarantee the correctness of higher order inferences. Yet we argue that if all agents understand the thought process of the first order inference, then the higher order inference principles are natural to be imposed.

6.2.3 Axioms and Characterizations for A^n

As discussed in the last section, an agent's awareness inference is characterized by awareness operators A^n. To achieve a reasonable prediction on agents' awareness inference, in this section, we adopt an axiomatic approach to impose restrictions on A^n. In this framework, we assume agents' subjective state spaces are exogenously determined. In other words, A^1 is taken as given.[13] Axioms are imposed on A^n for $n \geq 2$. These axioms in turn provide bounds on the power of awareness inference.

The first axiom requires that, if two states are indistinguishable in an agent's subjective state space, that agent should form the same conjecture with respect to these states. In other words, if state s and s' share the same projection on an agent's subjective state space, that agent must form the same conjecture over these states. Moreover, it is common knowledge that this axiom is applied to all agents. Thus, if agent i_1 in state s_1 conjectures that agent i_2 in state s_2 cannot distinguish states s_3 and s_3', then i_1 must conjecture that i_2's conjectures in s_2 about any agent's awareness in state s_3 and in state s_3' coincide.

Axiom 2. ***Awareness Measurability***

For all $n \geq 2$, $i_1, .., i_n \in I$ *and* $s_1, ..., s_n, s_n' \in S$, *let*

$$i^n = (i_1, ..., i_n) \ ,$$

$$s_1^n = (s_1, ..., s_{n-1}, s_n) \ ,$$

$$s_2^n = (s_1, ..., s_{n-1}, s_n') \ ,$$

[13] No constraint is imposed on A^1. It can be specified arbitrarily.

$$s^{n-1} = s_1^n \setminus s_n = s_2^n \setminus s_n^{'} \ .$$

If

$$P_{A^{n-1}(i^{n-1},s^{n-1})}(s_n) = P_{A^{n-1}(i^{n-1},s^{n-1})}(s_n^{'}) \ ,$$

then,

$$A^n(i^n, s_1^n) = A^n(i^n, s_2^n) \ .$$

The second axiom requires that an agent's conjectures only contain aspects he is aware of. In other words, agents cannot expand their subjective state spaces through higher order thinking about awareness.

Axiom 3. ***Awareness Limitation***

For any $n \geq 2$ and any sequence i^n and s^n, $A^n(i^n, s^n) \subseteq A^{n-1}(i^n \setminus i_n, s^n \setminus s_n)$.

The third axiom rules out the situation where agents correctly exclude states from wrong conjectures. To understand the necessity of this requirement, consider the following single agent example with an awareness structure satisfying Axiom 1, 2 and 3. The agent's subjective state space coincides the objective state space in any state. Namely, for all $s \in S$, $A^1((i,s)) = \mathcal{S}$. For $n \geq 2$, A^n are defined as follows: for any $s \in S$, if

$$s^n = (\underbrace{s, ..., s}_{n}) \ , \quad A^n(i^n, s^n) = \mathcal{S};$$

otherwise

$$A^n(i^n, s^n) = \emptyset \ .$$

Note that this definition implies Axiom 1. Moreover, since the agent's subjective

state space coincides with the objective state space in any state, Axiom 2 is trivially satisfied. To check Axiom 3, for any i^n and s^n, if

$$A^{n-1}(i^n \setminus i_n, i^n \setminus s_n) = \emptyset \ ,$$

then

$$A^n(i^n, s^n) = \emptyset \ .$$

If

$$A^n(i^n, s^n) = \mathcal{S} \ ,$$

then

$$A^{n-1}(i^n \setminus i_n, s^n \setminus s_n) = \mathcal{S} \ .$$

In this example, the agent can perfectly infer the true state through the first order awareness inference. However, this strong inference power stems from the agent mistakenly conjecturing that he is aware of nothing in any state other than the true state.

We impose Axiom 4 to rule out this type of "magical thinking". Specifically, if agent i_1 is aware of aspect S_j in state s_1 and agent i_2 is aware of aspect S_j in state s_2, then agent i_i in state s_1 would correctly conjecture that agent i_2 is aware of aspect S_j in state s_2. Moreover, if agent i_1 in state s_1 conjectures that agent i_2 in state s_2 and agent i_3 in state s_3 are both aware of aspect S_j, then agent i_1 in state s_1 conjectures that agent i_2 in state s_2 conjectures that agent i_3 in state s_3 is aware of aspect S_j. In other words, agents only make type-2 errors in conjectures and conjectures that

other agents also only make type-2 errors in conjectures.

Axiom 4. ***Inference from Correct Reasoning***

For all $i_1, .., i_n, i_{n+1} \in I$, $s_1, ...s_n, s_{n+1} \in S$, *let*

$$i_1^1 = (i_1),\ s_1^1 = (s_1)\ ,$$

$$i_2^1 = (i_2),\ s_2^1 = (s_2)\ ,$$

$$i^2 = (i_1, i_2),\ s^2 = (s_1, s_2)\ .$$

For $n \geq 2$, *let*

$$i_1^n = (i_1, ..., i_{n-1}, i_n),\ s_1^n = (s_1, ..., s_{n-1}, s_n)\ ,$$

$$i_2^n = (i_1, ..., i_{n-1}, i_{n+1}),\ s_2^n = (s_1, ..., s_{n-1}, s_{n+1})\ ,$$

$$i^{n+1} = (i_1, ..., i_n, i_{n+1}),\ s^{n+1} = (s_1, ..., s_n, s_{n+1})\ .$$

For all $S_j \in \mathcal{S}$ *and* $n \geq 1$, *if*

$$S_j \in A^n(i_1^n, s_1^n) \text{ and } S_j \in A^n(i_2^n, s_2^n)\ ,$$

then

$$S_j \in A^n(i^{n+1}, s^{n+1})\ .$$

Easy to see that Axioms 2-4 are independent. That is, any two axioms does not imply the third one. On the other hand, as shown in Lemma 7, Axiom 1 can be deduced from Axiom 3 and Axiom 4. Thus, it is suffice to focus on Axiom 2-4.

Lemma 7. *Axiom 3 and Axiom 4 imply Axiom 1.*

Proof. See appendix. □

Taking A^1 as given, we focus on awareness structures $\{A^n\}_{n=1}^{\infty}$ satisfying Axioms 2-4. These axioms cannot pin down an unique awareness structure. However, we can compare awareness structures satisfying Axioms 2-4 by their informativeness. Intuitively, since agents only make type-2 errors in their conjectures, an awareness operator A^n that maps sequences of agents and states to smaller sets of aspects indicates agents' more accurate understandings over subjective state spaces and other agents' conjectures at level n. Moreover, if agents have more accurate understandings of subjective state spaces and other agents' conjectures at level n under one awareness structure than another at all $n \geq 2$, roughly speaking, they have better understanding of awareness in the former awareness structure than the latter.

Formally, we introduce a partial order over $2^{\mathcal{S}}$ by set inclusion. This induces a partial order over awareness operators A^n. Specifically, for two operators $\hat{A}^n$ and $\tilde{A}^n$, $\hat{A}^n \leq \tilde{A}^n$ if and only if for all $i^n \in I^n$ and $s^n \in S^n$,

$$\hat{A}^n(i^n, s^n) \subseteq \tilde{A}^n(i^n, s^n) \ .$$

Furthermore, for two awareness structures $\{\hat{A}^n\}$ and $\{\tilde{A}^n\}$, $\{\hat{A}^n\} \leq \{\tilde{A}^n\}$ if and only if

$$\hat{A}^n \leq \tilde{A}^n \ ,\forall n \ .$$

Since we only define a partial order over the awareness structure, generally there

may not be a maximum element and a minimum element over the set of awareness structures satisfying certain conditions. However, the following theorem shows that if we consider all awareness structures satisfying Axioms 2-4, the least informative awareness structure $\{\bar{A}^n\}_{n=1}^{\infty}$ and the most informative awareness structure $\{\underline{A}^n\}_{n=1}^{\infty}$ exist.[14] We call $\{\bar{A}^n\}$ the naive awareness structure and $\{\underline{A}^n\}$ the sophisticated awareness structure for the obvious reason.

Theorem 8. ***Maximum and Minimum Awareness structures***
Fix A^1, there exists two awareness structures $\{\bar{A}^n\}_{n=1}^{\infty}$ and $\{\underline{A}^n\}_{n=1}^{\infty}$ satisfying Axioms 2-4 such that for any awareness structure $\{A^n\}_{n=1}^{\infty}$ satisfying Axioms 2-4, we have

$$\{\underline{A}^n\} \leq \{A^n\} \leq \{\bar{A}^n\} \ .$$

That is, for any $n \geq 2$, $i^n \in I^n$ and $s^n \in S^n$,

$$\underline{A}^n(i^n, s^n) \subseteq A^n(i^n, s^n) \subseteq \bar{A}^n(i^n, s^n) \ .$$

Moreover, $\{\bar{A}^n\}_{n=1}^{\infty}$ and $\{\underline{A}^n\}_{n=1}^{\infty}$ can be characterized as follows:

1. *For any $n \geq 2$, $i^n \in I^n$ and $s^n \in S^n$, $\bar{A}^n(i^n, s^n) = A^1(i_1, s_1)$*
2. *For any $n \geq 2$, $i^n \in I^n$ and $s^n \in S^n$, define an equivalence relation $\sim$ over S such that $s \sim \tilde{s}$ if*

$$P_{A^{n-1}(i^n \setminus i_n, s^n \setminus s_n)}(s) = P_{A^{n-1}(i^n \setminus i_n, s^n \setminus s_n)}(\tilde{s}) \ .$$

[14] Or in mathematical terms, the lattice of awareness structures satisfying Axioms 2-4 is bounded.

That is, state s and $\tilde{s}$ are equivalent if they share the same projection on the space generated by $A^{n-1}(i^n \setminus i_n, s^n \setminus s_n)$. Then,

$$\underline{A}^n(i^n, s^n) = \underline{A}^{n-1}(i^n \setminus i_n, s^n \setminus s_n) \bigcap \bigcup_{s:s\sim s_n} A^1((i_n, s)) \ .$$

Proof. See Appendix. □

6.2.4 The Naive and the Sophisticated Awareness Structures

In this section, we focus on discussing the naive awareness structure and the sophisticated awareness structure. The naive awareness structure corresponds to the least informative belief updating induced by awareness; the sophisticated awareness structure corresponds to the most informative belief updating induced by awareness. In this sense, we provide bounds on the informativeness of belief updating induced by awareness structure.

Under the naive awareness structure, agents' awareness inference has no power at all. Intuitively, if an agent thinks that all agents in all states are aware of the same aspects he is currently aware of, he cannot exclude any state and his conjectures about other agents' conjectures are not informative. This can be formally summarized into the following observation:

Observation 1. *Given the awareness structure $\{\bar{A}^n\}_{n=1}^{\infty}$, $L_i(s) = S$, $\forall i \in I, s \in S$.*

It is also worthwhile to consider the naive and the sophisticated awareness structure in a standard state space framework. In this situation, every agent is aware

of the objective state space in all states. Then by Theorem 8, the naive and the sophisticated awareness structure coincide.

Observation 2. *If $A^1(i,s) = \mathcal{S}$ for all $i \in I$, $s \in S$, then for all n, i^n and s^n, $\bar{A}^n(i^n, s^n) = \underline{A}^n(i^n, s^n) = \mathcal{S}$.*

These observations together suggest two premises for an agent to update his belief through first order awareness inference. First, agents' subjective state spaces depends non-trivially on the state realization. Second, the agent (at least partially) understands the state dependency of his subjective state spaces.

We now analyze awareness inference under the sophisticated awareness structure. First notice that the sophisticated awareness structure is indeed "sophisticated" in the sense that it enables an agent to obtain the most accurate inference on the state realization.

Proposition 31. *Fix A^1. Suppose agent i cannot exclude state s' in state s under the sophisticated awareness structure $\{\underline{A}^n\}$. Then agent i cannot exclude state s' in state s under any awareness structure $\{A^n\}$.*

Proof. See Appendix. □

One convenient feature of the sophisticated awareness structure is that awareness inference principles can be greatly simplified.

Proposition 32. *Given the sophisticated awareness structure $\{\underline{A}^n\}_{n=1}^{\infty}$, for any $j \geq 1$, let*

$$i^{n+j} = (i_1, ..., i_{n-1}, \underbrace{i_n, ..., i_n}_{j+1}) \ , \ s^{n+j} = (s_1, ..., s_{n-1}, \underbrace{s_n, ..., s_n}_{j+1}) \ ,$$

$\underline{A}^{n+j}(i^{n+j}, s^{n+j}) = \underline{A}^n(i^n, s^n)$ *where* $i^n = (i_1, ..., i_{n-1}, i_n)$, $s^n = (s_1, ..., s_{n-1}, s_n)$.

Proof. See Appendix. □

Specifically, the first order awareness inference principle can be simplified as follows:

Corollary 14. *Given the sophisticated awareness structure* $\{\underline{A}^n\}_{n=1}^{\infty}$, *for any* $s \in S$, *agent* i *at state* s *would exclude state* $s^{'} \in S$ *in state* s *if and only if*

$$\underline{A}^2(i^2, s^2) \neq A^1(i, s) \ , \ i^2 = (i, i), s^2 = (s, s^{'}) \ .$$

In other words, agent i *in state* s *will exclude state* $s^{'}$ *if and only if there exists* $S_j \in A^1(i, s)$ *such that*

$$\forall s^{''} \in \{\tilde{s} \mid P_{A^1(i,s)}(\tilde{s}) = P_{A^1(i,s)}(s^{'})\}, \ S_j \notin A^1(i, s^{''}) \ .$$

Proof. See Appendix. □

Moreover, under the sophisticated awareness structure, an agent's awareness inferences (both the first order and the higher order) are uniquely determined by his subjective state space, regardless of the actual state realization.

Corollary 15. *Under the sophisticated awareness structure* $\{\underline{A}^n\}$, *if* $A^1(i, s) = A^1(i, s^{'})$, *then for any*

$$i^n = \{i, i_2, ..., i_n\} \ , \ s_1^n = \{s, s_2, ..., s_n\} \ , \ s_2^n = \{s^{'}, s_2, ..., s_n\} \ ,$$

$\underline{A}^n(i^n, s_1^n) = \underline{A}^n(i^n, s_2^n)$. *Specifically, agent i makes the same first order awareness inference in state s and $s^{'}$; i.e.,* $L_i(s) = L_i(s^{'})$.

Proof. See Appendix. □

To understand the role played by the sophisticated awareness structure in this corollary, consider the first order awareness inference under an arbitrary awareness structure. If $A^1(i, s) = A^1(i, s^{'})$, by axiom 4, $s \in L_i(s^{'})$ and $s^{'} \in L_i(s)$. Yet $L_i(s)$ and $L_i(s^{'})$ can be different.

6.2.5 Exam Example Revisited

In this section, we revisit the economics exam example in the introduction and explicitly calculate students' posteriors under the sophisticated awareness structure. In this example $\mathcal{S} = \{S_1, S_2, S_3\}$ where $S_1 = \{L, M, H\}$, $S_2 = \{Y, N\}$ and $S_3 = \{Y, N\}$. S_1 represents students' type; S_2 and S_3 represent whether a student understands two important concepts (Nash Equilibrium and constrained maximization) or not. A student who does not understand a certain concept is also not aware of its relevance to the exam. A student who understands zero, one or two concepts is type L, M or H respectively. Type L ranks from 0 to 33 percentile; type M ranks from 34 to 66 percentile; type H ranks from 67 to 100 percentile. A student of an unknown type is asked to estimate his percentile in the exam.

Assume that each type is equally likely to exist and type M is equally likely to be aware of either of the two concept. Then prior to awareness inference, the probability

of each state is as follows:

$$\pi((L,N,N)) = \pi((H,Y,Y)) = \frac{1}{3} \ , \ \pi((M,Y,N)) = \pi((M,N,Y)) = \frac{1}{6} \ .$$

We use type names L, M_1, M_2 and H to refer corresponding states for the ease of notation. Specifically, M_1 stands for the state (M,Y,N) and M_2 stands for the state (M,N,Y). From the assumption,

$$A^1(i,L) = \{S_1\} \ , \ A^1(i,M_1) = \{S_1,S_2\} \ , \ A^1(i,M_2) = \{S_1,S_3\} \ , \ A^1(i,H) = \{S_1,S_2,S_3\} \ .$$

To pin down awareness inference, subjective state spaces in states with zero probability also need to be specified. We make the assumption that an agent (student) cannot be aware of anything in fictitious states. That is,

$$\forall s \notin \{L,M_1,M_2,H\} \ , \ A^1(i,s) = \emptyset \ .^{15}$$

Since we only discuss first order awareness inference here, with no confusion involved, use $A^2(s_1,s_2)$ to be an abbreviation for $A^2((i,s_1),(i,s_2))$. Remember that $A^2((i,s_1),(i,s_2))$ is agent i's conjecture of his subjective state space in state s_2 when the true state is s_1. Under the sophisticated awareness structure, the second order awareness operator's values over nonzero probability states can be pinned down as

[15]One caveat is that this assumption is not as innocuous as it looks. For example, if $A^1(i,s) = \mathcal{S}$ for $s \notin \{L,M_1,M_2,H\}$, the awareness inference result will be different.

follows.

$$\textbf{Type L}: A^2(L,L) = A^2(L,M_1) = A^2(L,M_2) = A^2(L,H) = \{S_1\} .$$

$$\textbf{Type } \boldsymbol{M_1}: A^2(M_1,L) = A^2(M_1,M_2) = \{S_1\} ,$$
$$A^2(M_1,M_1) = A^2(M_1,H) = \{S_1,S_2\} .$$

$$\textbf{Type } \boldsymbol{M_2}: A^2(M_2,L) = A^2(M_2,M_1) = \{S_1\} ,$$
$$A^2(M_2,M_2) = A^2(M_2,H) = \{S_1,S_3\} .$$

$$\textbf{Type H}: A^2(H,L) = \{S_1\} , A^2(H,M_1) = \{S_1,S_2\} ,$$
$$A^2(H,M_2) = \{S_1,S_3\} , A^2(H,H) = \{S_1,S_2,S_3\} .$$

Then by Corollary 14, with the sophisticated awareness structure, in all states the agent's priors and posteriors can be summarized as a 4-tuple $(p_L, p_{M_1}, p_{M_2}, p_H)$ where each number corresponds to the probability of a certain state. The awareness inference results are summarized in Table 6.1:

Table 6.1: Awareness Inference in the Economics Exam Example

	L	M_1	M_2	H
Prior	$(\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{3})$	$(\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{3})$	$(\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{3})$	$(\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{3})$
Type Ruled Out	None	L & M_2	L & M_1	L, M_1 & M_2
Posterior	$(\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{3})$	$(0, \frac{1}{3}, 0, \frac{2}{3})$	$(0, 0, \frac{1}{3}, \frac{2}{3})$	$(0, 0, 0, 1)$
Average Percentile	17	50	50	84
Percentile Estimation	50	72	72	84

Notice that in both states M_1 and M_2, the posterior probability of type M remains $\frac{1}{3}$. The reason is that although the posterior probability over the true state always increases after awareness inference, the agent in state M_1 (M_2) ruling out state M_2 (M_1) offsets the probability increase of type M.

6.3 Awareness Signal and Awareness Inference

In this section, we investigate an agent's belief updating when he learns new aspects and his subjective state spaces expand. Notice that this is very different from Bayesian learning. The agent is not receiving signals that enables him to update his belief with Bayes's rule. Instead, the agent is learning that he should take new elements and possibilities into consideration. In this framework, an agent cannot be aware of new aspects by forming conjectures base on the awareness structure. Thus, we model this learning process as an agent receiving a signal containing (possibly new) aspects. Upon receiving this signal, the agent becomes aware of new aspects in the signal and also learns the joint distribution of aspects in his current subjective state space and new aspects. We call this signal an awareness signal. It differs from

a Bayesian signal in that an awareness signal enlarges the agent's subjective state space but does not directly change the agent's belief on his previous state space.[16]

The expansion of an agent's subjective state spaces has two effects on belief updating. The first effect is straightforward. With a larger subjective state space, the agent has access to more Bayesian signals for belief updating. To better illustrated this point, it is helpful to re-frame the story of Holmes and Watson discussed in Li (2009) and many other papers about unawareness under this framework. In the case of Silver Blaze, Holmes and Watson tried to figure out whether there was intruder in the stable. Thus, the key aspect of interest is $S_1 = \{y, n\}$. y represents an intruder and n represents no intruder. To think about Bayesian updating, suppose the prior belief is that each situation happens with probability $\frac{1}{2}$. Another relevant aspect to this problem is whether the dog barked that night. If the dog barked, there must be an intruder and if the dog did not bark, no intruder exists. Denote this aspect by $S_2 = \{b, nb\}$. b stands for "barked" and nb stands for "not barked". The objective state space is $S_1 \times S_2$. Being a famous detective, in all states, Holmes' subjective state spaces coincide the objective state space. As soon as the police provided the Bayesian signal that the dog did not bark, Holmes was able to update his belief to the posterior that no intruder existed. On the other hand, Watson was not aware of the link between dog barking and existence of an intruder and his subjective state spaces are S_1 in all states. Thus, his belief remained at the prior even after the police said that the dog did not bark. Consider the situation where someone sent Watson

[16]For example, if we want to understand the severity of global warming, checking the data of carbon dioxide level can be considered as receiving a Bayesian signal; gaining knowledge about how human activity is related to global warming can be considered as receiving an awareness signal.

an awareness signal by telling him that dog barking or not indicates the existence of an intruder. Then, Watson's subjective state space becomes $S_1 \times S_2$ and he can utilize the Bayesian signal from the police to reach the correct conclusion that there was no intruder.

In this example, receiving an awareness signal merely serves as a trigger for Bayesian updating. Thus, this belief updating triggered by the awareness signal shares an essential feature with Bayesian updating. That is, the belief process is a martingale. Indeed, if we insist that the prior belief is "correct", then in another state of the world realized with probability $\frac{1}{2}$, the police would tell them the dog has barked. And after receiving the awareness signal, Watson would update his belief to assign probability one to "there was an intruder". This kind of belief updating triggered by an awareness signal is indistinguishable from Bayesian updating.

In this framework, receiving an awareness signal has another effect on belief updating. By expanding an agent's subjective state spaces, receiving the awareness signal effectively changes the A^1 operator to include more states. This may lead to new awareness structures due to their dependence on A^1. In this sense, receiving the awareness signal enables the agent to perform awareness inference multiple times with different awareness structures. In this way, the agent may exclude more states and his belief may become more accurate. Importantly, the effectiveness of belief updating through this channel depends on the agent's subjective state space and the belief process is not necessarily a martingale. We focus on discussing this novel belief updating channel.

In this section, fix the agent receiving an awareness signal to be agent i. An

awareness signal is a set of aspects $\phi \subseteq 2^{\mathcal{S}}$. We assume that receiving an awareness signal itself does not provide agent i any information about the state realization. That is, upon receiving the awareness signal, agent i thinks that he would receive the same awareness signal regardless of state realization.

This assumption enable us to focus solely on the informational aspect of an agent's state space expansion. Specifically, if the event of receiving awareness signal itself changes the agent's belief over state realizations, we can always decompose the signal into two parts. Receiving the signal itself serves as a standard Bayesian signal and the agent thinks that he would receive the same awareness signal under all possible states. Moreover, similar to Bayesian signal, we require awareness signals to be "sequence-free". That is, when an agent receives multiple awareness signals, the order of signals does not affect his belief updating. Following this assumption, receiving multiple awareness signals is equivalent to receiving the union of these signals as one awareness signal. Thus, with these assumptions we focus on considering the belief updating process of agent i receiving one awareness signal $\phi = \{\phi_1, ..., \phi_k\} \subseteq 2^{\mathcal{S}}$.

In this framework, agent i's belief updating induced by an awareness signal can be considered as conducting multiple awareness inferences under awareness structures with different A^1. Given that all these awareness structures satisfy Axioms 2-4, we can use results derived in section 6.2 to analyze the informativeness of an awareness. Formally, for any awareness signal $\phi = \{\phi_1, ..., \phi_k\}$, we can pick a sequence of aspects in ϕ. For the ease of notation, let the sequence be $(\phi_1, ..., \phi_k)$. This sequence induces an evolution of all agents subjective state spaces, i.e., an evolution of A^1. We use a

set of operators A^1_m to track this evolution. Specifically, for all $s \in S$,

$$A^1_0(i,s) = A^1(i,s)\ ,\ A^1_m(i,s) = A^1_{m-1}(i,s)\bigcup\{\phi_m\} \qquad \forall m = 1,...,k\ .$$

For $j \neq i$,

$$A^1_m(j,s) = A^1_0(j,s) = A^1(j,s)\ .$$

Then for each m, given A^1_m, we may consider an awareness structure $\{A^n_m\}$ satisfying Axioms 2-4.[17] This in turn pins down agents' conjectures after agent i learned aspects $\{\phi_1,...,\phi_m\}$. For each awareness structure at a given m, we can apply inference principles developed in section 6.2 to determine an agent's belief updating process. This transforms the belief updating induced by an awareness signal into a series of belief updating when agents have different underlying subjective state spaces.

However, different orders of elements in the set $\phi = \{\phi_1,...,\phi_k\}$ would lead to different operators A^1_m and generally different awareness inference results.[18] Since there is no good reason to pick any particular sequence over others, we consider awareness structures under all possible permutations of the set $\{\phi_1,...,\phi_k\}$. In practice, this corresponds to agent i making thought experiments over the order of learning different aspects to maximize the information gain. For example, let $\phi = \{\phi_1, \phi_2\}$. After receiving awareness signal ϕ, agent i would consider his and other agents' conjectures when he learns ϕ_1 first and then ϕ_2. To fully utilize this awareness signal, he would also consider his and other agents' conjectures when he learns ϕ_2 first and then ϕ_1.

[17]By applying axiom 4, we implicitly assume that if agent i receives an awareness signal containing S_i and agent j is aware of S_i, then agent j knows agent i receives this awareness signal.

[18]See Appendix for a concrete example.

Formally, for agent i, the first order awareness inference can be defined as follows:[19]

Definition 11. ***Inference Principle with Awareness Signals (First Order)*** *For all $s \in S$, agent i would exclude state $s^{'} \in S$ at state s if there exists a permutation of $\{\phi_1, ..., \phi_k\}$, $m = 0, 1, ..., k$ and integer n such that*

$$A_m^n(i^n, s^n) \neq A_m^1(i, s) \ , \ i^n = (\underbrace{i, ..., i}_{n}) \ , \ s^n = (s, \underbrace{s^{'}, ..., s^{'}}_{n-1}) \ .$$

The induced probability measure would be: for all $\kappa^{i,s} \in \prod A_k^1(i, s)$,

$$\pi_{A_k^1(i,s)}(\kappa^{i,s}) = \frac{\sum_{\hat{s} \in L_i(s), P_{A_k^1(i,s)}(\hat{s}) = s^{i,s}} \pi(\hat{s})}{\sum_{\tilde{s} \in L_i(s)} \pi(\tilde{s})} \ .$$

The higher order awareness inference principles can be defined in a similar manner as follows:

Definition 12. ***Inference Principle with Awareness Signals (Higher Order)***

For all $i_1, ..., i_l \in I$ and $s_1, ..., s_l, s_l^{'} \in S$, agent i_1 at state s_1 thinks that agent i_2 at state s_2 thinks that, ..., agent i_l at state s_l would exclude state $s_l^{'}$ if there exists a permutation of $\{\phi_1, ..., \phi_k\}$, $m = 0, 1, ..., k$ and $n = l, l+1, ...$ such that

$$A_m^n(i_1^n, s_1^n) \neq A_m^n(i_2^n, s_2^n), \quad i_1^n = i_2^n = (i_1, ..., i_{l-1}, \underbrace{i_l, ..., i_l}_{n-l+1}),$$

$$s_1^n = (s_1, ..., s_{l-1}, \underbrace{s_l, ..., s_l}_{n-l+1}), s_2^n = (s_1, ..., s_{l-1}, s_l, \underbrace{s_l^{'}, ..., s_l^{'}}_{n-l}) \ .$$

[19]Notice that the first order awareness inference for agents receiving no awareness signal remains the same as specified in section 6.2.

These inference principles are hard to check in general. If an awareness signal ϕ contains k aspects, there are $k!$ possible permutations. However, if we limit our attentions to the least and the most informative structures, the inference principles can be simplified. As some readers might have expected, if for any sequence of elements in ϕ, all awareness structures are naive (i.e., awareness structures $\{A_m^n\}$ attain the maximum for all m and all sequence of aspects in ϕ), the belief updating from this channel is trivial. It is straightforward and thus we omit the details. On the contrary, if all awareness structures are sophisticated (i.e., awareness structures $\{A_m^n\}$ attain the minimum for all m and all sequence of aspects in ϕ), inference principles have simple forms with clear interpretations.

6.3.1 Awareness Signal in the Sophisticated Case

In this section, we consider belief updating induced by an awareness signal when awareness structures $\{A_m^n\}$ attain minimum for all m and all sequences of aspects in ϕ. In other words, $A_m^n = \underline{A}_m^n$ derived in section 6.2 for all n, m and permutations of ϕ. This corresponds to the most informative belief updating induced by an awareness signal. We focus on discussing the first order awareness inference principle under this assumption. The higher order awareness inference principles' derivations are similar. Formally, the first order awareness inference principle has a simple form.

Theorem 9. *For any awareness signal $\phi = \{\phi_1, ..., \phi_k\}$, suppose $A_m^n = \underline{A}_m^n$ for all n,*

m and permutations of ϕ. Then agent i at state s will exclude state s' if and only if

$$A^1(i,s) \neq A^1(i,s) \bigcap \bigcup_{\tilde{s}:\tilde{s}\sim s'} A^1(i,\tilde{s}) \ ,$$

where the equivalence relation is induced by $A^1_k(i,s) = A^1(i,s) \bigcup \phi$. $A^1(i,s)$ is the set aspects agent i is aware of in state s before receiving the awareness signal ϕ.

Proof. See Appendix. □

It is informative to compare this theorem with the first order awareness inference under sophisticated awareness structure in section 6.2. With no awareness signal, agent i at state s excludes state s' if and only if

$$A^1(i,s) \neq A^1(i,s) \bigcap \bigcup_{\tilde{s}:\tilde{s}\sim s'} A^1(i,\tilde{s})$$

where the equivalence relation is induced by $A^1(i,s)$. The only difference between two inference principles lies in the subjective state spaces that generate the equivalence relations. With no awareness signal, two states are considered the same if they share the projection on the state space generated by $A^1(i,s)$. With an awareness signal, two states are considered equivalent only if they share a projection on the larger subjective state space generated by $A^1_k(i,s)$. In this sense, growing awareness improves an agent's inference power by increasing his ability to differentiate states.

This theorem also hints an important difference between belief updating induced by an awareness signal and a Bayesian signal. A Bayesian signal may have arbitrary level of precision, ranging from not informative at all to perfectly revealing. On the

other hand, an awareness signal can not achieve arbitrary level of informativeness. Generally, even with an awareness signal revealing the entire objective state space, certain states cannot be excluded by the awareness inference. The following corollary characterizes the informativeness bound on an awareness signal.

Corollary 16. *If $A^1(i,s) \subseteq A^1(i,s')$, agent i in state s can not exclude state s' by receiving any awareness signal. With the sophisticated awareness structure, if $A^1(i,s) \not\subseteq A^1(i,s')$, there exists an awareness signal upon receiving which agent i in state s can exclude s' through belief updating.*

Proof. See Appendix. □

6.4 Application: Persuasion with Awareness Signal

In this section, we present examples of applying this framework of awareness into persuasion games. These examples discuss whether awareness signals can be more effective than Bayesian signals in persuasion and highlight the difference between Bayesian inference and awareness inference. Consider a game where a sender can send a signal to a receiver to influence the receiver's decision. Literature on Bayesian persuasion such as Aumann et al. (1995), Gentzkow and Kamenica (2011) and Gentzkow and Kamenica (2016) discusses the situation when the sender uses a Bayesian signal. Under the framework, the sender may also consider sending an awareness signal. In the first example, the effectiveness of the awareness signal depends on a following exogenous Bayesian signal on an aspect that the receiver is unaware of previously.

This situation is similar to the Holmes and Watson story discussed in section 6.3. In this case, the belief process is a martingale and the posteriors induced by the awareness signal can be replicated by a Bayesian experiment. In the second example, the awareness signal changes the receiver's belief through awareness inference. Since the belief process induced by awareness inference may not be a martingale, persuasion by sending the awareness signal can be more effective than conducting any Bayesian experiment.

6.4.1 An Example without Awareness Inference

Consider a situation where the sender is an short-seller and the receiver is another investor. The short-seller takes short position on a specific stock and wish to persuade the receiver to sell that stock. The receiver can take three possible actions: buying(H), selling(L) and no change(U). Let u be the short-seller's utility function depending on the receiver's decision. $u(L) > u(U) > u(H)$. The objective state space is the product space $R \times M$. Aspect R stands for the future return of the stock and has possible realizations R_H and R_L. Aspect M stands for the corporate management condition and has possible realizations M_H and M_L. The probability distribution on the objective state space is $\pi((R_H, M_H)) = \pi((R_L, M_L)) = \frac{1}{2}$. That is, the condition of corporate management determines the future return. The sender is aware of both aspects while the receiver is only aware of aspect R regardless of the state realization. In this case, neither of them can update their beliefs through awareness inference. The sender's prior agrees to the probability measure π and the receiver's prior is $\pi_R(R_H) = \pi_R(R_L) = \frac{1}{2}$.

The receiver's decision depends on R. He will buy if $E(R) \geq \bar{R}$ and will sell if $E(R) \leq \underline{R}$. Otherwise he will hold his current position. Suppose that without any signal, $E(R) = \frac{1}{2}R_H + \frac{1}{2}R_L \in (\underline{R}, \bar{R})$; $R_H > \bar{R}$; $R_L < \underline{R}$. Suppose a Bayesian signal, for example, a product release conference, perfectly reveals the management situation of the company. The sender can send an awareness signal $\{M\}$ to let the receiver know about aspect M before this Bayesian signal realizes.[20] Then if $\frac{1}{2}u(L) + \frac{1}{2}u(H) > u(U)$, the sender would prefer to send the awareness signal $\{M\}$ to the receiver.

In this example, the awareness signal alters the receiver's decision by enabling the receiver to update his belief with the Bayesian signal. Thus, if the sender can use Bayesian experiment to induce any posterior respecting the martingale constraint, he can replicate the effect of sending this awareness signal by performing a Bayesian experiment. In this case, he can induce the same belief updating of the receiver by perform a Bayesian experiment that perfectly reveals the realization of R. In other words, when awareness inference does not play a role in belief updating, performing a Bayesian experiment always weakly dominates sending an awareness signal.

6.4.2 An Example with Awareness Inference

In this example, we show that when the receiver updates belief with awareness inference, sending an awareness signal might dominate any possible Bayesian signal. Similar to the previous example, a short-seller (the sender) j tries to persuade an-

[20]If the sender can send out the awareness signal after observing the Bayesian signal's realization, he can do better. Yet this is beyond the scope of this paper. We assume that the sender needs to commit to an awareness signal before the Bayesian signal realizes.

other investor (the receiver) i to sell a stock. We make three different assumptions from the previous example. First, we consider the situation when both agents are sophisticated. Second, no Bayesian signal revealing the realization of aspect M exists. With this assumption, an awareness signal can only change the receiver's belief through awareness inference. Third, the objective state space is $S = R \times r \times M$. The definitions of R and M are similar to the ones in the previous example. Aspect $r = \{r_H, r_L\}$ contains potential rumors and hearsays about the company. Event $\{r = r_H\}$ represents that some rumor exists; Event $\{r = r_L\}$ represents that rumors are very concerning.

The sender is aware of all relevant aspects regardless of the state realization, i.e., his subjective state spaces coincide the objective state space in all states. This assumption has two implications. First, since his subjective state spaces are not state-dependent, awareness inference cannot help him gain private information about the state realization. Second, through the second order awareness inference, he makes the correct conjecture about the receiver's awareness inference after an awareness signal is sent by him. The receiver is always aware of the aspect R and is never aware of the aspects M. The receiver is aware of the aspect R, i.e., the existence of rumors, in two situations: either when the company's management is not good (event $\{M = M_L\}$) and rumors are spreading or when rumors about the company are particularly concerning (event $\{r = r_L\}$). However, the receiver himself does not completely understand this relation because he is unaware of aspect M.

As in the previous example, the realization of R is perfectly determined by the

realization of M. Specifically, four states happen with positive probability:

$$s_1 = (R_L, r_L, M_L) \ , \ s_2 = (R_L, r_H, M_L) \ , \ s_3 = (R_H, r_L, M_H) \ , \ s_4 = (R_H, r_H, M_H) \ .$$

The probability distribution π over the objective state space satisfies

$$\pi(s_1) = \pi(s_3) = \frac{1}{2} - q \ , \ \pi(s_2) = \pi(s_4) = q \ , \ q \in (0, \frac{1}{2}) \ .$$

Aspects that the receiver is aware of under possible states are:

$$A^1(i, s_1) = A^1(i, s_2) = A^1(i, s_3) = \{R, r\} \ , \ A^1(i, s_4) = \{R\} \ .$$

Consider probabilities assigned to events $\{R = R_H\}$ and $\{R = R_L\}$ by the receiver without any awareness signal. Let $I^2 = \{i, i\}$, $S^2 = (s, s^{'})$. Easy to check that if $s \in \{s_1, s_2, s_3\}$, then for any $s^{'} \in S$,

$$A^2(I^2, S^2) = A^1(i, s) = \{R, r\} \ .$$

If $s = s_4$, then for any $s^{'} \in S$,

$$A^2(I^2, S^2) = A^1(i, s) = \{R\} \ .$$

This implies that no state can be excluded by the receiver, regardless of the state realization. Thus, in state s_1, s_2 or s_3, $\pi_{R \times r}(\{R = R_H\}) = \pi_{R \times r}(\{R = R_L\}) = \frac{1}{2}$. In state s_4, $\pi_R(\{R = R_H\}) = \pi_R(\{R = R_L\}) = \frac{1}{2}$.

Now we examine whether the sender can benefit from sending the receiver an awareness signal $\{M\}$. Upon receiving this signal, the receiver's subjective state space becomes $S = R \times r \times M$ in state s_1, s_2 or s_3 and $R \times M$ in state s_4. Since

$$A^1(i, s_4) \subset A^1(i, s_1) = A^1(i, s_2) = A^1(i, s_3) \ ,$$

by Corollary 16, the receiver will exclude s_4 in state s_1, s_2 or s_3 but he cannot exclude any state at s_4. Thus, at states other than s_4, the receiver's probabilistic estimation on relevant events will be

$$\pi_{R\times r\times M}(\{R = R_H\}) = \frac{1-2q}{2-2q}$$

and

$$\pi_{R\times r\times M}(\{R = R_L\}) = \frac{1}{2-2q} \ .$$

In state s_4, the receiver's belief remains to be

$$\pi_{R\times M}(\{R = R_H\}) = \pi_{R\times M}(\{R = R_L\}) = \frac{1}{2} \ .$$

We can then compare expected payoffs of the sender between sending awareness signal $\{M\}$ and sending the best possible Bayesian signal. Since $R_L < \underline{R}$, for q close enough to $\frac{1}{2}$,

$$\frac{1-2q}{2-2q}R_H + \frac{1}{2-2q}R_L < \underline{R} \ .$$

The receiver will take action L after receiving the awareness signal in states s_1, s_2,

s_3 and take action U in state s_4. The sender's expected utility by sending awareness signal $\{M\}$ is $(1-q)u(L)+qu(U)$. From the last example, if $u(U)-u(H)$ is small, the sender's expected utility with Bayesian persuasion is $(1-p)u(L)+pu(H)$ with

$$\frac{(\frac{1}{2}-p)R_H+\frac{1}{2}R_L}{1-p}=\underline{R}\ .$$

This implies that under certain values of q, awareness persuasion dominates all possible Bayesian signals. Specifically, when $q=p$, the sender is better off with awareness persuasion since $u(U)>u(H)$.

6.4.3 Discussion

The key difference between awareness persuasion and Bayesian persuasion lays in the fact that the belief updating induced by an awareness signal may not be a martingale. This is because the agent's ability to exclude states depend not only on the awareness signal but his subjective state space as well. Particularly, an agent in different states with different subjective state spaces would exclude different states even after receiving the same awareness signal. Thus, after receiving an awareness signal, the agent's expectation of posteriors does not necessarily equal to the prior.

In the previous example, the expectation of posteriors over the event $\{R=R_H\}$ is

$$q\times\frac{1}{2}+(1-q)\times\frac{1-2q}{2-2q}=\frac{1}{2}-\frac{q}{2}<\frac{1}{2}\ .$$

Since the belief induced by a Bayesian signal is always a martingale, awareness per-

suasion may dominate Bayesian persuasion by inducing a non-martingale belief process. One caveat: awareness persuasion depend not only on the awareness structure of the receiver, but the sender's ability to predict the receiver's awareness structure as well. In this example, the sender is aware of all aspects regardless of the state realization and can thus perfectly predict the receiver's awareness inference. It would be interesting for future research to consider awareness persuasion when the sender only partially understands the receiver's awareness structure.

This awareness structure provides another way to understand expertise in the persuasion literature. The sender receiving a private signal before persuasion seems to be the most straightforward way to model expertise. However, Alonso and Câmara (2018) show that the sender can never benefit from his expertise if he can perform all possible Bayesian experiments. A key element of the intuition relies on the fact that the belief process is a martingale. Thus, a sender with no private information can always replicate the receiver's belief process induced by a sender with private information. Yet after a sender receives the private information, he might have incentive to deviate from the optimal persuasion scheme. This leads to a lower ex ante payoff for a sender with private information. This framework provides an alternative way to model the sender's expertise and how can he benefit from it. If the expertise of the sender is modeled as knowing an additional aspects regardless of the state realization, the sender is better-off since he has more choice over sending awareness signal. Specifically, in the previous example, if the sender is not aware of aspect M, his expected payoff would be lower because he can only persuade the receiver through sending a Bayesian signal.

6.5 Conclusion

In this paper, we present a framework to link awareness to belief updating. We formally define awareness structures and induced awareness inference principles to demonstrate how awareness interacts with belief updating. Specifically, we derive bounds on the informativeness of awareness structures under three axioms. Moreover, this framework enables us to model an agent's belief updating after his awareness grows, i.e., the subjective state space expansion via learning new aspects. The key difference between Bayesian updating and updating through awareness is that the effectiveness of awareness inference is state-dependent. Thus, the belief process induced by awareness inference is not necessarily a martingale. This novel feature makes sending an awareness signal sometime more effective then sending any Bayesian signal in a persuasion game.

Chapter 7

Conclusion

With the rapid development of financial markets in the past decades, many new problems concerning market microstructure emerge. In this dissertation, I investigate three important topics related to the trend of electronic trading: (1) the competition between high-frequency traders and traditional market makers and its market quality implications, (2) the execution quality of large traders with increased trading frequency and difficulties over implementing reputation-based strategies and (3) insider trading with uncertainty over the insider's existence. I also explore two other elements under the frameworks of game and decision theory: (1) limited attention in obtaining signals and (2) unexpected factors and their interaction with belief updating. Both should be incorporated into future microstructure models to achieve a deeper understanding of the financial markets.

Appendix A

Appendix to Chapter 2

A.1 Base Case Proofs and Claims

A.1.1 Useful Results

Lemma 8. *$(1 - F(x))x$ is unimodal.*

Proof. Note that

$$\begin{aligned}[(1 - F(x))x]' &= 1 - F(x) - xf(x) \\ &= [1 - F(x)](1 - x\frac{f(x)}{1 - F(x)}) \ .\end{aligned}$$

$1 - F(x) \geq 0$ and $1 - x\frac{f(x)}{1-F(x)}$ is continuous and decreasing. Thus, either there exists a unique x^* such that $1 - x^*\frac{f(x^*)}{1-F(x^*)} = 0$ or $1 - x\frac{f(x)}{1-F(x)} > 0$ for all $x \in [0, \hat{x}]$. In the latter case let $x^* = \hat{x}$. Easy to see that for $x > x^*$, $[(1 - F(x))x]' < 0$; for $x < x^*$, $[(1 - F(x))x]' > 0$. □

A.1.2 No HFT

Proof of Theorem 1

Proof. First consider a relaxed problem with $d = w - q \in [-\bar{q}, w]$. Conjecture that the optimal policy is $d_t = w_t - \bar{q}$ and $x_t = x^*$ where $x^* = argmax(1 - F(x))x$, forall t. If this policy is indeed the optimal policy for this relax problem, then for $w_0 \geq \bar{q}$, this optimal policy is applicable and thus also optimal for the original more constrained problem. This proposition also implies that the market maker's payoff is linear in w_0 with $w_0 \geq \bar{q}$.

We use a method similar to one-shot deviation principle to establish the optimality of proposed policy. Notice that although the market maker discounts future dividends, the per-period dividend does not necessarily have a uniform bound. Thus, I directly check that this problem is continuous at infinity.

Consider two dividend and pricing policies $\{d_t, x_t\}_{t=0}^{\infty}$ and $\{\tilde{d}_t, \tilde{x}_t\}_{t=0}^{\infty}$. d_t, x_t, $\tilde{d}_t$, $\tilde{x}_t$ are functions of h^t, the history of the first $t-1$ periods.[1] We suppress the dependence for the ease of notation. Consider the case when $d_t = \tilde{d}_t$ and $x_t = \tilde{x}_t$ for $t \leq T$. Define the absolute value of the difference in expected payoffs between two policies to be D_T.

[1] We define $h^0 = \emptyset$.

We have

$$\begin{aligned} D_T &= |E_0(\sum_{i=T+1}^{\infty} \delta^i (d_i - \tilde{d}_i))| \\ &\leq |E_0(\sum_{i=T+1}^{\infty} \delta^i c_i)| + \delta^{T+1} \frac{1}{1-\delta} \bar{q} \\ &\leq \delta^{T+1} E_0(w_{T+1}) + \sum_{i=T+1}^{\infty} \delta^i \bar{x} E_G(q) + \delta^{T+1} \frac{1}{1-\delta} \bar{q} \\ &= \delta^{T+1} E_0(w_{T+1}) + \delta^{T+1} \frac{1}{1-\delta} \bar{x} E_G(q) + \delta^{T+1} \frac{1}{1-\delta} \bar{q} \ . \end{aligned}$$

The first inequality is because the worst dividend plan after period T is to pay $-\bar{q}$ for all periods. The second inequality is because for any period t, the expect profit is $(1 - F(x_t))x_t E_G(min(q, w_t - d_t))$.[2] This is uniformly bounded by $\bar{x} E_G(q)$. Thus, in each period, the expected dividend is bounded by $\bar{x} E_G(q)$ plus part of the market maker's net worth in period $T+1$. Notice that commit more shares cannot improve the expected dividend bound since $E_G(min(q, w)) \leq E_G(q)$. Thus, the expected discounted dividend payout is bounded by the case when the market maker pays dividend equal to the entire net worth in period $t = T+1$ and pays the upper bound of expected profit in each period.

Notice that $\delta^{T+1} \frac{1}{1-\delta} \bar{x} E_G(q) \to 0$ and $\delta^{T+1} \frac{1}{1-\delta} \bar{q} \to 0$ as $T \to \infty$. Moreover, $E_t(w_{t+1}) \leq w_t + \bar{x} E_G(q) + \bar{q}$. This implies that

$$\delta^{T+1} E_0(w_{T+1}) \leq \delta^{T+1} [w_0 + (T+1)(\bar{x} E_G(q) + \bar{q})] \ . \tag{A.1}$$

[2] E_G means q follows distribution G, I suppress the time notation because demands are *i.i.d.*

Thus, $\delta^{T+1}E_0(w_{T+1}) \to 0$ as $T \to \infty$. Thus, for any two policies that different only after period T, as $T \to \infty$, $D_T \to 0$.

Since this game is continuous at infinity, if there exists a profitable deviation, then there exists a profitable deviation such that the deviating policy is different from the candidate policy for finite periods. Consider a deviation where the the deviating policy is different from the candidate policy for n periods. For $t \geq n$, the deviating policy switches back to the candidate policy $\hat{d}_t = w_t - \bar{q}$ and $\hat{x}_t = x^*$. Consider the market maker in period $t = n$ with net worth w_n. Suppose the deviating policy specifics $\hat{d}_n = w_n - \hat{w}$ and $x_n = \hat{x}_n$. Then in period n, the difference between expected payoffs of two policies is

$$\begin{aligned} E_n(d_n - \hat{d}_n + \delta d_{n+1} - \delta \hat{d}_{n+1}) &= \hat{w} - \bar{q} + \delta(1 - F(x^*))x^* E_G(min(q, \bar{q})) \\ &\quad - \delta(1 - F(\hat{x}_n))\hat{x}_n E_G(min(q, \hat{w})) - \delta(\hat{w} - \bar{q}) \\ &\geq (1 - \delta)(\hat{w} - \bar{q}) \\ &\quad + \delta(1 - F(x^*))x^*[E_G(min(q, \bar{q})) - E_G(min(q, \hat{w}))] \ . \end{aligned}$$

The inequality follows from $(1 - F(\hat{x}_n))\hat{x}_n \leq (1 - F(x^*))x^*$.

Define

$$A(y) = (1 - \delta)(y - \bar{q}) + \delta(1 - F(x^*))x^*[E_G(min(q, \bar{q})) - E_G(min(q, y))] \ .$$

Then

$$A^{'}(y) = 1 - \delta - \delta(1 - F(x^*))x^*(1 - G(y)) \ ,$$

$$A^{''}(y) = g(y) > 0 \ .$$

Since $A^{'}(y)$ is monotone, $A^{'}(y) = 0$ has at most one solution and upon which $A(y)$ achieves minimum. Note that $A^{'}(y) = 0$ implies

$$\frac{\delta}{1-\delta}(1 - F(x^*))x^*(1 - G(y)) = 1 \ .$$

Thus, $A(y)$ achieves minimum at $y = \bar{q}$ and $A(\bar{q}) = 0$. Thus,

$$E_n(d_n - \hat{d}_n + \delta d_{n+1} - \delta \hat{d}_{n+1}) \geq 0 \ . \tag{A.2}$$

This implies that if there exists a profitable deviation such that the deviating policy differs from the candidate policy for n periods, then in period n, the market maker should adopt the candidate policy. Same reasoning then shows that the market maker should adpot the candidate policy in period $n - 1$. The backward induction goes back to period 1. Since n is arbitrary and this problem is continuous at infinity, no profitable deviation exists and the candidate policy is optimal.

□

Existence of Value Function

Proposition 33. *There's a unique V such that it is continuous and strictly increasing in w.*

Proof. We focus on $V(w)$ for $w \in [0, \hat{w}]$. Moreover, since $V(w) = w - \bar{q} + V(\bar{q})$ Define

operator T to be

$$\begin{aligned}(Tl)(w) =& sup_{d,x} d + \delta\{F(x)l(w-d) \\ &+ (1-F(x))[\int_0^{w-d} (l(min(\bar{q}, w-d+xq)) \\ &+ max(0, w-c+xq-\bar{q})g(q))dq \\ &+ (1-G(w-d))(l(min(\bar{q}, (1+x)(w-d))) \\ &+ max(0, (1+x)(w-d)-\bar{q}))]\}\end{aligned} \tag{A.3}$$

satisfying $c \in [0, w]$.

First check that for large enough $\bar{K}$, $l(w) \leq \bar{K} \implies Tl(w) \leq \bar{K}$. Thus, the value function is bounded and Blackwell condition is applicable. Easy to check T satisfies monotonicity and discounting.

By contract mapping theorem, operator T has a unique fixed point V. Easy to see T maps increasing functions to strictly increasing functions. This implies V must be increasing. □

A.1.3 Sequential Pricing

Proof of Lemma 1

Proof. If $x_m > x^*$, since $argmax_x(1-F(x))x = x^*$, the HFT's optimal strategy is to set $x_h = x^*$. Consider the situation when $x_m \leq x^*$. For $x_h \leq x_m$, the HFT's expected net profit is

$$(1-F(x_h))x_h k(q_h) \ ,$$

which attains maximum at $x_h = x_m$ by lemma A.1.1. For $x_h > x_m$, the HFT's expected net profit is

$$(1 - F(x_h))x_h[k(q_h + q_m) - k(q_m)] \ ,$$

which attains maximum at $x_h = x^*$. □

Proof of Lemma 2

Proof. First notice that $x_m > x^*$ cannot be optimal. If $x_m > x^*$, the HFT's best response is to set $x_h = x^*$ and the market maker's expected net profit is

$$\begin{aligned}&(1 - F(x_m))x_m[\pi(k(q_h + q_m) - k(q_h)) + (1 - \pi)k(q_m)] \\ &< (1 - F(x^*))x^*[\pi(k(q_h + q_m) - k(q_h)) + (1 - \pi)k(q_m)] \ .\end{aligned}$$

This implies the market maker will be better off by setting $x_m = x^*$.

Next, there is a unique $\underline{x} < x^*$ such that if $x_m = \underline{x}$, the HFT is indifferent between $x_h = x^*$ and $x_h = x_m$. For any $x_m < x^*$, the HFT's expected net profit with $x_h = x^*$ is

$$(1 - F(x^*))x^*[k(q_h + q_m) - k(q_m)] \ ;$$

the HFT's expected net profit with $x_h = x_m$ is

$$(1 - F(x_m))x_m k(q_h) \ .$$

Since $(1 - F(x))x$ is increasing for $x \in [0, x^*]$ and $k(q_h + q_m) - k(q_m) < k(q_h)$, there

exists a unique $\underline{x} \in (0, x^*)$ such that

$$(1 - F(\underline{x}))\underline{x}k(q_h) = (1 - F(x^*))x^*[k(q_h + q_m) - k(q_m)] \ ,$$

or equivalently,

$$a(\underline{x})k(q_h) = k(q_h + q_m) - k(q_m) \ .$$

Finally, check that any other pricing strategy of the market maker is dominated either by $x_m = x^*$ or $x_m = \underline{x}$. If $x_m \in (\underline{x}, x^*)$, the HFT would set $x_h = x_m$. The market maker's expected net profit is

$$\begin{aligned}&(1 - F(x_m))x_m[\pi(k(q_h + q_m) - k(q_h)) + (1 - \pi)k(q_m)] \\ &< (1 - F(x^*))x^*[\pi(k(q_h + q_m) - k(q_h)) + (1 - \pi)k(q_m)] \ .\end{aligned}$$

Thus, he would be better off switch to $x_m = x^*$. For $x_m \in (0, \underline{x})$, the HFT would set $x_h = x^*$. The market maker's expected net profit is

$$(1 - F(x_m))x_m k(q_m) < (1 - F(\underline{x}))\underline{x}k(q_m) \ .$$

This suggests that he would be better off to set $x_m = \underline{x}$. □

Proof of Proposition 1

Proof. For any q_m, the tight spread can be determined by the equation

$$a(\underline{x}(q_m)) = \frac{k(q_m + q_h) - k(q_m)}{k(q_h)} \ . \tag{A.4}$$

The tight spread strategy strategy is optimal if

$$a(\underline{x}(q_m))k(q_m) \geq \pi[k(q_m + q_h) - k(q_h)] + (1 - \pi)k(q_m) \ . \tag{A.5}$$

Subtract $k(q_m)$ from both sides,

$$\frac{k(q_m + q_h) - k(q_m) - k(q_h)}{k(q_h)} k(q_m) \geq \pi[k(q_m + q_h) - k(q_h) - k(q_m)] \ . \tag{A.6}$$

Since $k(q_m + q_h) - k(q_h) - k(q_m) < 0$ for $q_m > 0$, $q_h > 0$, we have

$$\frac{k(q_m)}{k(q_h)} \leq \pi \ . \tag{A.7}$$

□

Proof of Theorem 2

Proof. Consider a relaxed problem where $d_t \in [-\bar{q}, w_t]$. Given HFT's best response , this problem can be reduced to a decision problem of the market maker. Suppose the policy proposed in this theorem is not optimal. Using the same argument as in the proof of theorem 1, this game is continuous at infinity. Thus, I can focus on considering a finite period deviation. Consider a better policy with deviation for at most n periods. In period n, I only need to consider the difference of dividends in period n and $n + 1$. If $d_n \neq w_n - q_m$, by Proposition 1, the market maker's optimal strategy is to set $x_m = \hat{x}_m(q_m)$ and get expected net profit $M(q_m)$. This is exactly the original policy. Suppose $d_n = w_n - \hat{w}$. Since the market maker's maximum

expected profit in period n is $M(\hat{w})$,

$$w_n - \hat{w} + \delta[M(\hat{w}) + (\hat{w} - q_m)] > w_n - q_m + \delta M(q_m) \ .$$

This implies

$$\frac{\delta}{1-\delta} M(\hat{w}) - \hat{w} > \frac{\delta}{1-\delta} M(q_m) - q_m \ .$$

Since $q_m = argmax_{w \in [0,\bar{q}]} \frac{\delta}{1-\delta} M(w) + (w_0 - w)$, if such $\hat{w}$ exists, it must be $\hat{w} > \bar{q}$. Since $M(w) = max((1 - F(x^*))x^*[k(w + q_h) - k(q_h)], (1 - F(\underline{x}(w))\underline{x}(w))k(w))$ is continuous and differentiable almost everywhere. Easy to see that $\frac{\delta}{1-\delta} M^{'}(w) \leq 1$ for $w \geq \bar{q}$. Thus, if $\hat{w} > \bar{q}$,

$$\frac{\delta}{1-\delta} M(\bar{q}) - \bar{q} \geq \frac{\delta}{1-\delta} M(\hat{w}) - \hat{w} \ .$$

This is because

$$\frac{\delta}{1-\delta} M(\hat{w}) = \frac{\delta}{1-\delta} M(\bar{q}) + \int_{\bar{q}}^{\hat{w}} \frac{\delta}{1-\delta} M^{'}(x) dx \ .$$

This implies that any n period deviation can be dominated by a $n-1$ period deviation for all n. Repeating this argument implies that no finite period deviation exists and establishes the optimality of the proposed policy. Since $w_0 > \bar{q}$, the proposed policy is implementable in the original problem and is thus optimal. The HFT's optimality condition is satisfied since the HFT always plays the best response. □

Proof of Corollary 2

Proof. Two conditions are derived from the first order condition of $max_{w\in[0,\bar{q}]}\frac{\delta}{1-\delta}M(w)+(w_0-w)$. To see the market maker never fully exit the market, notice that $\underline{x}\to x^*$ when $x_m\to 0$. Since $\bar{q}>0$, $\frac{\delta}{1-\delta}(1-F(x^*))x^*>1$. Then there always exists a $q_m>0$ such that $\frac{\delta}{1-\delta}(1-F(x^*))x^*(1-G(q_m))=1$. □

Proof of Theorem 3

Proof. For the ease of notation, let q_m^π be the equilibrium capital commitment of the market maker when the HFT's entry probability is π. Let $\underline{x}(q)$ be the tight spread when the market maker's shareholding is q. Notice that $\underline{x}$ does not depend on π. Consider a sequential pricing game with $\pi=1$. If in the steady state equilibrium, the market maker uses the wide spread strategy with shareholding q_m^1, then by Theorem 2, for any $q\in[0,\bar{q}]$,

$$\frac{\delta}{1-\delta}(1-F(x^*))x^*[k(q_m^1+q_h)-k(q_h)]+(w_0-q_m^1)\geq\frac{\delta}{1-\delta}(1-F(\underline{x}(q)))\underline{x}(q)k(q)+(w_0-q)\ .$$

That is, adopting the wide spread strategy with shareholding q_m^1 is better than using the tight spread strategy at any level of shareholding. Then for $\pi<1$ and any q,

$$\begin{aligned}&\frac{\delta}{1-\delta}(1-F(x^*))x^*\{\pi[k(q_m^\pi+q_h)-k(q_h)]+(1-\pi)k(q_m^\pi)\}+(w_0-q_m^\pi)\\ >&\frac{\delta}{1-\delta}(1-F(x^*))x^*[k(q_m^1+q_h)-k(q_h)]+(w_0-q_m^1)\\ \geq&\frac{\delta}{1-\delta}(1-F(\underline{x}(q)))\underline{x}(q)k(q)+(w_0-q)\ .\end{aligned}$$

Thus, for $\pi < 1$, the market maker's equilibrium strategy must still be the wide spread strategy. This corresponds to the case where $\hat{\pi} = 1$.

If the market maker is using the tight spread strategy at a $\pi_1 < 1$, then for $\pi_2 > \pi_1$, by a similar argument with Proposition 2, the market maker would still use the tight spread strategy. Moreover, $q_m^{\pi_1} = q_m^{\pi_2}$ and thus $\underline{x}(q_m^{\pi_1}) = \underline{x}(q_m^{\pi_2})$ and the market maker has the same equilibrium payoff. Denote this equilibrium payoff when the market maker is using a tight spread strategy by V^{tight}. Define

$$V_\pi^{wide} = \frac{\delta}{1-\delta}(1 - F(x^*))x^*[\pi(k(q + q_h) - k(q_h)) + (1 - \pi)k(q)] + (w_0 - q)$$

where q satisfies

$$\frac{\delta}{1-\delta}(1 - F(x^*))x^*[1 - \pi G(q + q_h) - (1 - \pi)G(q)] = 1 \ .$$

V_π^{wide} is the equilibrium payoff for the market maker if the wide spread strategy is adopted in the equilibrium. V_π^{wide} is continuous and decreasing with respect to π. Moreover, V_0^{wide} goes to the monopolistic payoff. Since $V^{tight} > V_1^{wide}$ and V^{tight} is bounded away from the monopolistic payoff, there exist $\hat{\pi} \in (0, 1)$ such that $V_{\hat{\pi}}^{wide} = V^{tight}$. By previous argument, the market maker adopts the wide spread strategy if $\pi < \hat{\pi}$ and tight spread strategy if $\pi > \hat{\pi}$.

In the tight spread region,

$$L = (1 - F(x_m))k(q_m) + \pi(F(x^*) - F(x_m))(k(q_m + q_h) - k(q_m)) \ .$$

Since in the tight spread region, q_m and $x_m = \underline{x}(q_m)$ is not changing with respect to π, L is increasing in π.

For the third statement, consider a game at $\pi = \hat{\pi} < 1$. Two equilibrium shareholdings for market maker, w_m^{tight} and w_m^{wide} both exist. If the market maker chooses shareholding q_m^{tight} (q_m^{wide}), he will play the tight (wide) spread strategy in the equilibrium. By Proposition 1, $k(q_m^{wide}) \geq \hat{\pi} k(q_h) \geq k(q_m^{tight})$. This implies $q_m^{wide} \geq q_m^{tight}$. For $\pi < \hat{\pi}$, $q_m > q_m^{wide} \geq q_m^{tight}$. This establishes that the market maker always have a higher equilibrium shareholding in the wide spread region. □

Proof of Proposition 3

Proof. Notice that in the wide spread region, L is continuous in π. Moreover, if the wide spread region is $[0, 1]$, liquidity is the same at $\pi = 0$ and $\pi = 1$. These two observations imply the proposition. □

Proof of Proposition 4

Proof. For any $w \geq 0$, given G is an exponential distribution, $k(s) = E_G(min(q, s)) = E_G(q)G(s)$. By theorem 1, when no HFT exists, the market maker's capital commitment $\bar{q}$ satisfies $\frac{\delta}{1-\delta}(1 - F(x^*))x^*(1 - G(\bar{q})) = 1$. By corollary 2, when the market maker posts a wide spread in the equilibrium, his capital commitment satisfies $\frac{\delta}{1-\delta}(1 - F(x^*))x^*[(1 - \pi)(1 - G(q_m)) + \pi(1 - G(q_m + q_h))] = 1$. Thus, $G(\bar{q}) = \pi G(q_m + q_h) + (1 - \pi)G(q_m)$.

Then,

$$\begin{aligned} k(\bar{q}) &= E_G(q)G(\bar{q}) \\ &= E_G(q)(\pi G(q_m + q_h) + (1-\pi)G(q_m)) \\ &= \pi k(q_m + q_h) + (1-\pi)k(q_m) \ . \end{aligned} \tag{A.8}$$

This implies that liquidity does not depend on π in the wide spread region and is equal to the liquidity in a monopolistic market. □

Proof of Theorem 4

Proof. Let's consider the first statement. Since I take other parameters as fixed and only change π, I represent liquidity by $L(\pi)$ and the market maker's capital commitment by $q_m(\pi)$ to make their dependences on π explicit while suppressing all other dependences.

As $\pi \to 0$, the market maker's payoff by posting the wide spread converges to the monopolistic payoff. By continuity of the market maker's payoff, for π small enough, the market maker would post a wide spread in the steady state equilibrium. In the wide spread region, the market maker's capital commitment $q_m(\pi)$ satisfies

$$\frac{\delta}{1-\delta}(1-F(x^*))x^*[(1-\pi)(1-G(q_m(\pi))) + \pi(1-G(q_m(\pi)+q_h))] = 1 \ . \tag{A.9}$$

Take derivative with respect to π,

$$G(q_m(\pi)) - G(q_m(\pi)+q_h) - \pi g(q_m(\pi)+q_h)q_m^{'}(\pi) - (1-\pi)g(q_m(\pi))q_m^{'}(\pi) = 0 \ . \tag{A.10}$$

Collecting terms to get

$$q_m^{'}(\pi) = \frac{G(q_m(\pi)) - G(q_m(\pi) + q_h)}{\pi g(q_m(\pi) + q_h) + (1-\pi) g(q_m(\pi))} \ . \tag{A.11}$$

In the wide spread region, $L(\pi) = (1 - F(x^*))[(1-\pi)k(q_m(\pi)) + \pi k(q_m(\pi) + q_h)]$. Then

$$\begin{aligned} \frac{1}{1-F(x^*)} L^{'}(\pi) =& k(q_m(\pi) + q_h) - k(q_m(\pi)) + \pi(1 - G(q_m(\pi) + q_h)) q_m^{'}(\pi) \\ &+ (1-\pi)(1 - G(q_m(\pi))) q_m^{'}(\pi) \ . \end{aligned} \tag{A.12}$$

Easy to see this function is continuous in π. Consider $L^{'}(\pi)$ at $\pi = 0$. Since $q_m(0) = \bar{q}$,

$$\frac{1}{1-F(x^*)} L^{'}(0) = k(\bar{q} + q_h) - k(\bar{q}) + (1 - G(\bar{q})) \frac{G(\bar{q}) - G(\bar{q} + q_h)}{g(\bar{q})} \ . \tag{A.13}$$

$L^{'}(0) < 0$ if and only if

$$\frac{G(\bar{q} + q_h) - G(\bar{q})}{k(\bar{q} + q_h) - k(\bar{q})} > \frac{g(\bar{q})}{1 - G(\bar{q})} \ . \tag{A.14}$$

Use integration by parts,

$$\begin{aligned} k(s) &= s(1-G(s)) + \int_0^s qg(q)dq \\ &= s(1-G(s)) + sG(s) - \int_0^s G(q)dq \\ &= s - \int_0^s G(q)dq \\ &= \int_0^s (1-G(q))dq \ . \end{aligned} \tag{A.15}$$

Thus, $L^{'}(0) < 0$ if and only if

$$\frac{\int_{\bar{q}}^{\bar{q}+q_h} g(q)dq}{\int_{\bar{q}}^{\bar{q}+q_h}(1-G(q))dq} > \frac{g(\bar{q})}{1-G(\bar{q})} \ . \tag{A.16}$$

Let

$$I(x) = \int_{\bar{q}}^{\bar{q}+x} g(q)dq - \frac{g(\bar{q})}{1-G(\bar{q})}\int_{\bar{q}}^{\bar{q}+x}(1-G(q))dq \ .$$

Inequality (A.16) holds if and only if $I(q_h) > 0$. Notice that $I(0) = 0$. Moreover,

$$I^{'}(x) = g(\bar{q}+x) - \frac{g(\bar{q})}{1-G(\bar{q})}(1-G(\bar{q}+x)) \ .$$

Since $\frac{g(x)}{1-G(x)}$ is increasing, for $x > 0$, $I^{'}(x) > 0$. Thus, $I(q_h) > 0$ and $L^{'}(0) < 0$. Then by continuity of $L^{'}(\pi)$, there exists a small region around 0 such that liquidity is decreasing in π.

Notice that the calculation above works for the situation when $\bar{q} + q_h$ is in the support of G. If $\bar{q}+q_h$ is not in the support of G, replace $\bar{q}+q_h$ with the upper-bound of G's support yields the same result.

For the increasing part, it is suffice to consider the situation where $\pi = 1$ is in the tight spread region. Since liquidity is increasing with π in the tight spread region, there exists $\tilde{\pi}$ such that liquidity is increasing for $\pi \in [\tilde{\pi}, 1]$. This finish the proof of the first statement.

Now I consider the second statement. Fix $\pi = 1$. Notice that for any fixed $q_m > 0$,

$$a(\underline{x}) = \frac{k(q_m + q_h) - k(q_m)}{k(q_h)} \to 1 - G(q_m) < 1 \text{ as } q_h \to 0 \ .$$

This implies that the market maker's payoff by using the tight spread strategy is bounded away from the monopolistic payoff as $q_h \to 0$. On the other hand, if the market maker uses the wide spread strategy, easy to see as $q_h \to 0$, the expected payoff converges to the monopolistic payoff. Thus, for small enough q_h, the market maker would use the wide spread strategy at the steady state even when $\pi = 1$. This finish the proof of the second statement. □

A.1.4 Simultaneous Pricing

Proof of Proposition 5

Proposition 5 can be divided into following claims.

Claim 1. *Players never propose spreads greater than x^*.*

Proof. If a player propose a spread greater than x^*, regardless of the other player's strategy, switching to proposing x^* yields a strictly larger payoff. □

Claim 2. *Neither players would use pure strategies in an equilibrium.*

Proof. Suppose the market maker posts spread $x_m = x$ in a equilibrium. The HFT's optimal strategy would be posting $x_h = x^*$, $x_h = x$ or a mix between these two price. Then the market maker would achieve higher payoff by undercutting the HFT's lowest possible price for a small enough ϵ. Contradiction

Suppose the HFT post spread $x_h = x$ in an equilibrium. Then in an equilibrium the market maker can only post x^*. (Undercutting will lead to no equilibrium because the payoff of the market maker is not continuous at x.) This implies $x_h \neq x^*$ in the equilibrium. However, if $x_h < x^*$, given the market maker is posting $x_m = x^*$, the HFT would be better off posting $x_h = x^*$. Contradiction. □

Suppose there exists a mixed strategy equilibrium. Denote the infimum and supremum of the spread posted by the market maker (HFT) by $\underline{x}_m(\underline{x}_h)$ and $\bar{x}_m(\bar{x}_h)$.

Claim 3. *$\underline{x}_m = \underline{x}_h$ and neither the market maker nor the HFT would post this spread with positive probability in an equilibrium.*

Proof. If $\underline{x}_m \neq \underline{x}_h$, the player with smaller spread lower-bound could raise the lower-bound by a small enough amount to achieve a higher payoff. Denote this common lower-bound by $\underline{x}$. If the HFT posts this spread with positive probability, rather than posting $\underline{x}$, the market maker would be strictly better off undercutting the HFT for a small amount.

Suppose the market maker posts $\underline{x}$ with positive probability. Let $B(x, r)$ be a open ball centered at x with radius r. First note that $\forall \epsilon > 0$, $\exists x_h \in B(\underline{x}, \epsilon)$ such that x_h is in HFT's mixed strategy's support. If not, since $\underline{x}$ is posted by the HFT with zero probability, the market maker can increase $\underline{x}_m$ by ϵ to achieve higher payoff.

Then for small enough ϵ, HFT's profit of posting $\underline{x}+\epsilon$ is strictly smaller than posting $\underline{x}$. Contradiction. □

Claim 4. *(No Holes)* $\nexists a, b \in (\underline{x}, \bar{x}_m)$, $a < b$ *such that* $(a,b) \bigcap X_m = \emptyset$. *A similar claim holds for* X_h.

Proof. Suppose this claim is false. Without loss of generality, let (a,b) be the maximum interval satisfying the claimed property. That is, $(a,b) \bigcap X_m = \emptyset$ and for any $a' < a$ and $b' > b$, $(a',b) \bigcap X_m \neq \emptyset$ and $(a,b') \bigcap X_m \neq \emptyset$.

By claim 1, $\bar{x}_m, \bar{x}_h \leq x^*$. Notice that if $(a,b) \notin X_m$, then $(a,b) \notin X_h$. This is because if $x \in (a,b)$ and $x \in X_h$, the HFT may increase x by a small amount to increase her payoff.

Then notice that $a \notin X_m$. This is because posting $x_m \in (a,b)$ will achieve a higher payoff given $(a,b) \notin X_h$. Moreover, $a \notin X_h$ by a similar argument.

Given that spread a is not posted by the HFT and the market maker with positive probability, when $x_m \to a$ from below, the payoff goes to the payoff of posting $x_m = a$ by continuity, which is smaller than posting $x_m \in (a,b)$. Since (a,b) is a maximum interval satisfying $(a,b) \bigcap X_m = \emptyset$, $\forall \epsilon > 0$, $B(a,\epsilon) \bigcap X_m \neq \emptyset$. This contradicts the equilibrium definition that $x_m \in X_m$ is a best response to the HFT's pricing strategy. □

Claim 5. $\bar{x}_m = \bar{x}_h = x^*$.

Proof. Suppose that $\bar{x}_m < \bar{x}_h$. Then $(\bar{x}_m, \bar{x}_h) \bigcap X_h = \emptyset$ since posting $x_h = \bar{x}_h$ yields a higher payoff. This contradicts Claim 4. Similarly, it is impossible that $\bar{x}_m > \bar{x}_h$. If $\bar{x}_m = \bar{x}_h < x^*$, $\bar{x}_m \notin X_m$ since $x_m = x^*$ would yield higher payoff. Since $\bar{x}_m \notin X_m$, by

the same argument, $\bar{x}_h \notin X_h$. However, then by the continuity argument, for small enough ϵ, $x_m \in B(\bar{x}_m, \epsilon)$ will be dominated by posting $x_m = x^*$. Contradiction. □

Claim 6. *For all $x \in (\underline{x}, x^*) \bigcap X_m$, x is not proposed by the market maker with positive probability in an equilibrium. For all $x \in (\underline{x}, x^*) \bigcap X_h$, x is not proposed by the HFT with positive probability in an equilibrium.*

Proof. We prove by contradiction. Suppose that the market maker posts spread x with positive probability. Then by claim 4, $\forall \epsilon > 0$, $B(x + \epsilon, \epsilon) \bigcap X_h \neq \emptyset$. However, by continuity, when ϵ is small, the payoff posting that spread is dominated by posting x. Contradiction. If the HFT posts spread x, note that the market maker's profit when posting a spread approaching x from the left is larger than the profit when posting a spread approaching x from the right. This leads to a contradiction. □

Proof of theorem 5

Proof. The proof of the first part is the same as the proof of Theorem 2. For the second and the third statement, note that expected payoffs of the market maker are the same in all one-shot games. Thus, in the equilibrium the market maker commits the same amount of capital to the market. The HFT's payoffs can be calculated from the corresponding one-shot game. □

Proof of proposition 7

Proof. For the first statement, notice that

$$L_{se} = (1 - F(x^*))[\pi k(q_m + q_h) + (1 - \pi)k(q_m)] \ .$$

Compare this to L_{sim} in Theorem 5 to reach the conclusion.

Notice that I have shown that L_{se} is increasing in π. Thus, the third statement is merely a corollary of the second statement. If π is in the tight spread region, in equilibrium, $k(q_m) \leq \pi k(q_h)$ and $a(\underline{x})$ is not changing with π. Moreover, q_m also remains constant with respect to π. Then by the market maker's indifference condition, for all $x \in (\underline{x}, x^*)$,

$$a(x)\{(1 - \pi)k(q_m) + \pi[H_h(x)(k(q_m + q_h) - k(q_h)) + (1 - H_h(x))k(q_m)]\} \quad \text{(A.17)}$$

is constant for all π in the tight spread region. This implies for any given x, $\pi H_h(x)$ is constant for all π in the tight spread region. This together with Theorem 5 implies that $L_{sim} - L_{se}$ is constant. It also implies that in the tight spread region, increase in π only benefits buyers with buying thresholds higher than $1 + x^*$. □

A.2 Extension: Costly Entry

A.2.1 Sequential Pricing

Proof of Proposition 8

Proof. If $C \geq \bar{C} = \pi(1 - F(x^*))x^*k(q_h)$, the expected return of the HFT cannot cover the cost even when the HFT undercuts the market maker at spread x^*. Thus, the HFT will not enter the market regardless of the market maker's spread. In equilibrium, the market maker would choose $x_m = x^*$.

Now consider the situation where $C < \bar{C}$. In this case, if the market maker posts the wide spread x^*, the HFT would attempt to enter the market and undercut the market maker upon entry. Moreover, the HFT would not choose to enter and undercut the market maker if the market maker posts the deterring spread x satisfying $\pi(1-F(x))xk(q_h) = C$. If the market maker posts a spread higher than the deterring spread x, the HFT will always enter since she can always undercut the market maker and earn a expected payoff higher than C.

If $k(q_m) < \pi k(q_h)$, given the HFT chooses to enter the market, the market maker's optimal spread is the tight spread satisfying $(1-F(x))xk(q_h) = (1-F(x^*))x^*[k(q_m+q_h)-k(q_m)]$. Moreover, as long as the HFT does not undercut the market maker, the market maker always prefers to set the spread x_m higher (given $x_m \leq x^*$). Thus, in equilibrium, the market maker will compare the tight spread and the deterring spread and pick the greater one. Specifically, if $C > \pi(1 - F(x^*))x^*[k(q_m + q_h) - k(q_m)]$, posting the deterring spread is more profitable. Otherwise, posting the tight spread

is more profitable. Furthermore, when facing the tight spread, the HFT is indifferent between posting the monopolistic spread and undercutting the market maker. Then when the market maker posts the deterring spread, upon entering, the HFT is better off undercutting the market maker. This implies that when the market maker posts the deterring spread, the HFT will choose not to try to enter the market. The discussion for $k(q_m) > \pi k(q_h)$ follows the similar logic and is thus omitted. □

Proof of Theorem 6

Proof. Let $\underline{x}^d$ satisfies $\pi(1 - F(\underline{x}^d))\underline{x}^d k(q_h) = C$ for $C \in [0, \bar{C}]$. Let q_m^d satisfies $\frac{\delta}{1-\delta}(1 - F(\underline{x}^d))\underline{x}^d(1 - G(q_m^d)) = 1$. This is the equilibrium capital commitment if the market maker uses a deterring entry strategy. The equilibrium payoff is $V_C(w_0) = \frac{\delta}{1-\delta}(1 - F(\underline{x}^d))\underline{x}^d k(q_m^d) + (w_0 - q_m^d)$. Easy to see that this quantity is increasing in C. Easy to see that when $C \geq \bar{C}$, this quantity becomes monopolistic payoff. Let the market maker's equilibrium payoff when $C = 0$ be $V_0(w_0)$. There exist a unique $\hat{C}$ such that $V_{\hat{C}}(w_0) = V_0(w_0)$. Thus, for $C > \hat{C}$, the market maker is using the deterring strategy in the equilibrium.

When the market maker is using the deterring strategy, suppose the HFT chooses to participate, then she optimally set $x_h = x^*$. Since (1) the HFT is not undercutting the market maker, and (2) when the HFT participates, her optimal pricing strategy does not depend on C, when $C = 0$, the market maker can use the same equilibrium strategy to achieve a higher expected payoff. Contradiction. Thus, the HFT does not choose to participate. □

A.2.2 Simultaneous Pricing

Proof of Proposition 9

Proof. First consider the case where $C > \bar{C}$. In this case, the HFT's expect profit can never cover the cost regardless of the market maker's pricing strategy. Thus, $\eta = 0$ and the market maker sets $x_m = x^*$.

Now consider the situation when $C \in [0, \bar{C}]$. Suppose the HFT chooses $\eta = 1$ and plays a mixed pricing strategy as in a game $(q_m, q_h, \pi, 0)$. By Proposition 6, the HFT's expected profit is $\pi(1 - F(x^*))x^* a(\underline{x})(\pi)k(q_h)$. If $\pi(1 - F(x^*))x^* a(\underline{x})(\pi)k(q_h) \geq C$, since C is paid at the end of the period, the equilibrium characterized by Proposition 6 still holds.

If $\pi(1 - F(x^*))x^* a(\underline{x})(\pi)k(q_h) < C < \bar{C}$, note that $\eta \neq 0$ in the equilibrium. This is because if $\eta = 0$, the market maker would post $x_m = x^*$. The HFT has incentive to deviate to $\eta = 1$. Thus, I need to consider an equilibrium where the HFT mixes between participating. In other words, $\eta \in (0, 1)$. η can be pinned down by the indifference condition that the HFT earns zero profit when trying to enter the market.

First consider the situation $k(q_m) \geq \pi k(q_h)$. By Proposition 6, if the HFT tries to enter with probability η, $\underline{x}$ is determined by

$$(1 - \eta\pi)k(q_m) + \eta\pi(k(q_m + q_h) - k(q_h)) = a(\underline{x})k(q_m) \ . \tag{A.18}$$

Notice that $\underline{x}$ is decreasing in η and $\underline{x} \to x^*$ as $\eta \to 0$. Thus, there exist a unique

$\eta \in (0,1)$ such that $\eta\pi(1-F(x^*))x^*a(\underline{x})(\eta\pi)k(q_h) = \eta C$ where $\underline{x}$ is the lower-bound of the mixed strategy in the game $(q_m, q_h, \eta\pi, 0)$. If the HFT participates with probability η and posts spread according to H_h in the game $(q_m, q_h, \eta\pi, 0)$, the market maker has no incentive to deviate from posting spread according to H_m in the game $(q_m, q_h, \eta\pi, 0)$. If the market maker sets price according to H_m, upon entering, the HFT has no incentive to deviate from posting spread according to H_h. Moreover, the HFT earns zero expected profit for trying to enter. Thus, the HFT has no incentive to deviate from η.

Next consider the situation $k(q_m) < \pi k(q_h)$. Notice that $\underline{x}$ remains constant in this region. Let $\bar{\eta}$ satisfies $k(q_m) = \bar{\eta}\pi k(q_h)$. By the same argument, there exists a unique $\eta \in (0, \bar{\eta})$ such that $k(q_m) > \pi\eta k(q_h)$ and $\pi(1-F(x^*))x^*a(\underline{x})(\pi)k(q_h) = C$ where $\underline{x}$ is the lower-bound of the mixed strategy in the game $(q_m, q_h, \eta\pi, 0)$. The rest of the verification is the same. □

Proof of Corollary 4

Proof. This proof essentially involves only comparing the market maker's payoffs under two settings with different parameter values. Fix a game (q_m, q_h, π, C). First consider the case when $k(q_m) \geq \pi k(q_h)$. In the one-shot simultaneous pricing game,

$$a(\underline{x})(\pi) = 1 - \pi + \frac{k(q_m + q_h) - k(q_h)}{k(q_m)}\pi \ .$$

If $\pi(1-F(x^*))x^*a(\underline{x})(\pi)k(q_h) \geq C$, by Proposition 9, in the simultaneous pricing game, the HFT participates in high-frequency with probability 1. The market

maker enjoys the same expected payoff as in the simultaneous pricing one-shot game $(q_m, q_h, \pi, 0)$, which equals to $(1-\pi)k(q_m) + \pi(k(q_m+q_h) - k(q_h))$. By Proposition 8, the market maker receives the same expected payoff in the sequential pricing one-shot game . For $\pi(1-F(x^*))x^*a(\underline{x})(\pi)k(q_h) < C$, in the simultaneous pricing game, the market maker receives payoff $(1-F(x^*))x^*a(\underline{x})(\eta\pi)k(q_m)$ by the indifference condition where

$$a(\underline{x})(\eta\pi) = \frac{C}{\pi(1-F(x^*))x^*k(q_h)} \ .$$

Thus, the market maker's expected payoff is $\frac{C}{\pi k(q_h)}k(q_m)$, which equals to the expected payoff in a one-shot sequential pricing game by Proposition 8..

Next consider the case when $k(q_m) < \pi k(q_h)$. In a one-shot simultaneous pricing game,

$$a(\underline{x})(\pi) = \frac{k(q_m+q_h) - k(q_m)}{k(q_h)} \ .$$

If $\pi(1-F(x^*))x^*a(\underline{x})(\pi)k(q_h) = \pi(1-F(x^*))x^*(k(q_m+q_h)-k(q_m)) \geq C$, in a simultaneous pricing game, the market maker's expected payoff is $(1-F(x^*)x^*)a(\underline{x})k(q_m)$. This is the same as the expected payoff in a sequential pricing game. If $\pi(1-F(x^*))x^*a(\underline{x})(\pi)k(q_h) < C$, in a simultaneous pricing game, the market maker receives payoff $(1-F(x^*))x^*a(\underline{x})(\eta\pi)k(q_m)$ where

$$a(\underline{x})(\eta\pi) = \frac{C}{\pi(1-F(x^*))x^*k(q_h)} \ .$$

Thus, the market maker's expected payoff is $\frac{C}{\pi k(q_h)}k(q_m)$, which equals to the expected payoff in the sequential pricing game. □

A.3 Flipping

A.3.1 Proof of Proposition 11

Proof. Notice that

$$r^{'}(q_f) = (1 - F(x^*))x^*(1 - G(q_m - q_f)) - x_m \ ,$$

$$r^{''}(q_f) = (1 - F(x^*))x^* g(q_m - q_f) > 0 \ .$$

This implies the maximum is achieved at the boundary $q_f = 0$ or $q_f = q_m$. □

A.3.2 Proof of Lemma 3

Proof. Inequality (2.13) guarantees that the HFT is better off with $q_f = q_m$ than with $q_f = 0$ when setting $x_h = x^*$. Inequality (2.14) guarantees that the HFT is better off with $w = q_m$ than undercutting the market maker. Since the market maker is better off choosing the highest possible spread given the HFT is flipping orders, one of the inequalities must be binding.

Moreover, if inequality (2.13) binds,

$$x_m^f = \frac{(1 - F(x^*))x^* k(q_m)}{q_m} \ .$$

Otherwise, since $(1 - F(x))x$ is increasing in $[0, x^*]$, there exists a unique

$$x_m^f \in (0, \frac{(1 - F(x^*))x^* k(q_m)}{q_m})$$

such that

$$(1 - F(x^*))x^* k(q_m + q_h) = x_m^f q_m + (1 - F(x_m^f))x_m^f k(q_h) \ .$$

If inequality (2.14) binds,

$$(1 - F(x^*))x^* k(q_m) > x_m^f q_m \tag{A.19}$$

and

$$(1 - F(x^*))x^* k(q_m + q_h) = x_m^f q_m + (1 - F(x_m^f))x_m^f k(q_h) \ . \tag{A.20}$$

Then,

$$(1 - F(x^*))x^*[k(q_m + q_h) - k(q_m)] < (1 - F(x_m^f))x_m^f k(q_h) \ . \tag{A.21}$$

Thus,

$$x_m^f > \underline{x} \ . \tag{A.22}$$

In this case, the tight spread strategy is never optimal because the market maker can raise the spread to x_m^f to achieve higher expected payoff. □

A.3.3 Proof of Proposition 12

Proof. Consider the situation when $\pi = 1$. If inequality (2.13) is binding, the market maker's expected payoff with flipping is

$$x_m^f q_m = (1 - F(x^*))x^* k(q_m) \ .$$

This is the highest possible payoff. If inequality (2.14) is binding, by Lemma 3, the tight spread strategy is dominated. Moreover,

$$x_m^f q_m = (1-F(x^*))x^* k(q_m+q_h) - (1-F(x_m^f))x_m^f k(q_h) > (1-F(x^*))x^*(k(q_m+q_h)-k(q_h)) \ .$$

Thus, setting $x_m = x_m^f$ is better than setting $x_m = x^*$. By continuity, for π large enough, it is always optimal to induce flipping. □

A.3.4 Proof of Proposition 13

Proof. The proof is omitted since it is similar to the existence result proved in previous sections. □

A.4 Extension: Supply Schedule and Induced Limit Order Book

A.4.1 Proof of Proposition 14

Proof. Obviously, it is not optimal for the market maker to sell any share at a spread higher than x^*. Then without loss of generality, I only consider the situation where the market maker set spreads lower than x^*. The proof consists of two steps. I first show that if the market maker can supply shares with n spreads $x_1, ..., x_n$ with $\sum_{i=1}^{n} q_i = q_m$, then he should optimally set $x_1 = ... = x_n = x^*$. Then I show that the market maker's payoff under any supply schedule $\Psi(x)$ can be approximated with arbitrary precision by a n-spreads supply plan with a large enough n.

Consider the situation when $n = N$. Without loss of generality, suppose $x_1 \leq x_2 \leq ... \leq x_N \leq x^*$. Define $q_0 = 0$. The market maker's expected payoff is

$$\sum_{i=1}^{N}(1 - F(x_i))x_i[k(\sum_{j=0}^{i} q_j) - k(\sum_{r=0}^{i-1} q_r)] \ .$$

Note that the market maker can increase his expected payoff by setting $x_1 = x_2$ since $(1 - F(x))x$ is increasing in $x \in [0, x^*]$. This reduce the problem to $n = N - 1$ situation. By induction, for arbitrary n, $x_1 = ... = x_n = x^*$ is the optimal supply schedule.

Next consider the approximation procedure under arbitrarily fixed q_m. For arbitrary $\Psi(x)$, divide its support into n intervals $\{I_1, ..., I_n\}$. The I_i interval is from $\frac{i-1}{n}th$ quantile to $\frac{i}{n}th$ quantile. Consider a new supply schedule that supply shares at

n spreads. Specifically, in the new schedule, the market maker supplies q_i shares at spread x_i for $i = 1, ..., n$. Let $x_i = E_\Psi(x|x \in I_i)$; $q_i = \frac{q_m}{n}$ for all i. Under any buyer's demand and buying threshold, realized profits of this new schedule and schedule Ψ differ by at most a factor of $\frac{q_m}{n}$, which goes to 0 as $n \to \infty$. Thus, expected profit from any supply schedule Ψ can be approximated to an arbitrarily close level by a schedule with n spreads when n is large enough. This establish the fact that the optimal supply schedule is to sell all shares at the spread x^*. □

A.4.2 Proof of Corollary 5

Proof. The first statement is a straightforward result from Proposition 14. For the second statement, if the dividend payout is non-zero, the market maker can always achieve a higher payoff by refraining from paying dividend and supply the extra amount of shares at the spread x^* and payout the total return from the extra shares in the next period. □

A.4.3 Proof of Proposition 15

Proof. From the analysis of the baseline model, any single spread pricing strategy is dominated either by the wide spread strategy or the tight spread strategy. Thus, I only need to show that, when the market maker can submit a supply curve, using the wide spread strategy or the tight spread strategy is not optimal.

Suppose for some π and q_h there exists a steady state equilibrium with capital commitment q_m and supply schedule $\Psi(x) = I_{\{x \geq x^*\}}$. Then upon entering the market,

the HFT would set spread $x_h = x^*$. The market maker's expected dividend payout each period would be

$$\pi(1 - F(x^*))x^*[k(q_m + q_h) - k(q_h)] + (1 - \pi)(1 - F(x^*))x^*k(q_m) \ .$$

Consider a deviation of the market maker by selling ϵ shares at the spread x_ϵ satisfying

$$(1 - F(x^*))x^*(k(\epsilon + q_h) - k(\epsilon)) = (1 - F(x_\epsilon))x_\epsilon k(q_h)$$

and $q_m - \epsilon$ shares at the spread x^*. Then the HFT would still set spread $x_h = x^*$ and the market maker's expected dividend payout would be

$$\begin{aligned} div(\epsilon) = (1 - F(x_\epsilon))x_\epsilon k(\epsilon) + \pi(1 - F(x^*))x^*[k(q_m + q_h) - k(q_h \\ + \epsilon)] + (1 - \pi)(1 - F(x^*))x^*[k(q_m) - k(\epsilon)] \ . \end{aligned}$$

Easy to check

$$div^{'}(0) = (1 - F(x^*))x^*\pi G(q_h) > 0 \ .$$

Thus, the market maker can deviate in pricing to achieve a higher expected payoff. Contradiction.

For the tight spread strategy, a similar argument can show that the market maker can achieve higher expected payoff by increasing the spread of a small amount of shares. This completes the proof. □

A.4.4 Proof of Proposition 16

Lemma 9. *In any steady state equilibrium, the HFT set $x_h = x^*$.*

Proof. Suppose not, then $\lim_{x \to x_h^-} \Psi(x) < 1$. The market maker would achieve a higher expected payoff by sell $q_m(\lim_{x \to x^{*-}} \Psi(x) - \lim_{x \to x_h^-} \Psi(x))$ shares at the spread x^*. □

Proof. **Proposition 16** First note that if $\Psi(x^*) < 1$, the market maker can become better off by selling all shares with spreads higher than x^* at spread x^*.

Suppose $\Psi(x)$ has a mass point at $x < x^*$. If the HFT is strictly prefers posting $x_h = x^*$, then there exists an ϵ such that the market maker can sell these shares at the spread $x + \epsilon$ to achieve higher payoff. If the HFT is indifferent, then there must exist an ϵ such that the HFT is strictly prefer setting $x_h = x^*$ than setting $x_h = x + \epsilon$. If the HFT is indifferent between posting x and x^*, since x is a mass point, there exists a $\epsilon > 0$ such that the HFT strictly prefers setting $x_h = x^*$ to setting $x_h \in (x, x + \epsilon)$. The market maker can then improve his pricing by selling all shares within the spread range $(x, x + \epsilon)$ and some shares at the spread x to the spread $x + \epsilon$.

The next step is to show that for any Ψ violating the indifference condition of the HFT, the market maker can always find a better pricing plan. Specifically, I consider this problem holding $q_m \Psi(x^{*-})$ and q_m constant. First notice that $\underline{x}$ can be uniquely pinned down by

$$(1 - F(x^*)x^*)k(q_h + q_m \Psi(x^{*-})) - k(q_m \Psi(x^{*-})) = (1 - F(\underline{x})\underline{x})k(q_h) \ .$$

Denote the pricing distribution satisfying the HFT's indifference condition by $\underline{\Psi}(x)$. Then for all $x \in [\underline{x}, x^*]$, $\Psi(x) \geq \underline{\Psi}(x)$. Otherwise the HFT will not set $x_h = x^*$ and the pricing distribution cannot be optimal at the steady state. Suppose $\Psi \neq \underline{\Psi}$, let $\acute{x} = inf_x\{\Psi(x) > \underline{\Psi}(x)\}$. Since $\Psi(x)$ does not have mass point, there exists $\xi > 0$ such that $\Psi(x) > \underline{\Psi}(x)$ for $x \in (\acute{x}, \acute{x} + \xi]$ and $\Psi(\acute{x} + \xi) > \Psi(\acute{x})$. Then by the same approximation and moving mass argument, the market maker is better off selling shares in the spread interval $(\acute{x}, \acute{x} + \xi)$ at the spread $\acute{x} + \xi$. □

A.5 Capital Commitment when G has Non-decreasing Hazard Rate

This section provides a detailed analysis of the market maker's capital commitment strategy when the buyer's demand G follows a distribution with increasing hazard rate. Particularly, under any fixed HFT shareholding q_h, the market maker has a unique optimal steady state tight spread strategy.

Proposition 34. *Let* $B = \frac{\delta}{1-\delta}(1 - F(x^*))x^* > 1$.[3] *If* G *has non-decreasing hazard rate,* $max_y B\frac{k(y+q_h)-k(y)}{k(q_h)}k(y) + (w_0 - y)$ *has a unique solution* $q_m \in [0, \bar{q}]$.

The connection between this proposition and the tight spread strategy is clear. Notice that $a(\underline{x}) = \frac{k(q_m+q_h)-k(q_m)}{k(q_h)}$. By posting spread x_m satisfying $(1 - F(x_m))x_m = (1 - F(x^*))x^* a(\underline{x})$, a short-run HFT with q_h shares has no incentive to undercut the market maker.

[3] If $B \leq 1$, the market maker would not make that market even as a monopolist.

Proof. The first order condition is

$$W^{'}(y) = \frac{B}{k(q_h)}[(1-G(y))(k(y+q_h)-k(y))-(G(y+q_h)-G(y))k(y)]-1 = 0 \ . \quad \text{(A.23)}$$

When $y = 0$, $W^{'}(0) = B - 1 > 0$. When $y \geq \bar{q}$, $W^{'}(y) < B(1 - G(\bar{q}))\frac{k(y+q_h)-k(y)}{k(q_h)} - 1$. Since $B(1 - G(\bar{q})) = 1$, $W^{'}(y) < 0$. By continuity, $W^{'}(y)$ cross zero at least once for $y \in [0, \bar{q}]$. If I can show that $W^{'}$ only cross zero once, then a unique maximizer exists.

Consider any q_m such that $W^{'}(q_m) = 0$. We have

$$(1-G(q_m))[k(q_m+q_h)-k(q_m)]-(G(q_m+q_h)-G(q_m))k(q_m) = k(q_h)(1-G(\bar{q})) > 0 \ . \quad \text{(A.24)}$$

Thus,

$$\frac{k(q_m + q_h) - k(q_m)}{k(q_m)} > \frac{G(q_m + q_h) - G(q_m)}{1 - G(q_m)} \ . \quad \text{(A.25)}$$

Next I show that $W^{''}(q_m) < 0$. $W^{''}(q_m) < 0$ is equivalent to

$$\begin{aligned} &g(q_m)[k(q_m + q_h) - k(q_m)] + k(q_m)[g(q_m + q_h) - g(q_m)] \\ &+2(1 - G(q_m))[G(q_m + q_h) - G(q_m)] > 0 \ . \end{aligned} \quad \text{(A.26)}$$

Since $G(q_m + q_h) - G(q_m) > 0$, a sufficient condition for inequality (A.26) is

$$g(q_m)[k(q_m + q_h) - k(q_m)] > k(q_m)[g(q_m) - g(q_m + q_h)] \ . \quad \text{(A.27)}$$

Since G has non-decreasing hazard rate, $\frac{g(q_m+q_h)}{1-G(q_m+q_h)} \geq \frac{g(q_m)}{1-G(q_m)}$. Thus, $g(q_m) - g(q_m +$

$q_h) \leq \frac{G(q_m+q_h)-G(q_m)}{1-G(q_m)} g(q_m)$. This implies inequality (A.25) is sufficient for inequality (A.27).

In sum, there exists a $q_m \in [0, \bar{q}]$ such that $W^{'}(q_m) = 0$. Moreover, for any q_m such that $W^{'}(q_m) = 0$, $W^{''}(q_m) < 0$. This implies that $W(y)$ has a unique maximum. □

Proposition 35. *Suppose G has non-decreasing hazard rate. Consider two simultaneous pricing games where the market maker has discount rate δ_1 in the first game and discount rate δ_2 in the second game. Suppose $\delta_1 > \delta_2$ and all other parameters are the same. Let q_m^1 (q_m^2) be the market maker's steady state capital commitment in the first game (the second game). Then $q_m^1 > q_m^2$.*

Proof. Let $B_1 = \frac{\delta_1}{1-\delta_1}(1 - F(x^*))x^*$; $B_2 = \frac{\delta_2}{1-\delta_2}(1 - F(x^*))x^*$. If the market maker is using the wide spread strategy in both games, then $q_m^1 > q_m^2$ directly follows from the first order condition. If the market maker is using the tight spread strategy in both games, then by the first order condition,

$$\frac{B_1}{k(q_h)}[(1 - G(q_m^1))(k(q_m^1 + q_h) - k(q_m^1)) - (G(q_m^1 + q_h) - G(q_m^1))k(q_m^1)] - 1 = 0 \ .$$

Since $B_1 > B_2$, we have

$$\frac{B_2}{k(q_h)}[(1 - G(q_m^1))(k(q_m^1 + q_h) - k(q_m^1)) - (G(q_m^1 + q_h) - G(q_m^1))k(q_m^1)] - 1 < 0 \ .$$

Then by Proposition 34, there exists a unique $q_m^2 < q_m^1$ such that

$$\frac{B_2}{k(q_h)}[(1 - G(q_m^2))(k(q_m^2 + q_h) - k(q_m^2)) - (G(q_m^2 + q_h) - G(q_m^2))k(q_m^2)] - 1 = 0 \ ,$$

and q_m^2 maximize the market maker's expected payoff given he is using a tight spread strategy in the steady state. If the market maker is using the wide spread strategy in the first game and the tight spread strategy in the second game, combine the result about with Theorem 3 yield the result that $q_m^1 > q_m^2$. This covers all situations when the market maker is using the wide spread strategy in the first game.

Now consider the situation where the market maker is using the tight spread strategy in the first game. Let q_t^1 and q_w^1 (q_t^2 and q_w^2) be the market maker's share-holding under the optimal tight and wide spread strategy in the first (second) game. Since the market maker is using the tight spread strategy in the first game, $q_m^1 = q_t^1$. By the discussion above, $q_t^1 > q_t^2$; $q_w^1 > q_w^2$. If $q_t^1 > q_w^2$, the claim is true. Thus, we only consider the case when $q_t^1 \leq q_w^2$.

From the optimality condition,

$$\frac{\delta_1}{1-\delta_1} M(q_t^1) + (w_0 - q_t^1) \geq \frac{\delta_1}{1-\delta_1} M(q_w^1) + (w_0 - q_w^1) > \frac{\delta_1}{1-\delta_1} M(q_w^2) + (w_0 - q_w^2) \ , \tag{A.28}$$

where $M(\cdot)$ is the expected profit of the market maker in a one-shot game. If $M(q_t^1) > M(q_w^2)$, since $q_t^1 \leq q_w^2$, we have

$$\frac{\delta_2}{1-\delta_2} M(q_t^2) + (w_0 - q_t^2) > \frac{\delta_2}{1-\delta_1} M(q_t^1) + (w_0 - q_t^2) > \frac{\delta_2}{1-\delta_2} M(q_w^2) + (w_0 - q_w^2) \ .$$

Thus, the market maker would use the tight spread strategy in the second game and $q_m^1 > q_m^2 = q_t^2$.

If $M(q_t^1) \leq M(q_w^2)$, from equation A.28 and $\frac{\delta_1}{1-\delta_1} > \frac{\delta_2}{1-\delta_2}$, we also have

$$\frac{\delta_2}{1-\delta_2} M(q_t^2) + (w_0 - q_t^2) > \frac{\delta_1}{1-\delta_2} M(q_t^1) + (w_0 - q_t^2) > \frac{\delta_2}{1-\delta_2} M(q_w^2) + (w_0 - q_w^2) \ .$$

This implies $q_m^1 > q_m^2 = q_t^2$ and concludes the proof. □

This result is important for the simultaneous pricing game extension. Notice that the equilibrium I construct in the simultaneous pricing game might not be sub-game perfect. In the sub-game where the market maker commits less capital than the steady state level, it might not be optimal for the market maker to stick to the strategy specified in the equilibrium since net worth may have additional benefit. However, if I assume G has non-decreasing hazard rate, this is not a problem since I can consider a game where the HFT is uncertain about the market maker's discount rate δ and infers it from the market maker's capital commitment decision. This result guarantees the existence of a separating equilibrium where in equilibrium, the market maker's discount rate is perfectly signaled by his capital commitment decision.[4] In this sense, the equilibrium I propose coincide with a perfect Bayesian equilibrium in this extended game.

Proposition 36. *Suppose G has non-decreasing hazard rate. If $q_h \geq \frac{\bar{q}}{2}$, $argmax W(y) \in [0, \frac{\bar{q}}{2}]$.*

Proof. By Proposition 34, if $W'(\frac{\bar{q}}{2}) \leq 0$, then $argmax_y W(y) \in [0, \frac{\bar{q}}{2}]$. Thus, it is sufficient to show that for all $q_h \geq \frac{\bar{q}}{2}$, $W'(\frac{\bar{q}}{2}) \leq 0$.

[4]When the market maker's capital commitment cannot be mapped to any δ, any off path belief can be specified. For example, the HFT may assume that the market maker is maximizing the short term profit.

This is equivalent to

$$k(q_h)(1-G(\bar{q}))+(G(\frac{\bar{q}}{2}+q_h)-G(\frac{\bar{q}}{2}))k(\frac{\bar{q}}{2})-[1-G(\frac{\bar{q}}{2})][k(\frac{\bar{q}}{2}+q_h)-k(\frac{\bar{q}}{2})] \geq 0 \ . \quad \text{(A.29)}$$

When $q_h = \frac{\bar{q}}{2}$, the LHS of inequality (A.29) becomes

$$(1 - G(\frac{\bar{q}}{2}))[2k(\frac{\bar{q}}{2}) - k(\bar{q})] \ . \quad \text{(A.30)}$$

This quantity is greater than zero since $2k(\frac{\bar{q}}{2}) > k(\bar{q})$. Denote the LHS of inequality (A.29) by $J(q_h)$. If $J(q_h)$ is increasing in q_h, the lemma is proved.

$$J^{'}(q_h) = (1 - G(q_h))(1 - G(\bar{q})) + g(\frac{\bar{q}}{2})k(\frac{\bar{q}}{2}) - [1 - G(\frac{\bar{q}}{2})](1 - G(\frac{\bar{q}}{2} + q_h)) \ . \quad \text{(A.31)}$$

A sufficient condition of $J^{'}(q_h) \geq 0$ is $\frac{1-G(\bar{q})}{1-G(\frac{\bar{q}}{2})} \geq \frac{1-G(\frac{\bar{q}}{2}+q_h)}{1-G(q_h)}$. Since $q_h \geq \frac{\bar{q}}{2}$, it is sufficient to have $\frac{1-G(\bar{q}+z)}{1-G(\frac{\bar{q}}{2}+z)}$ decreasing in z. Take derivative to get

$$- g(\bar{q} + z)(1 - G(\frac{\bar{q}}{2} + z)) + g(\frac{\bar{q}}{2} + z)(1 - G(\bar{q} + z)) \leq 0 \ . \quad \text{(A.32)}$$

This condition is satisfied due to the increasing hazard rate of G. □

A.6 Social Planner's Perspective on Welfare

In this section, I briefly discuss the social planner's perspective on welfare. In practice, the social planner can be either a policy maker or an exchange, aiming at

maximizing market participants' welfare. I assume that the social planner can control q_m, the market maker's capital commitment and π, the HFT's entry probability. In this case, the social planner provides the market maker operating capital q_m at period 0 to make the market and the market maker pays the profit from market making as dividend.[5] For simplicity, I focus on the situation where the HFT trades faster than the market maker.

If the social planner aims at maximizing liquidity, he has two possible policies. Either he relies on the market maker to supply liquidity by setting $q_m = \infty$ and $\pi = 0$.[6] In this market, $L = (1 - F(x^*))E_G(q_b) = (1 - F(x^*))k(\infty)$. Alternatively, the social planner can rely on both the market maker and the HFT to supply liquidity by setting $q_m = q_h$ and $\pi = 1$. In this market, $L = (1 - F(x^*))k(2q_h) + [F(x^*) - F(\underline{x})]k(q_h)$ where $\underline{x}$ is uniquely pinned down by $a(\underline{x})k(q_h) = k(2q_h) - k(q_h)$. The social planner would choose the policy that provides higher liquidity. Intuitively, when q_h is large, the social planner tends to use the latter policy. In either case, the social planner is committing more capital than the profit maximizing market maker.

[5]Alternatively, I may assume the social planner set a mandatory capital commitment level for the market maker. Two settings lead to similar qualitative results.

[6]π can be any number between 0 and 1 and market liquidity will be the same.

Appendix B

Appendix to Chapter 3

B.1 Model

In this section, we provide details on additional cases for the N-period model discussed in Section 3.4. We also provide proofs for the results in Section 3.4.

B.1.1 Benchmark Cases

In this subsection, we formally provide two benchmark cases which we refer to in the main text.

First, we present the equilibrium for the N period model of Section 3.4 but with perfectly competitive market makers, i.e., market makers that do not take their own price impact into account. In this equilibrium, prices must be constant over trading periods, otherwise there would be an opportunity for unlimited profits by buying when prices are low and selling when prices are high.

Second, we present the equilibrium in a dealer market. Specifically, market makers have the same preferences as in the model of Section 3.4 but instead trading occurs from the trader making take-it-or-leave-it offers to each of the I market makers, which the market makers can either accept or reject. Since there is no asymmetric information over any parameters in the model, the trader can make an offer where the market maker is just indifferent between trading and not trading.

The following propositions formally state the equilibria.

Proposition 37. *If market makers engage in perfect competition in that they do not take their own price impact into account, there exists an equilibrium such that* $\{z_t\}_{t=1}^{N}$ *is any sequence satisfying* $\sum z_t = z$*; for market maker* i*, the equilibrium demand schedules are* $x_{i,N} = \frac{\mu - p_N}{\rho\sigma^2} - \sum_{t=1}^{N-1} x_{i,t}$*;* $x_{i,t}$ *is any number for* $p_t = \mu - \rho\sigma^2 \frac{z}{I}$ *and* $x_t = \infty$ *for* $p_t < \mu - \rho\sigma^2 \frac{z}{I}$ *for* $t < N$ *(similarly,* $x_t = -\infty$ *if* $p_t > \mu - \rho\sigma^2 \frac{z}{I}$*). In a symmetric equilibrium,* $x_{i,t} = \frac{z_t}{I}$ *and* $p_t^* = \mu - \rho\sigma^2 \frac{z}{I}$ *for all* t*. The liquidation value in achieved by the trader in this market is* $V_c = \mu z - \rho\sigma^2 \frac{z^2}{I}$*.*

Proof. Notice that the equilibrium asset prices must be the same in all periods. If a market maker expects that the equilibrium asset price is higher in the future, he would have incentive to buy infinite amount of asset in the current period and sell infinite amount of asset in the future, or vice versa. Thus, we can focus on the equilibrium asset price in the last period without loss of generality. Equalizing market maker i's marginal value to the last period's price p_N yields his demand schedule $x_{i,N} = \frac{\mu - p_N}{\rho\sigma^2} - \sum_{t=1}^{N-1} x_{i,t}$. From the market clearing condition, $p_N^* = \mu - \rho\sigma^2 \frac{z}{I}$. Since equilibrium prices in all periods have to be the same, $p_t^* = \mu - \rho\sigma^2 \frac{z}{I}$ for all t. Since the equilibrium asset price remains constant over periods, the trader is indifference

among all supply schedules and the liquidation value is $V_c = p_t^* \cdot z = \mu z - \rho\sigma^2 \frac{z^2}{I}$.

Proposition 38. *If the trader operates in a dealer market with knowledge of the market makers' common risk aversion coefficient ρ, she offers $\frac{z}{I}$ shares to each market maker at the bundling price $\mu\frac{z}{I} - \frac{\rho\sigma^2}{2}\frac{z^2}{I^2}$. $\mu z - \frac{\rho\sigma^2}{2}\frac{z^2}{I}$ the first-best liquidation value that can be achieved by the trader.*

Proof. Since market makers have the same level of risk-aversion, the trader optimally chooses to offer $\frac{z}{I}$ shares to each market maker. Each market maker's valuation for $\frac{z}{I}$ shares is $\mu\frac{z}{I} - \frac{\rho\sigma^2}{2}\frac{z^2}{I^2}$, which coincides the trader's take-it-or-leave-it offer.

B.1.2 Proofs

Proof of Proposition 18. Suppose the demand schedule takes the form of

$$x_{i,t} = a_{i,t} - b_t p_t \ ,t = 1, 2, ..., N \ . \tag{B.1}$$

In the Nth period, the demand schedule is

$$x_{i,N}(p_N) = \frac{\gamma}{\rho\sigma^2}(\mu - p_N) - \gamma \sum_{t=1}^{N-1} x_{it} \ .$$

The derivation is omitted since it is similar to the derivation of the second period

demand schedule in a two period model. Market clearing conditions imply that

$$z_N = \frac{I\gamma}{\rho\sigma^2}(\mu - p_N) - \gamma \sum_{t=1}^{N-1} z_t \ .$$

Equivalently,

$$p_N^* = \mu - \frac{\rho\sigma^2}{\gamma}\frac{z_N}{I} - \rho\sigma^2 \sum_{t=1}^{N-1} \frac{z_t}{I} \ .$$

Notice that as long as $z_t, t = 1, ..., N$ are fixed, p_N^* is not changing with respect to x_{it}.

For $t < N$, conjecture that

$$b_t = \frac{\gamma}{(1-\gamma)^{2(N-t)}\rho\sigma^2}$$

and $x_{i,t}$ takes a special form of

$$x_{i,t}(p_t) = f_t - b_t p_t - \gamma \sum_{j=1}^{t-1} x_{i,j} \tag{B.2}$$

Suppose this conjecture holds for periods $t+1, ..., N$. Then at period t, market maker i chooses $x_{i,t}$ to maximize

$$\mu \left(\sum_{k=1}^{t-1} x_{i,k} + x_{i,t} + \sum_{j=t+1}^{N} x_{i,j}(p_j^*) \right) - \frac{\rho\sigma^2}{2} \left(\sum_{k=1}^{t-1} x_{i,k} + x_{i,t} + \sum_{j=t+1}^{N} x_{i,j}(p_j^*) \right)^2 \tag{B.3}$$

$$- (\tilde{p}_t + \lambda_t x_{i,t}) x_{i,t} - \sum_{j=t+1}^{N} x_{i,j}(p_j^*) p_j^* \ . \tag{B.4}$$

Notice that p_j^* for $j \geq t+1$ is not changing with respect to $x_{i,t}$.

The first order condition implies that

$$\mu(1-\gamma)^{N-t} - \rho\sigma^2 \left(\sum_{k=1}^{t-1} x_{i,k} + x_{i,t} + \sum_{j=t+1}^{N} x_{i,j}(p_j^*) \right) (1-\gamma)^{N-t} \tag{B.5}$$

$$- p_t - \lambda_t x_{i,t} + \sum_{j=t+1}^{N} \gamma(1-\gamma)^{j-t-1} p_j^* = 0 \ . \tag{B.6}$$

Collecting terms to get[1]

$$\frac{1}{(1-\gamma)^{2(N-t)}\rho\sigma^2 + \lambda_t} = b_t \ .$$

Moreover,

$$\lambda_t = \frac{1}{(I-1)b_1} = \frac{1-\gamma}{b_t} \ .$$

This verifies that

$$b_t = \frac{\gamma}{(1-\gamma)^{2(N-t)}\rho\sigma^2} \ .$$

Plug b_t back to the first order condition to get

$$\frac{x_{i,t}}{b_t} = \mu(1-\gamma)^{N-t} - (1-\gamma)^{2(N-t)}\rho\sigma^2 \sum_{k=1}^{t-1} x_{i,k} \tag{B.7}$$

$$- \rho\sigma^2(1-\gamma)^{N-t}[\sum_{j=t+1}^{N} (1-\gamma)^{N-j}(f_j - b_j p_j^*)] \tag{B.8}$$

$$- p_t + \sum_{j=t+1}^{N} \gamma(1-\gamma)^{j-t-1} p_j^* \ . \tag{B.9}$$

[1]Note that $d(x_{i,t} + \sum_{j=t+1}^{N} x_{i,j}(p_j^*))/dx_{i,t} = (1-\gamma)^{N-t}$

This implies

$$x_{i,t} = \frac{\gamma}{(1-\gamma)^{N-t}\rho\sigma^2}\mu - \gamma\sum_{k=1}^{t-1} x_{i,k} - \frac{\gamma}{(1-\gamma)^{N-t}}[\sum_{j=t+1}^{N}(1-\gamma)^{N-j}(f_j - b_j p_j^*)] \quad \text{(B.10)}$$

$$- b_t p_t + b_t \sum_{j=t+1}^{N} \gamma(1-\gamma)^{j-t-1} p_j^* \ . \quad \text{(B.11)}$$

Combined with the expression of $x_{i,N}$, this confirms the conjecture.

Easy to see that

$$b_t \sum_{j=t+1}^{N} \gamma(1-\gamma)^{j-t-1} p_j^* = \frac{\gamma}{1-\gamma}\sum_{j=t+1}^{N}\frac{b_j p_j^*}{(1-\gamma)^{j-t}} = \frac{\gamma}{(1-\gamma)^{N-t+1}}\sum_{j=t+1}^{N}(1-\gamma)^{N-j} b_j p_j^* \ .$$

Notice that by market clearing condition

$$f_j - b_j p_j^* = \frac{z_j}{I} + \frac{\gamma}{I}\sum_{k=1}^{j-1} z_k \ .$$

Thus,

$$\sum_{j=t+1}^{N}(1-\gamma)^{N-j}(f_j - b_j p_j^*) = \frac{1}{I}\sum_{j=t+1}^{N}\left((1-\gamma)^{N-j} z_j + \gamma(1-\gamma)^{N-j}\sum_{k=1}^{j-1} z_k\right) \ .$$

That is,

$$\sum_{j=t+1}^{N}(1-\gamma)^{N-j}(f_j - b_j p_j^*) = \frac{1}{I}\left(\sum_{j=t+1}^{N} z_j + (1-(1-\gamma)^{N-t})\sum_{k=1}^{t} z_k\right) \quad \text{(B.12)}$$

$$= \frac{z}{I} - (1-\gamma)^{N-t}\sum_{k=1}^{t}\frac{z_k}{I} \ . \quad \text{(B.13)}$$

Notice that

$$-\sum_{j=t+1}^{N}(1-\gamma)^{N-j}b_j p_j^* = \frac{z}{I} - (1-\gamma)^{N-t}\sum_{k=1}^{t}\frac{z_k}{I} - \sum_{j=t+1}^{N}(1-\gamma)^{N-j}f_j \ .$$

Thus,

$$f_t = \frac{\gamma}{(1-\gamma)^{N-t}\rho\sigma^2}\mu - \frac{2\gamma-\gamma^2}{(1-\gamma)^{N-t+1}}\frac{z}{I} + \frac{2\gamma-\gamma^2}{1-\gamma}\sum_{k=1}^{t}\frac{z_k}{I} + \gamma\sum_{j=t+1}^{N}\frac{f_j}{(1-\gamma)^{j-t+1}} \ . \tag{B.14}$$

Equivalently,

$$\begin{aligned} f_t =& \frac{\gamma}{(1-\gamma)^{N-t}\rho\sigma^2}\mu - \frac{(2\gamma-\gamma^2)[1-(1-\gamma)^{N-t}]}{(1-\gamma)^{N-t+1}}\frac{z}{I} \\ & - \frac{2\gamma-\gamma^2}{1-\gamma}\sum_{j=t+1}^{N}\frac{z_j}{I} + \gamma\sum_{j=t+1}^{N}\frac{f_j}{(1-\gamma)^{j-t+1}} \ . \end{aligned} \tag{B.15}$$

Notice that $f_N = b_N\mu$. Then

$$f_{N-1} = \frac{\gamma}{(1-\gamma)^2\rho\sigma^2}\mu - \frac{2\gamma-\gamma^2}{(1-\gamma)^2}\frac{z}{I} + \frac{2\gamma-\gamma^2}{1-\gamma}\sum_{k=1}^{N-1}\frac{z_k}{I} \ .$$

Explicitly,

$$f_t = b_t\mu - \sum_{j=t+1}^{N}[\gamma + \frac{\gamma(1-\gamma)}{(1-\gamma)^{2(j-t)}}]\frac{z_j}{I} - \gamma(\frac{1}{(1-\gamma)^{2(N-t)}} - 1)\frac{z}{I} \ .$$

This means that the demand schedule is

$$\begin{aligned} x_{i,t}(p_t) =& \frac{\gamma}{(1-\gamma)^{2(N-t)}\rho\sigma^2}(\mu - p_t) - \sum_{j=t+1}^{N}[\gamma + \frac{\gamma(1-\gamma)}{(1-\gamma)^{2(j-t)}}]\frac{z_j}{I} \\ & - \gamma(\frac{1}{(1-\gamma)^{2(N-t)}} - 1)\frac{z}{I} - \gamma\sum_{j=1}^{t-1} x_{i,j} \end{aligned} \tag{B.16}$$

and in the equilibrium,

$$p_t^* = \mu - \rho\sigma^2\frac{z}{I} - \frac{1}{b_t}\left[\sum_{j=t+1}^{N}\frac{\gamma(1-\gamma)}{(1-\gamma)^{2(j-t)}}\frac{z_j}{I} + (1-\gamma)\frac{z_t}{I}\right] \tag{B.17}$$

$$= \mu - \rho\sigma^2\frac{z}{I} - \frac{\lambda_t z_t}{I} - \gamma\sum_{j=t+1}^{N}\frac{\lambda_j z_j}{I}\ . \tag{B.18}$$

Proof of Proposition 19. The proof proceeds in several steps. (1) The SPNE supply schedule is sub-game invariant. In other words, fixing the supply quantity, selling periods and market makers' conjectures, if a supply schedule is a part of the SPNE in a game, it is also a part of the SPNE as a larger game's sub-game. (2) If market makers' conjectures are scalable with respect to the number of shares left to be supplied, the SPNE supply schedule is also scalable. (3) Using no deviation conditions to uniquely pin down the inductive equation and construct market makers' conjectures which are linear in the number of shares left to be supplied.

(1) From the expression of p^*, given z, the equilibrium asset price only depends

on current period's supply and supplies in subsequent periods. This leads to the equivalence lemma. Specifically, suppose $z_1, ..., z_t$ is the SPNE supply schedule in a t period execution problem with total supply z. Consider a t period execution sub-game from period $N-t+1$ to period N in a N period execution problem with total supply $\tilde{z}$. If $\tilde{z} - \sum_{i=1}^{N-t} \tilde{z}_i = z$, then $\tilde{z}_{N-t+j} = z_j$ for $j = 1, ..., t$ is the SPNE supply schedule in this sub-game. The reversed statement is also true.

Proof Without loss of generality, I check there is no profitable one-shot deviation at period $N-t+1$. When $t = 1$, this is obvious. Suppose this claim holds for $t \leq n$. For $t = n+1$, since $z_1, ..., z_t$ is a SPNE supply schedule, for all $\acute{z}_1 = z_1 + k$,

$$\sum_{j=1}^{n+1} p_j^* z_j \geq \sum_{j=1}^{n+1} \acute{p}_j^* \acute{z}_j \tag{B.19}$$

where $\acute{z}_t$, $t = 2, ..., n+1$ is the SPNE supply profile with $\sum_{t=1}^{n+1} \acute{z}_t = z - \acute{z}_1$. In the first period, expecting supply schedule $z_1, ...z_{n+1}$, the market maker's demand schedule is

$$x_{i,1}(p_1) = b_1(\mu - p_1) - \sum_{j=2}^{n+1} [\gamma + \frac{\gamma(1-\gamma)}{(1-\gamma)^{2(j-1)}}] \frac{z_j}{I} - \gamma(\frac{1}{(1-\gamma)^{2n}} - 1)\frac{z}{I} \ .$$

where $b_1 = \frac{\gamma}{(1-\gamma)^{2n}\rho\sigma^2}$.[2] If the trader deviates by supplying $\acute{z}_1 = z_1 + k$, for each market maker, the market clearing quantity becomes $x_{i,1} = \frac{z_1+k}{I}$. This implies the

[2]In the proof of this section, b_t specifically refer to the b function in a $n+1$ period problem

market clearing price becomes

$$\acute{p}_1^* = \mu - \rho\sigma^2\frac{z}{I} - \frac{1}{b_1}\left[\sum_{j=2}^{n+1}\frac{\gamma(1-\gamma)}{(1-\gamma)^{2(j-1)}}\frac{z_j}{I} + (1-\gamma)\frac{z_1}{I} + \frac{k}{I}\right] \tag{B.20}$$

$$= \mu - \rho\sigma^2\frac{z}{I} - \frac{\lambda_1 z_1}{I} - \gamma\sum_{j=2}^{n+1}\frac{\lambda_j z_j}{I} - \frac{\lambda_1}{(1-\gamma)I}k \ . \tag{B.21}$$

After the first period deviation, equilibrium prices from period 2 to $n+1$ are

$$\acute{p}_t^* = \mu - \rho\sigma^2\frac{z}{I} - \frac{1}{b_t}\left[\sum_{j=t+1}^{n+1}\frac{\gamma(1-\gamma)}{(1-\gamma)^{2(j-t)}}\frac{\acute{z}_j}{I} + (1-\gamma)\frac{\acute{z}_t}{I}\right] \tag{B.22}$$

$$= \mu - \rho\sigma^2\frac{z}{I} - \frac{\lambda_t\acute{z}_t}{I} - \gamma\sum_{j=t+1}^{n+1}\frac{\lambda_j\acute{z}_j}{I} \ . \tag{B.23}$$

Next consider the t period sub-game. I claim that $\tilde{z}_{N-t+j} = z_j$ for $j = 1, ..., t$ is the SPNE supply schedule in this sub-game. For a deviation $\acute{\tilde{z}}_{N-t+1} = z_1 + k$, $\acute{\tilde{z}}_{N-t+j} = \tilde{z}_j$ is a SPNE supply schedule. Moreover, $\tilde{p}^*_{N-t+j} = p_t^* - \rho\sigma^2\frac{\tilde{z}-z}{I}$, $\acute{\tilde{p}}^*_{N-t+j} = \tilde{p}_t^* - \rho\sigma^2\frac{\tilde{z}-z}{I}$ and $\sum_{j=1}^{t} z_{N-t+j} = z$. This means for any $\acute{\tilde{z}}_{N-t+1} = z_1 + k$,

$$\begin{aligned}\sum_{j=1}^{n+1}\tilde{p}^*_{N-n-1+j}\tilde{z}_{N-n-1+j} &= \sum_{j=1}^{n+1}p_j^* z_j - \rho\sigma^2\frac{z(\tilde{z}-z)}{I} \\ \geq \sum_{j=1}^{n+1}\acute{\tilde{p}}^*_{N-n-1+j}\acute{\tilde{z}}_{N-n-1+j} &= \sum_{j=1}^{n+1}\acute{p}_j^*\acute{z}_j - \rho\sigma^2\frac{z(\tilde{z}-z)}{I} \ .\end{aligned}$$

The reversed statement can be proved in the similar manner.

(2) Then notice that we have the following scaling lemma: Suppose $z_1, ..., z_t$ is a SPNE supply schedule in a t period execution problem with total supply z and

market maker's conjectures are linear functions with respect to the quantity of shares left.[3] Then for any number d, $d \cdot z_1, ..., d \cdot z_t$ is a SPNE supply schedule in a t period execution problem with total supply $d \cdot z$ if market makers' conjectures are scaled by the factor d.

Proof We consider two situations: $d = 0$ and $d \neq 0$. If $d = 0$, recall that in the scheduling case, the optimal supply schedule is supplying 0 in all periods. Furthermore, the trader's equilibrium liquidation value in the no scheduling case (if an equilibrium exists) is upper-bounded by the liquidation value with scheduling. Thus, we only need to check that supplying 0 in all periods is the SPNE supply schedule without scheduling. This can be proved by contradiction. Without loss of generality, suppose the trader is deviating from selling 0 in the first period by selling a positive quantity. From the market clearing price (in either the sequential case or the simultaneous case), the unit selling price is smaller than μ. In subsequent periods, under any SPNE supply schedule, the unit cost of buying back shares sold, which is lower-bounded by the optimal buying back schedule with scheduling, is higher than μ. Thus, the trader does no has a profitable one-shot deviation.

Now consider $d \neq 0$. When $t = 1$, the claim is trivial. Suppose this claim holds for $t \leq n$. Then for $t = n + 1$, if $z_1, ..., z_t$ is the SPNE supply schedule, by one-shot deviation principle, there is no $k \neq 0$ such that the trader is willing to deviate in the first period by supplying $z_1 + k$. When the total supply is $d \cdot z$, from the inductive hypothesis, if the trader deviate by supplying $d \cdot z_1 + d \cdot k$ in the first period, the supply schedule in the following periods is the schedule from the previous problem

[3]Here linear, different from affine, implies the graph of the function has zero as the intercept.

after deviation of size k scaling by the factor d. Simple algebra shows the no deviation condition of the previous problem is equivalent to the no deviation condition in the latter problem.

(3) We first consider the SPNE schedule where $z_1 \neq z$ and $z_N \neq 0$ in a game with more than one period of trading. To simplify notation, denote $p_t(k)$ and $z_t(k)$ to be the supply and the price in period t when the trader deviates from the SPNE schedule by supplying additional k units of asset in period 1. By definition, $z_1(k) = z_1 + k$. By step (1) and (2), if $z_2, ..., z_N$ is a SPNE supply schedule, after the one-shot deviation of size k, $z_t(k) = z_t - \frac{z_t}{\sum_{i=2}^{N} z_i} k$.[4] Thus, $z_1^{'}(k) = 1$ and $z_t^{'}(k) = -\frac{z_t}{\sum_{i=2}^{N} z_i}$.

From the calculation above

$$p_1(k) = \mu - \rho\sigma^2 \frac{z}{I} - \frac{\lambda_1 z_1}{I} - \gamma \sum_{j=2}^{N} \frac{\lambda_j z_j}{I} - \frac{\lambda_1}{(1-\gamma)I} k$$

and for $t \geq 2$,

$$p_t(k) = \mu - \rho\sigma^2 \frac{z}{I} - \frac{\lambda_t z_t(k)}{I} - \gamma \sum_{j=t+1}^{N} \frac{\lambda_j z_j(k)}{I} \ .$$

No one-shot deviation means that $\sum p_t(k) z_t(k)$ is maximized at $k = 0$. From the first order condition

$$\sum_{t=1}^{N} \left[p_t^{'}(0) z_t(0) + p_t(0) z_t^{'}(0) \right] = 0 \ .$$

[4]Here I slightly abuse the notation by using z_t to both represent a quantity and a function. I always regard $z_t(k)$ as a function.

Equivalently,

$$-\frac{\lambda_1 z_1}{1-\gamma}+\sum_{t=2}^{N}\left[\frac{\lambda_t z_t^2}{\sum_{i=2}^{N} z_i}+\gamma z_t\sum_{j=t+1}^{N}\frac{\lambda_j z_j}{\sum_{i=2}^{N} z_i}\right]$$
$$-\lambda_1 z_1-\gamma\sum_{j=2}^{N}\lambda_j z_j+\sum_{t=2}^{N}\frac{z_t}{\sum_{i=2}^{N} z_i}\left[\lambda_t z_t+\gamma\sum_{j=t+1}^{N}\lambda_j z_j\right]=0$$

Since $\alpha_t=\frac{z_{N-t+1}}{z_N}$ and $A_t=\sum_{i=1}^{t}\alpha_t$,

$$\frac{\lambda_1(2-\gamma)}{1-\gamma}\alpha_N=2\sum_{t=2}^{N}\frac{\alpha_{N-t+1}}{A_{N-1}}\left[\lambda_t\alpha_{N-t+1}+\gamma\sum_{j=t+1}^{N}\lambda_j\alpha_{N-j+1}\right]-\gamma\sum_{t=2}^{N}\lambda_t\alpha_{N-t+1}$$

Note that $\frac{\lambda_t}{\lambda_1}=\frac{b_1}{b_t}=\frac{1}{(1-\gamma)^{2(t-1)}}$. Thus,

$$\frac{2-\gamma}{1-\gamma}\alpha_N=2\sum_{t=2}^{N}\frac{\alpha_{N-t+1}}{A_{N-1}}\left[\frac{\alpha_{N-t+1}}{(1-\gamma)^{2(t-1)}}+\gamma\sum_{j=t+1}^{N}\frac{\alpha_{N-j+1}}{(1-\gamma)^{2(j-1)}}\right]-\gamma\sum_{t=2}^{N}\frac{\alpha_{N-t+1}}{(1-\gamma)^{2(t-1)}}$$

Changing the subscript to get

$$\frac{2-\gamma}{1-\gamma}\alpha_N=2\sum_{t=1}^{N-1}\frac{\alpha_t}{A_{N-1}}\left[\frac{\alpha_t}{(1-\gamma)^{2(N-t)}}+\gamma\sum_{j=1}^{t-1}\frac{\alpha_j}{(1-\gamma)^{2(N-j)}}\right]-\gamma\sum_{t=1}^{N-1}\frac{\alpha_t}{(1-\gamma)^{2(N-t)}}\ .$$

Since $p_t^{''}(k)=z_t^{''}(k)=0$, the second order condition requires

$$\sum_{t=1}^{N}p_t^{'}(0)z_t^{'}(0)<0\ .$$

This is equivalent to

$$-\frac{\lambda_1}{1-\gamma} - \sum_{t=2}^{N}(\lambda_t \frac{z_t}{\sum_{i=2}^{N} z_i} + \gamma \sum_{j=t+1}^{N} \lambda_j \frac{z_j}{\sum_{i=2}^{N} z_i})\frac{z_t}{\sum_{i=2}^{N} z_i} < 0 \ .$$

This is automatically satisfied when $z_t > 0$ for all t.

Now we check that in the equilibrium the trader never purchase shares in execution. This is equivalent to $\alpha_t > 0$ for all t. $\alpha_1 = 1 > 0$ by definition. Let

$$D_{n-1} = \sum_{t=1}^{n-1} \frac{\alpha_t}{A_{n-1}} \left[\frac{\alpha_t}{(1-\gamma)^{2(n-t)}} + \gamma \sum_{j=1}^{t-1} \frac{\alpha_j}{(1-\gamma)^{2(n-j)}} \right]$$

Suppose $\alpha_n > 0$, this is equivalent to $D_{n-1} > \gamma \sum_{t=1}^{n-1} \frac{\alpha_t}{(1-\gamma)^{2(n-t)}}$. We want to show

$$D_n > \gamma \sum_{t=1}^{n} \frac{\alpha_t}{(1-\gamma)^{2(n+1-t)}} \tag{B.24}$$

since this implies $\alpha_{n+1} > 0$. Inequality (B.24) is equivalent to

$$\frac{A_{n-1}}{A_n} D_{n-1} + \frac{2\alpha_n}{A_n} \left[\alpha_n + \gamma \sum_{j=1}^{n-1} \frac{\alpha_j}{(1-\gamma)^{2(n-j)}} \right] > \gamma \sum_{t=1}^{n-1} \frac{\alpha_t}{(1-\gamma)^{2(n-t)}} + \gamma \alpha_n$$

A sufficient condition is

$$\frac{2\alpha_n}{A_n}\left[\alpha_n+\gamma\sum_{j=1}^{n-1}\frac{\alpha_j}{(1-\gamma)^{2(n-j)}}\right]>\frac{\alpha_n}{A_n}\gamma\sum_{t=1}^{n-1}\frac{\alpha_t}{(1-\gamma)^{2(n-t)}}+\gamma\alpha_n \tag{B.25}$$

$$\iff 2\left[\alpha_n+\gamma\sum_{j=1}^{n-1}\frac{\alpha_j}{(1-\gamma)^{2(n-j)}}\right]>\gamma\sum_{t=1}^{n-1}\frac{\alpha_t}{(1-\gamma)^{2(n-t)}}+A_n \tag{B.26}$$

$$\iff 2\alpha_n+\gamma\sum_{t=1}^{n-1}\frac{\alpha_t}{(1-\gamma)^{2(n-t)}}>A_n\ . \tag{B.27}$$

Since $A_n=\sum_{i=1}^{n}\alpha_i$ and $\gamma\geq\frac{1}{2}$, the sufficient condition is satisfied.

An important implication is that in this case the trader would prefer selling in a perfectly competitive market. It is also easy to check that $z_1\neq z$ and $z_N\neq 0$ in any SPNE supply schedule.[5]

Proof of Proposition 20. For generic N, when the supplier can commit, the supplying problem is equivalent to finding $z_1,...,z_N$ adding up to z to minimize

$$\sum_{t=1}^{N}\frac{z_t}{b_t}\left[\sum_{j=t+1}^{N}\frac{\gamma(1-\gamma)}{(1-\gamma)^{2(j-t)}}\frac{z_j}{I}+(1-\gamma)\frac{z_t}{I}\right]\cdot I\ .$$

This implies that for all t,

$$2(1-\gamma)\frac{z_t}{b_t}+\frac{1}{b_t}\sum_{j=t+1}^{N}\frac{\gamma(1-\gamma)}{(1-\gamma)^{2(j-t)}}z_j+\frac{1}{b_t}\sum_{k=1}^{t-1}\gamma(1-\gamma)z_k=\Lambda \tag{B.28}$$

[5]Apply induction starting from a one-period game with $\alpha_N=1$

where Λ is the multiplier on $z - \sum_{j=1}^{N} z_j = 0$.

Multiplying $\frac{b_N}{1-\gamma}$ on both sides to get

$$2(1-\gamma)^{2(N-t)} z_t + \sum_{j=t+1}^{N} \gamma(1-\gamma)^{2(N-j)} z_j + \gamma(1-\gamma)^{2(N-t)} \sum_{k=1}^{t-1} z_k = C \ .$$

Plug in t and $t+1$ and take the difference to get

$$z_{t+1} = (1-2\gamma) z_t - \gamma^2 \sum_{k=1}^{t-1} z_k$$

for $t \geq 2$ and

$$z_2 = (1-2\gamma) z_1 \ .$$

This together with $\sum_{k=1}^{N} z_k = z$ pins down the optimal supply schedule.

Conjecture that $z_t = (1-\gamma)^{t-2}(1-t\gamma) z_1$ for $t \geq 2$. Suppose this holds from $t = 2$ to $t = k$. Notice that $\sum_{i=1}^{k} z_i = k(1-\gamma)^{k-1} z_1$. Thus

$$z_{k+1} = (1-\gamma)^2 z_k - k(1-\gamma)^{k-1}\gamma^2 z_1 = (1-\gamma)^{k-1}[1-(k+1)\gamma] z_1 \ .$$

This confirms the conjecture. Since $\sum_{j=1}^{N} z_j = z$, $z_t = \frac{(1-t\gamma)z}{N(1-\gamma)^{N-t+1}}$. Notice that since $\gamma \geq \frac{1}{2}$ only z_1 is positive; $z_2, ..., z_N$ are all taking negative positions.

From the calculation above,

$$\Lambda = \frac{(2-(N+1)\gamma)\rho\sigma^2}{N\gamma} z \ .$$

This means

$$\begin{aligned}
&(1-\gamma)\frac{z_t}{b_t}+\frac{1}{b_t}\sum_{j=t+1}^{N}\frac{\gamma(1-\gamma)}{(1-\gamma)^{2(j-t)}}z_j\\
=&\Lambda-(1-\gamma)\frac{z_t}{b_t}-\frac{1}{b_t}\sum_{k=1}^{t-1}\gamma(1-\gamma)z_k\\
=&\frac{(2-(N+1)\gamma)\rho\sigma^2}{N\gamma}z-\frac{(1-t\gamma)(1-\gamma)^{N-t}\rho\sigma^2}{N\gamma}z-\frac{(t-1)\gamma(1-\gamma)^{N-t}\rho\sigma^2}{N\gamma}z\\
=&\frac{(2-(N+1)\gamma-(1-\gamma)^{N-t+1})\rho\sigma^2}{N\gamma}z
\end{aligned}$$

Thus,

$$p_t^*=\mu-\frac{(2-\gamma-(1-\gamma)^{N-t+1})\rho\sigma^2}{N\gamma}\frac{z}{I}\ .$$

p_t^* is increasing in t. Moreover, for $N\geq 2$, $p_1^*>\mu-\frac{\rho\sigma^2}{\gamma}\frac{z}{I}$, the selling price in a one period trading game. In other words, though the trading volume becomes larger, the price impact is smaller.

The trader's liquidation value is

$$V=\mu z-\frac{1+\frac{N-1}{2}\gamma}{N\gamma}\rho\sigma^2\frac{z^2}{I}\ . \tag{B.29}$$

One interesting observation is that the optimal supply schedule does not depend on μ, σ and ρ. Moreover, it is perfectly scalable. Essentially it only depends on the number of market makers trading periods.

Proof of Proposition 21 Denote the Lagrangian multiplier on the non-negative constraint of z_t by Λ_t. Taking the first order conditions and after some algebraic manipulation, we have

$$z_{t+1} = (1-2\gamma)z_t - \gamma^2 \sum_{k=1}^{t-1} z_k - \frac{\Lambda_t - \Lambda_{t+1}}{(2-\gamma)\lambda_{t+1}}$$

for $t \geq 2$ and

$$z_2 = (1-2\gamma)z_1 - \frac{\Lambda_1 - \Lambda_2}{(2-\gamma)\lambda_2} \ .$$

Conjecture that $z_1 > 0$. This implies that $\Lambda_2 > 0$ and $z_2 = 0$. Inductively, we have $z_t = 0$ for $t \geq 2$. Thus, $z_1 = z$. Then apply induction on N to see that $z_1 > 0$ (contrasting $z_1 = 0$) is indeed the optimal trading schedule. Consequently, if the trader engages in scheduling but cannot flip orders, the optimal execution is to sell everything at the first period. Notice that in this case the execution is worse than the perfect competition situation. To see the liquidation is better than the non-scheduling case, notice that in the non-scheduling case the trader never oversells.

B.2 Model Extensions

B.2.1 Multiple Predators with Scheduling

In this appendix, we first provide details on the model of Section 3.5 with multiple predators. Then we show how to obtain a closed form of the trader's and the predator's trading schedules when $I_p = 1$.

Proposition 39. *Denote each predator's supply schedule at period t by q_t. The trading schedules of the trader and the predators are determined by the following equations*

$$(2-\gamma)z_{t+1}+(1-\gamma)I_p q_{t+1} = (1-2\gamma)(2-\gamma)z_t+(1-\gamma)^2 I_p q_t-\gamma^2(2-\gamma)\sum_{k=1}^{t-1} z_k \ . \quad \text{(B.30)}$$

$$(1-\gamma)z_{t+1}+[I_p(1-\gamma)+1]q_{t+1} = (1-\gamma)^2 z_t+[(1-\gamma)^2(I_p+1)-\gamma]q_t-\gamma^2(2-\gamma)\sum_{k=1}^{t-1} q_k \ . \quad \text{(B.31)}$$

together with constraints $\sum_{t=1}^{N} z_t = \mathrm{z}$ and $\sum_{t=1}^{N} q_t = 0$.

Proof. Let the multiplier on the trader's supply constraint be Λ and the multiplier on each predator's supply constraint be Γ. The trader's problem is equivalent to finding a supply schedule $\{z_t\}$ adding up to z to minimize

$$\sum_{t=1}^{N} z_t[\frac{\lambda_t(z_t+I_p q_t)}{I}+\gamma\sum_{j=t+1}^{N}\frac{\lambda_j(z_j+I_p q_j)}{I}]\cdot I \ .$$

Each predator's problem is equivalent to find a supply schedule $\{\tilde{q}_t\}$ adding up to 0 to minimize

$$\sum_{t=1}^{N}\tilde{q}_t\left\{\frac{\lambda_t[z_t+\tilde{q}_t+(I_p-1)q_t]}{I}+\gamma\sum_{j=t+1}^{N}\frac{\lambda_j[z_j+\tilde{q}_t+(I_p-1)q_t]}{I}\right\}\cdot I$$

with the symmetric constraint $\tilde{q}_t = q_t$.

For any t, the first order condition for the trader is

$$\lambda_t(2z_t + I_p q_t) + \gamma \sum_{j=t+1}^{N} \lambda_j(z_j + I_p q_j) + \gamma\lambda_t \sum_{k=1}^{t-1} z_k = \Lambda \ ; \tag{B.32}$$

the first order condition for the predator is

$$\lambda_t[z_t + (I_p + 1)q_t] + \gamma \sum_{j=t+1}^{N} \lambda_j(z_j + I_p q_j) + \gamma\lambda_t \sum_{k=1}^{t-1} q_k = \Gamma \ . \tag{B.33}$$

From equation (B.32),

$$(2-\gamma)z_{t+1}+(1-\gamma)I_p q_{t+1} = (1-2\gamma)(2-\gamma)z_t+(1-\gamma)^2 I_p q_t-\gamma^2(2-\gamma)\sum_{k=1}^{t-1} z_k \ . \tag{B.34}$$

From equation (B.33),

$$(1-\gamma)z_{t+1}+[I_p(1-\gamma)+1]q_{t+1} = (1-\gamma)^2 z_t+[(1-\gamma)^2(I_p+1)-\gamma]q_t-\gamma^2(2-\gamma)\sum_{k=1}^{t-1} q_k \ . \tag{B.35}$$

Now consider the special case when $I_p = 1$. Let $a_t = \sum_{i=1}^{t}(z_i + q_i)$. This implies

$$(3 - 2\gamma)(a_{t+1} - a_t) = [3(1 - \gamma)^2 - \gamma](a_t - a_{t-1}) - \gamma^2(2 - \gamma)a_{t-1} \ . \tag{B.36}$$

Equivalently,

$$(3 - 2\gamma)a_{t+1} = 3(1 - \gamma)(2 - \gamma)a_t - (3 - \gamma)(1 - \gamma)^2 a_{t-1} \ . \tag{B.37}$$

Two roots of the characteristic polynomial are $1-\gamma$ and $\frac{(1-\gamma)(3-\gamma)}{3-2\gamma}$. Moreover, $a_N = z$ and $(3-2\gamma)a_2 = 3(1-\gamma)(2-\gamma)a_1$. This means[6]

$$a_t = \frac{\frac{(3-\gamma)^t}{(3-2\gamma)^t} - 1}{\frac{(3-\gamma)^N}{(3-2\gamma)^N} - 1} \cdot \frac{z}{(1-\gamma)^{N-t}} \ .$$

Let $c_t = \sum_{i=1}^{t}(z_i - q_i)$. We have

$$c_{t+1} - c_t = [(1-\gamma)^2 - \gamma](c_t - c_{t-1}) - \gamma^2(2-\gamma)c_{t-1} \ . \tag{B.38}$$

Equivalently,

$$c_{t+1} = (1-\gamma)(2-\gamma)c_t - (1-\gamma)^3 c_{t-1} \ . \tag{B.39}$$

Two roots of the characteristic polynomial are $1-\gamma$ and $(1-\gamma)^2$. Moreover, $c_N = z$ and $c_2 = (1-\gamma)(2-\gamma)c_1$. This means[7]

$$c_t = \frac{(1-\gamma)^t - (1-\gamma)^{2t}}{(1-\gamma)^N - (1-\gamma)^{2N}} z \ . \tag{B.40}$$

Notice that $a_0 = c_0 = 0$. Thus, $z_t = \frac{a_t + c_t}{2} - \frac{a_{t-1} + c_{t-1}}{2}$ and $q_t = \frac{a_t - c_t}{2} - \frac{a_{t-1} - c_{t-1}}{2}$ for all t.

We can also show that the pattern of overselling and buying back by the trader (and in aggregate) is robust. Let $K_1 = 1 - \gamma < 1$ and $K_2 = \frac{3-\gamma}{3-2\gamma} > 1$. Since

[6] $a_t = B[(\frac{(1-\gamma)(3-\gamma)}{3-2\gamma})^t - (1-\gamma)^t]$with parameter B determined by the terminal condition.

[7] $c_t = B[(1-\gamma)^t - (1-\gamma)^{2t}]$ with parameter B determined by the terminal condition.

$\gamma \in [\frac{1}{2}, 1)$, $K_1(K_2+1) < 1$.

$$\frac{da_t}{dt} > 0 \implies K_2^t < \frac{lnK_1}{lnK_1 + lnK_2} \ .$$

This means a_t is unimodal. Moreover, notice that $a_1 > a_N = z$ because $K_2^N - 1 = (K_2 - 1)K \sum_{t=0}^{N-1} K_2^t$, which implies

$$K_2 - 1 > (K_2^N - 1)K_1^{N-1} \ .$$

To show that the trader's schedule also exhibits an oversell-buyback pattern, it is suffice to have $q_1 < 0$.

$$\begin{aligned} q_1 < 0 &\iff a_1 < c_1 \\ &\iff K_1^N + K_2(1 - K_1^N) < K_1 + K_2^N(1 - K_1) \\ &\iff K_2 - K_1 < K_2^N(1 - K_1) + K_1^N(K_2 - 1) \ . \end{aligned}$$

Notice the when $N = 1$, $LHS = RHS$. Moreover,

$$K_2^N(1 - K_1) + K_1^N(K_2 - 1) < K_2^{N+1}(1 - K_1) + K_1^{N+1}(K_2 - 1)$$

since

$$K_1^N(K_1 - 1)(K_2 - 1) < K_2^N(K_1 - 1)(K_2 - 1) \ .$$

This implies $q_1 < 0$ when $N > 1$.

B.2.2 Multiple Traders with Scheduling

In this section, we consider an extension with scheduling and multiple traders. An important observation is that traders stick to the oversell-buyback pattern in liquidation. Let the number of traders be $I_s \geqslant 2$ and assume that each of them needs to sell $\frac{z}{I_s}$ shares. We focus on the symmetric equilibrium.

Proposition 40. *In the equilibrium, each trader submit the order* $z_t^i = \frac{\gamma(1-\gamma)^{t-1}-\theta\gamma(1-\theta\gamma)^{t-1}}{(1-\theta\gamma)^N-(1-\gamma)^N}\frac{z}{I_s}$ *where* $\theta = \frac{2-\gamma}{I_s(1-\gamma)+1} \leq 1$. *The aggregate trading schedule involves overselling and buying back.*

Proof. The first order condition with respect to the first trader implies

$$\frac{(1-\gamma)u_t}{b_t} + \frac{(1-\gamma)z_t^1}{b_t} + \frac{1}{b_t}\sum_{j=t+1}^{N}\frac{\gamma(1-\gamma)}{(1-\gamma)^{2(j-t)}}u_j + \frac{1}{b_t}\sum_{k=1}^{t-1}\gamma(1-\gamma)z_k = \Lambda$$

where $u_t = \sum_{i=1}^{I_s} z_t^i$ and Λ is the multiplier over the constraint $\sum_{t=1}^{N} z_t^1 = \frac{z}{I_s}$. By symmetry, $u_t = I_s z_t^1$. Thus,

$$(I_s+1)(1-\gamma)^{2(N-t)}z_t^1 + I_s\sum_{j=t+1}^{N}\gamma(1-\gamma)^{2(N-j)}z_j^1 + r(1-\gamma)^{2(N-t)}\sum_{k=1}^{t-1}z_k^1 = C.$$

This means

$$[(I_s+1)-I_s\gamma]z_{t+1}^1 = [(I_s+1)(1-\gamma)^2-\gamma]z_t^1 - \gamma^2(2-\gamma)\sum_{k=1}^{t-1}z_k^1.$$

Let $z_t = \sum_{i=1}^{I_s} z_t^i$ and $\theta = \frac{2-\gamma}{I_s(1-\gamma)+1} \leq 1$. Then we have

$$z_{t+1} = [1-(\theta+1)\gamma]z_t - \theta\gamma^2 \sum_{k=1}^{t-1} z_k = (1-\theta\gamma)(1-\gamma)z_t - \theta\gamma^2 \sum_{k=1}^{t} z_k.$$

Let $a_k = \sum_{k=1}^{t} z_k$. For $\theta \neq 1$,

$$a_n = \frac{(1-\theta\gamma)^n - (1-\gamma)^n}{(1-\theta\gamma)^N - (1-\gamma)^N} z.$$

Thus, $z_t^i = \frac{z_t}{I_s}$ where

$$z_t = a_t - a_{t-1} = \frac{\gamma(1-\gamma)^{t-1} - \theta\gamma(1-\theta\gamma)^{t-1}}{(1-\theta\gamma)^N - (1-\gamma)^N} z.$$

To see that the aggregate order involves overselling and buying back, notice that$a_N = z$. Thus is is suffice to show that $a_1 > z$. This can be proved by induction on N and using the fact that $x^{n-1} - x^n$increases in x for $n \geq 2$ and $x \in (0, \frac{1}{2}]$.

Appendix C

Appendix to Chapter 4

C.1 Instantaneous Stage Game Setting

We adopt a variation of the instantaneous stage game from Anderson and Smith (2013), as shown below. The instantaneous payoffs to the market maker per unit of trading rate are given by the entries of the matrices, the opposite of which are the insider's payoffs. At each instant, the insider, knowing which game she is playing, can choose actions B (market buy order) and S (market sell order) with respective intensities $\alpha \in [0, +\infty)$ and $\beta \in [0, +\infty)$. It is straightforward to see that the state contingent payoff for the insider at instant t depends only on $\theta_t = \alpha_t - \beta_t \in (-\infty, \infty)$, the net buying (selling) rate. The market maker can choose between two actions a and b, with a referring to setting the price to 1 and b to setting the price to 0. Restricting the market maker's action to a or b is without loss of generality as long as we allow him to mix between the two actions. Let p_t denote the probability weight on action a at time t. Then mixing with probability p_t can be equivalently

interpreted as setting the price to p_t. If the insider does not exist, the market maker plays a trivial single-player game. The payoffs per unit of trading rate of the possible games are given in the following matrices.

Table C.1: With Insider and $v = 1$

	a (set $p = 1$)	b (set $p = 0$)
B(Buy)	0	-1
S(Sell)	0	1

Table C.2: With Insider and $v = 0$

	a (set $p = 1$)	b (set $p = 0$)
B(Buy)	1	0
S(Sell)	-1	0

Table C.3: No Insider

	a (set $p = 1$)	b (set $p = 0$)
Null	0	0

Two underlying assumptions deserve further elaboration. First, we model the stage game played at each instant as a simultaneous-move game. On one hand, as argued in Kyle (1985b), if the insider can submit orders after observing the price, she would trade unbounded quantities, since her quantity traded does not affect the immediate execution price. This trivializes the problem and leads to unrealistic predictions. On the other hand, allowing the market maker to set the price after

observing the instantaneous order flow $dY_t = \theta_t dt + \sigma dZ_t$ does not give him extra information for pricing the asset. This is because the potential insider's trading is of order dt, dominated by the magnitude of liquidity trading that is of order $dt^{\frac{1}{2}}$. As a result, at instant t, the market maker sets the price based only on the information conveyed by the entire order path up to time t. Hence, it is without loss of generality to adopt the simultaneous-move stage game in this continuous-time model. Second, we assume that the market maker's objective is to maximize his expected payoff from playing the zero-sum game with the potential insider, rather than making profit from exploiting the liquidity trader. This is reasonable, given his job of market making, and further justifies the setting of the simultaneous-move stage game. Otherwise, if the market maker maximizes his gains from both the insider and the liquidity trader, he would set the price to 1 when observing $dY_t > 0$ and to 0 when observing $dY_t < 0$. In this way, he can achieve infinite expected payoff by essentially "front-running" the investors. In practice, this kind of practice is deemed unethical and thus prohibited.

C.2 Proofs

C.2.1 Proof of Lemma 4

Proof. If $\pi_0\theta_0 + \pi_1\theta_1 = 0$, we have

$$\begin{aligned}\lambda_0 &= \frac{\pi_0[(1-\pi_0)\theta_0 - \pi_1\theta_1]}{\sigma^2} \\ &= \frac{\pi_0[(\theta_0 - \pi_0\theta_0 - \pi_1\theta_1]}{\sigma^2} \\ &= \frac{\pi_0\theta_0}{\sigma^2}\end{aligned}$$

and

$$\begin{aligned}\lambda_1 &= \frac{\pi_1[(1-\pi_1)\theta_1 - \pi_0\theta_0]}{\sigma^2} \\ &= \frac{\pi_1[(\theta_1 - \pi_1\theta_1 - \pi_0\theta_0]}{\sigma^2} \\ &= \frac{\pi_1\theta_1}{\sigma^2}.\end{aligned}$$

Hence,

$$\lambda_0 + \lambda_1 = \frac{\pi_0\theta_0 + \pi_1\theta_1}{\sigma^2} = 0 \ .$$

Therefore,

$$\begin{aligned}d\pi_u &= -(d\pi_0 + d\pi_1) \\ &= -\left(\lambda_0 + \lambda_1\right) \cdot dY \\ &= 0 \ .\end{aligned}$$

□

C.2.2 Proof of Proposition 23

Proof. We guess and verify an equilibrium in which the potentially existing insider's expected trading rate is always $\phi = 0$. We then show that this equilibrium is essentially unique.

Suppose the expected trading rate is always $\phi = 0$. Lemma **4** suggests that the belief pair (π_0, π_1) always moves along a -45-degree line through the initial state $(\hat{\pi}_0, \hat{\pi}_1)$, and we can use π_1 , the probability that the insider exists and that the true value is 1, as the state variable. Then the HJB equation for the insider knowing that the true value equals $\upsilon \in \{0, 1\}$ is

$$rV_\upsilon = sup_{\theta_\upsilon \in (-\infty,\infty)} \theta_\upsilon u_\upsilon(p) + \lambda_1 \cdot (\theta_\upsilon - \phi) V_\upsilon^{'} + \frac{1}{2}\sigma^2 \lambda_1^2 V_\upsilon^{''} \ , \tag{C.1}$$

with

$$\lambda_1 = \frac{\pi_1[(1-\pi_1)\theta_1 - (1-\hat{\pi}_u - \pi_1)\theta_0]}{\sigma^2} \ , \tag{C.2}$$

$$u_0(p) = -p \ , \tag{C.3}$$

and

$$u_1(p) = 1 - p \ . \tag{C.4}$$

Since the insider's trading rates are unconstrained in this model, to obtain a finite

maximum for equation (C.1) in equilibrium, it must be

$$u_0(p) + \lambda_1 V_0^{'} = 0 \tag{C.5}$$

and

$$u_1(p) + \lambda_1 V_1^{'} = 0 \ . \tag{C.6}$$

Plugging $\phi = 0$, equations (C.5) and (C.6) into equation (C.1), we obtain

$$\frac{rV_0(\pi_1)}{\sigma^2} = \frac{1}{2}\lambda_1^2 V_0^{''}(\pi_1) \ , \tag{C.7}$$

and

$$\frac{rV_1(\pi_1)}{\sigma^2} = \frac{1}{2}\lambda_1^2 V_1^{''}(\pi_1) \ . \tag{C.8}$$

Define $\Lambda(\pi_1) = V_0(\pi_1) - V_1(\pi_1)$. Then from (C.5) $-$ (C.6), we get

$$\lambda_1 = \frac{1}{\Lambda^{'}(\pi_1)} \ . \tag{C.9}$$

Let $\rho = \frac{4r}{\sigma^2}$. Substitute equations (C.9) into equations (C.7) and (C.8) to get

$$\rho V_0(\pi_1) = 2\frac{V_0^{''}(\pi_1)}{(\Lambda^{'}(\pi_1))^2} \ , \tag{C.10}$$

and

$$\rho V_1(\pi_1) = 2\frac{V_1^{''}(\pi_1)}{(\Lambda^{'}(\pi_1))^2} \ . \tag{C.11}$$

From equations (C.10) and (C.11), we get

$$\rho\Lambda(\pi_1) = 2\frac{\Lambda''(\pi_1)}{(\Lambda'(\pi_1))^2} \ . \tag{C.12}$$

The general solution to equation (C.12) is

$$\Lambda(\pi_1) = \frac{2}{\sqrt{\rho}}\Phi^{-1}(2c\cdot(\pi_1 - \frac{1}{2}) + 2\cdot c\cdot n) \ , \tag{C.13}$$

where

$$\Phi(s) = 2[\int_0^s e^{-t^2}dt]/\sqrt{\pi} \ . \tag{C.14}$$

By symmetry, $V_0(\frac{1-\hat{\pi}_u}{2}) = V_1(\frac{1-\hat{\pi}_u}{2})$. Together with equation (C.13), this implies $n = \hat{\pi}_u/2$. To pin down c, notice that when $\pi_1 = 0$, the market maker sets $p = 0$ so that the type-0 insider gets 0 expected payoff. This in turn implies that the type-1 insider gains an infinite expected payoff. Hence, $\Lambda = V_0 - V_1$ goes to negative infinity as π_1 approaches 0. This implies $-c + \hat{\pi}_u c = -1$, i.e., $c = \frac{1}{1-\hat{\pi}_u}$. Thus,

$$\Lambda(\pi_1) = \frac{2}{\sqrt{\rho}}\Phi^{-1}(\frac{2}{1-\hat{\pi}_u}\cdot(\pi_1 - \frac{1-\hat{\pi}_u}{2})) \ . \tag{C.15}$$

We claim that the solution of V_0 is

$$V_0(\pi_1) = \frac{1}{1-\hat{\pi}_u}[\pi_1\Lambda(\pi_1) + \frac{2}{\rho\Lambda'(\pi_1)}] \ . \tag{C.16}$$

To check that this is indeed the solution, we take the derivative on both sides to get

$$V_0^{'}(\pi_1) = \frac{1}{1-\hat{\pi}_u}[\pi_1\Lambda^{'}(\pi_1) + \Lambda(\pi_1) - \frac{2\Lambda^{''}(\pi_1)}{\rho(\Lambda^{'}(\pi_1))^2}] \ . \tag{C.17}$$

From equation (C.12), we get

$$V_0^{'}(\pi_1) = \frac{1}{1-\hat{\pi}_u}\pi_1\Lambda^{'}(\pi_1) \ . \tag{C.18}$$

We differentiate again and divide both sides by $(\Lambda^{'}(\pi_1))^2$ to get

$$\frac{V_0^{''}(\pi_1)}{(\Lambda^{'}(\pi_1))^2} = \frac{1}{1-\hat{\pi}_u}\left[\frac{\pi_1\Lambda^{''}(\pi_1)}{(\Lambda^{'}(\pi_1))^2} + \frac{1}{\Lambda^{'}(\pi_1)}\right] \ . \tag{C.19}$$

We plug the above expression into equation (C.12) to obtain

$$\frac{V_0^{''}(\pi_1)}{(\Lambda^{'}(\pi_1))^2} = \frac{1}{1-\hat{\pi}_u}\frac{\rho}{2}\left[\pi_1\Lambda(\pi_1) + \frac{2}{\rho\Lambda^{'}(\pi_1)}\right] \ . \tag{C.20}$$

By the definition of V_0, equation (C.20) is equivalent to equation (C.10). Moreover, when $\pi_1 \to 0$, we have $\pi_1\Lambda(\pi_1) \to 0$ and $\rho\Lambda^{'}(\pi_1) \to \infty$, which confirms $V_0 \to 0$; when $\pi_1 \to 1-\hat{\pi}_u$, we have $\pi_1\Lambda(\pi_1) \to \infty$, which confirms $V_0 \to \infty$. To check V_0 and V_1 are indeed symmetric, note that

$$V_0(1-\pi_1-\hat{\pi}_u) = \frac{1}{1-\hat{\pi}_u}[(1-\pi_1-\hat{\pi}_u)\Lambda(1-\pi_1-\hat{\pi}_u) + \frac{2}{\rho\Lambda^{'}(1-\pi_1-\hat{\pi}_u)}] \ . \tag{C.21}$$

The symmetry of Λ suggests $\Lambda(1-\pi_1-\hat{\pi}_u) = -\Lambda(\pi_1)$ and $\Lambda^{'}(1-\pi_1-\hat{\pi}_u) = \Lambda^{'}(\pi_1)$.

This, together with equation (C.21), implies

$$V_0(1-\pi_1-\hat{\pi}_u) = \frac{1}{1-\hat{\pi}_u}[\pi_1\Lambda(\pi_1) + \frac{2}{\rho\Lambda'(\pi_1)}] - \Lambda(\pi_1) = V_0(\pi_1) - \Lambda(\pi_1) = V_1(\pi_1) \ . \tag{C.22}$$

Finally, we need to pin down $p(\pi_1)$. Plugging equation (C.9) into equations (C.5) and (C.6), we get

$$p(\pi_1) = \frac{V_0^{'}(\pi_1)}{\Lambda'(\pi_1)} \ , \tag{C.23}$$

which, together with equation (C.18), further implies

$$p(\pi_1) = \frac{\pi_1}{1-\hat{\pi}_u} = \frac{\pi_1}{\pi_0+\pi_1} = \frac{\frac{\pi_1}{\pi_0}}{1+\frac{\pi_1}{\pi_0}} \ . \tag{C.24}$$

It is straightforward to see that $p(\frac{1-\hat{\pi}_u}{2}) = p(\frac{\pi_0+\pi_1}{2}) = \frac{1}{2}$, $p(0) = 0$ and $p(1-\hat{\pi}_u) = p(\pi_0+\pi_1) = 1$. The insider's equilibrium strategies, θ_0 and θ_1, can be pinned down from $\phi = \pi_0\theta_0 + \pi_1\theta_1 = 0$ and $\lambda_1 = \frac{\pi_1[(1-\pi_1)\cdot\theta_1 - \pi_0\cdot\theta_0]}{\sigma^2} = \frac{1}{\Lambda'(\pi_1)}$. Since the existence probability of the insider remains constant, $1-\hat{\pi}_u = \pi_0+\pi_1$. To represent all quantities as functions of $\frac{\pi_1}{\pi_0}$, we may simply substitute $1-\hat{\pi}_u$ to be $\pi_0+\pi_1$ to get the desired result. For example, to get $V_0(\pi_0,\pi_1)$, substitute equation (C.15) into equation (C.16) to get

$$V_0(\pi_0,\pi_1) = \frac{\pi_1\sigma}{(1-\hat{\pi}_u)\sqrt{r}}\Phi^{-1}(\frac{2}{1-\hat{\pi}_u}\pi_0 - 1) + \frac{\sigma}{\sqrt{r}}\Phi'(\Phi^{-1}(\frac{2}{1-\hat{\pi}_u}\pi_0 - 1)) \ .$$

Then since $1-\hat{\pi}_u = \pi_0 + \pi_1$,

$$V_0(\pi_0,\pi_1) = \frac{\frac{\pi_1}{\pi_0}\sigma}{(1+\frac{\pi_1}{\pi_0})\sqrt{r}}\Phi^{-1}(\frac{2}{1+\frac{\pi_1}{\pi_0}}-1) + \frac{\sigma}{\sqrt{r}}\Phi'(\Phi^{-1}(\frac{2}{1+\frac{\pi_1}{\pi_0}}-1)) \ .$$

All other quantities can be calculated following a similar procedure. According to Back and Baruch (2004), when the insider definitely exists, the asset price would converge to 1 (0) if the type-1 (type-0) insider exists with probability 1 as $t \to \infty$. Since the market maker and the insider's strategies are the same along states with the same $\frac{\pi_1}{\pi_0}$, in this model the asset price would also converge to 1 (0) if the type-1 (type-0) insider exists with probability 1 as $t \to \infty$. This indicates the belief convergence result.

To check that the market maker's strategy is optimal, notice that p has no direct impact on the belief-updating process. Since the insider's strategy depends only on the market maker's belief π_0 and π_1, a change in price p does not alter the insider's future buying/selling behavior. Thus, under the optimality condition, the market maker is setting p to maximize his instantaneous payoff. Further notice that the insider's expected trading rate is zero in equilibrium for all t. Thus, in equilibrium the market maker is indifferent to setting the asset price at any level. To check that the insider's strategy is optimal, notice that for any type of insider in state (π_0,π_1) in Π, her value function, the market maker's pricing strategy and the market maker's price updating rule are the same as that type of insider in state $(\frac{\pi_0}{\pi_0+\pi_1}, \frac{\pi_1}{\pi_0+\pi_1})$. Thus, the optimality of the insider's trading strategy stems from the verification theorem when the insider definitely exists.

Uniqueness stems from the zero-sum feature of this game. Consider more general strategy profiles for the market maker and the insider. The market maker's pricing strategy at time t depends on the entire path of Y_t. The insider's strategy at time t depends on the entire history at time t, including the path of Z_t, the pricing history and her order history. An equilibrium is a strategy profile such that (i) given the insider's strategy $\theta^g_{v,t}$, the market maker's pricing strategy p^g_t maximizes his expected payoff; (ii) given the market maker's strategy p^g_t, the insider's strategy $\theta^g_{v,t}$ maximizes her expected payoff. The superscript g indicates that we are now considering this general notion of equilibrium. Note that the Markov equilibrium we obtain also satisfies this equilibrium definition.

Define $\mathcal{V}(\theta^g_{0,t}, \theta^g_{1,t}, p^g_t)$ to be the ex ante expected payoff of the potentially existing insider, given the strategy tuple $(\theta^g_{0,t}, \theta^g_{1,t}, p^g_t)$. Suppose there are two equilibria characterized by $(\theta^g_{0,t}, \theta^g_{1,t}, p^g_t)$ and $(\bar{\theta}^g_{0,t}, \bar{\theta}^g_{1,t}, \bar{p}^g_t)$, respectively. From the market maker's perspective, given the insider's strategy, his equilibrium strategy should weakly dominate all other strategies, so

$$-\mathcal{V}(\theta^g_{0,t}, \theta^g_{1,t}, p^g_t) \geq -\mathcal{V}(\theta^g_{0,t}, \theta^g_{1,t}, \bar{p}^g_t) \ . \tag{C.25}$$

Similarly, from the insider's perspective,

$$\mathcal{V}(\bar{\theta}^g_{0,t}, \bar{\theta}^g_{1,t}, \bar{p}^g_t) \geq \mathcal{V}(\theta^g_{0,t}, \theta^g_{1,t}, \bar{p}^g_t) \ . \tag{C.26}$$

Inequalities (C.25) and (C.26) together imply that

$$\mathcal{V}(\bar{\theta}^g_{0,t}, \bar{\theta}^g_{1,t}, \bar{p}^g_t) \geq \mathcal{V}(\theta^g_{0,t}, \theta^g_{1,t}, p^g_t) \ .$$

A similar argument suggests that

$$\mathcal{V}(\bar{\theta}^g_{0,t}, \bar{\theta}^g_{1,t}, \bar{p}^g_t) \leq \mathcal{V}(\theta^g_{0,t}, \theta^g_{1,t}, p^g_t) \ .$$

The above two inequalities imply that

$$\mathcal{V}(\bar{\theta}^g_{0,t}, \bar{\theta}^g_{1,t}, \bar{p}^g_t) = \mathcal{V}(\theta^g_{0,t}, \theta^g_{1,t}, p^g_t) \ .$$

Therefore, the two equilibria have the same payoff. The Markov equilibrium we proposed also satisfies the general definition of equilibrium. Thus, the equilibrium we proposed in Proposition 23 is essentially unique. □

We hereby provide a direct illustration of the reason why the expected trading rate of the insider always equals zero when the insider definitely exists.

Lemma 10. *Suppose the insider always exists; that is, $\pi_1 + \pi_0 = 1$. There exists no equilibrium such that the insider's expected trading rate is not zero in any state and the trading rates of the type-0 and type-1 insider are continuous.*[1]

Proof. We prove by contradiction. Suppose the claim is false, then there exists a

[1]Notice that it is without loss of generality to focus on the situation where the insider's trading strategy is continuous. Without the continuity assumption, we can always change the insider's trading rate on a zero measure set without affecting the equilibrium.

state π_1^o such that the expected trading rate $\pi_1^o\theta_1 + (1-\pi_1^o)\theta_0 \neq 0$ in state π_1^o. Without loss of generality, suppose $\pi_1^o\theta_1 + (1-\pi_1^o)\theta_0 > 0$. Moreover, in equilibrium, it is impossible that $\theta_1 = \theta_0$ in any state. Otherwise that state is an absorbing state of the market maker's belief and at least one type of insider can deviate to achieve infinite payoff. Suppose $\theta_1 > \theta_0$ in state π_1^o. By continuity there exists an interval $[\underline{\pi}_1, \bar{\pi}_1]$ such that the expected trading rate is positive and $\theta_1 > \theta_0$ in this interval. When the insider definitely exists,

$$d\pi_1 = \lambda_1 \cdot (dY_t - \phi dt)$$

and

$$\lambda_1 = \frac{\pi_1 (1-\pi_1) \cdot (\theta_1 - \theta_0)}{\sigma^2} .$$

Thus, we have $\lambda_1 \in [\underline{\lambda}_1, \bar{\lambda}_1]$ with $\underline{\lambda}_1 > 0$ in this interval. Here we suppress all variables' dependences on π_1 for ease of notation.

Consider the value function of the type-1 insider. Since payoffs in all equilibria are the same, we have $V_1(\underline{\pi}_1) > V_1(\pi_1^o) > V_1(\bar{\pi}_1)$ and $V_1(\underline{\pi}_1) - V_1(\pi_1^o)$ is a constant. We construct a contradiction by showing that there exists a trading strategy for the type-1 insider such that $V_1(\underline{\pi}_1) - V_1(\pi_1^o)$ is arbitrarily small.

First consider the asset price in this interval. Since the market maker's pricing decision does not affect his belief updating process, he chooses the asset price to maximize his instantaneous payoff. Since the expected trading rate is positive in the interval $[\underline{\pi}_1, \bar{\pi}_1]$, the market maker optimally sets the price to $p = 1$. Thus, for the type-1 insider, before the market maker's belief leaving this interval, her

instantaneous payoff is zero at any trading rate. Next consider the type-1 insider's trading problem. In her information set, the belief evolves according to

$$d\pi_1 = \lambda_1 \cdot [(\theta_1 - \phi)dt + \sigma dZ_t] \ .$$

where θ_1 can be picked by her arbitrarily. Moreover, since the market maker cannot detect the insider's deviation, λ_1 and ϕ are only functions of π_1 in equilibrium and do not change with the insider's deviation in her trading rate.

Suppose the market maker's initial belief is $\pi_1 = \pi_1^o$. Consider the type-1 insider's deviation such that $\theta_1 - \phi = \mu\lambda_1$ for an arbitrary μ when $\pi_1 \in [\underline{\pi}_1, \bar{\pi}_1]$. Then according to the type-1 insider, the market maker's belief process is

$$d\pi_1 = \mu\lambda_1^2(\pi_1)dt + \sigma\lambda_1(\pi_1)dZ_t \ .$$

Let the characteristic operator of this process be $\mathcal{A}$; i.e.,

$$\mathcal{A}f(x) = \mu f^{'}(x) + \frac{1}{2}\sigma^2 f^{''}(x)$$

for any twice differentiable function f. Denote $\tau = inf\{t > 0; \pi_{1,t} \notin (\underline{\pi}_1, \bar{\pi}_1)\}$. Consider the function $g(x) = x$. By Dynkin's formula, for an arbitrary positive time T,

$$E(g(\pi_{1,T\wedge\tau})) = g(\pi_1^o) + E(\int_0^{T\wedge\tau} \mathcal{A}g(\pi_{1,s})ds) \ .$$

That is

$$E(\pi_{1,T\wedge\tau}) = \pi_1^o + \mu E(\int_0^{T\wedge\tau} \lambda_1^2(\pi_{1,s})ds) \ .$$

This implies that for all $T > 0$

$$\bar{\pi}_1 \geq \pi_1^o + \mu \underline{\lambda}_1^2 E(T \wedge \tau) \ .$$

Take $T \to \infty$, we get $E(\tau) < \infty$. Thus, $\tau < \infty$ almost surely. Moreover, since

$$\pi_1^o + \mu \bar{\lambda}_1^2 E(\tau) \geq \underline{\pi}_1 \ ,$$

as $\mu \to -\infty$, $E(\tau) \to 0$.

Next we estimate the probability of the market maker's belief hitting each boundary. Consider the following differential equation induced by the operator $\mathcal{A}$:

$$\mathcal{A}f(\pi_1) = \mu \lambda_1^2(\pi_1) f^{'}(\pi_1) + \frac{1}{2}\sigma^2 \lambda_1^2(\pi_1) f^{''}(\pi_1) = 0 \ .$$

Since $\lambda_1 \geq \underline{\lambda}_1 > 0$, this equation is equivalent to

$$\mu f^{'}(\pi_1) + \frac{1}{2}\sigma^2 f^{''}(\pi_1) = 0 \ .$$

One solution of this differential equation is $f(\pi_1) = e^{-\frac{2\mu}{\sigma^2}\pi_1}$. Let $h(\pi_1) = \frac{f(\pi_1) - f(\underline{\pi}_1)}{f(\bar{\pi}_1) - f(\underline{\pi}_1)}$. Then $\mathcal{A}h(\pi_1) = 0$. By Dynkin's formula,

$$E(h(\pi_{1,\tau})) = 1 \cdot P(\pi_{1,\tau} = \bar{\pi}_1) + 0 \cdot P(\pi_{1,\tau} = \underline{\pi}_1) = h(\pi_1^o) \ .$$

Thus,

$$P(\pi_{1,\tau} = \bar{\pi}_1) = \frac{e^{-\frac{2\mu}{\sigma^2}\pi_1^o} - e^{-\frac{2\mu}{\sigma^2}\underline{\pi}_1}}{e^{-\frac{2\mu}{\sigma^2}\bar{\pi}_1} - e^{-\frac{2\mu}{\sigma^2}\underline{\pi}_1}}$$

When $\mu \to -\infty$, easy to see $P(\pi_{1,\tau} = \bar{\pi}_1) \to 0$. Moreover, since $E(\tau) \to 0$ as $\mu \to -\infty$, by Markov inequality, for any $T > 0$, $P(\tau > T) \to 0$ as $\mu \to -\infty$. Thus, as $\mu \to -\infty$, for any $T > 0$,

$$P(\pi_{1,\tau} = \underline{\pi}_1, \tau \leq T) \geq 1 - P(\pi_{1,\tau} = \bar{\pi}_1) - P(\tau > T) \to 1$$

Since the type-1 insider does not receive flow payoff in the interval $[\underline{\pi}_1, \bar{\pi}_1]$ and her value function is non-negative, for an arbitrarily small $T > 0$, the type-1 insider can deviate by picking a large enough μ such that $V_1(\pi_1^o) > (1 - e^{-rT})V_1(\underline{\pi}_1)$. This contradicts the fact that $V_1(\underline{\pi}_1) - V_1(\pi_1^o)$ is a positive constant. □

C.2.3 Proof of Proposition 24

Proof. Notice that for a given $\hat{\pi}_u$, the asset price can be expressed as a function of π_1. Specifically, $p(\pi_1) = \frac{1}{1-\hat{\pi}_u}\pi_1$. Thus, $\lambda_p = \frac{1}{1-\hat{\pi}_u}\lambda_1$, where λ_1 is the sensitivity of π_1 with respect to the order innovation. Then by Proposition 23, we have

$$\lambda_p = \frac{1}{1-\hat{\pi}_u}\lambda_1 = \frac{1}{1-\hat{\pi}_u}\frac{1}{\Lambda'(\pi_1)} = \frac{\sqrt{\rho}}{4} \cdot \Phi'(\Phi^{-1}(\frac{2}{1-\hat{\pi}_u}\pi_1 - 1)) \tag{C.27}$$

Notice that the price sensitivities are the same in states where $\frac{\pi_1}{\pi_0}$ are equal. Then the comparative statics follow directly from the results over λ_1 when the insider exists with probability 1. Specifically, when the insider definitely exists, λ_1 is a concave function of π_1, symmetric around $\pi_1 = \frac{1}{2}$ and achieves maximum at $\pi_1 = \frac{1}{2}$. See Back and Baruch (2004) Theorem 1 for this result. This completely characterizes

$L = \frac{1}{\lambda_p}$. □

C.2.4 Belief Updating with Two Channels of Inside Information

This section derives the belief updating rules for the model in Section 4.4 and shows that the belief processes for the type-1 insider's existence and the type-0 insider's existence are consistent with the belief processes specified in the baseline model.

Lemma 11. *As in the baseline model in Section 4.2, the belief-updating processes* $\pi_{1,t}$ *and* $\pi_{0,t}$ *are characterized by equations* (4.3) *and* (4.2).

Proof. We focus only on the belief updating process $\pi_{1,t}$. The proof for the belief updating process $\pi_{0,t}$ is the same. For ease of notation, we suppress the time dependence of $\pi_{1,t}$ and $\pi_{0,t}$. By filtering theory,

$$d\pi_1^f = \lambda_1^f (dY_t - \phi dt) \ ,$$

where

$$\begin{aligned} \lambda_1^f &= \frac{\pi_1^f[(1-\pi_1^f)\theta_1 - \pi_1^r\theta_1 - \pi_0^f\theta_0 - \pi_0^r\theta_0]}{\sigma^2} \\ &= \frac{\pi_1^f[(1-\pi_1)\theta_1 - \pi_0\theta_0]}{\sigma^2} \ . \end{aligned}$$

Similarly,

$$d\pi_{1,t}^r = \lambda_1^r (dY_t - \phi dt) \ ,$$

where

$$\begin{aligned}\lambda_1^r &= \frac{\pi_1^r[(1-\pi_1^r)\theta_1 - \pi_1^f\theta_1 - \pi_0^f\theta_0 - \pi_0^r\theta_0]}{\sigma^2} \\ &= \frac{\pi_1^r[(1-\pi_1)\theta_1 - \pi_0\theta_0]}{\sigma^2} \ .\end{aligned}$$

Since $\pi_1 = \pi_1^f + \pi_1^r$,

$$d\pi_1 = d\pi_1^f + d\pi_1^r = (\lambda_1^f + \lambda_1^r)(dY_t - \phi dt) \ .$$

Note that

$$\begin{aligned}\lambda_1^f + \lambda_1^r &= \frac{\pi_1^f[(1-\pi_1)\theta_1 - \pi_0\theta_0]}{\sigma^2} + \frac{\pi_1^r[(1-\pi_1)\theta_1 - \pi_0\theta_0]}{\sigma^2} \\ &= \frac{\pi_1[(1-\pi_1)\theta_1 - \pi_0\theta_0]}{\sigma^2}\end{aligned}$$

Thus, in Section 4.4, $d\pi_1 = \lambda_1(dY_t - \phi dt)$. This coincides with the belief-updating rule specified in Section 4.2. □

C.2.5 Proof of Proposition 25

Proof. Notice that in this model, the asset price p and liquidity L depend only on π_1/π_0, the ratio of existence probabilities between the type-1 insider and the type-0 insider. Thus, changes in the price and liquidity after a regulation can be determined

by analyzing the ratio change from

$$\frac{\pi_{1,t}}{\pi_{0,t}} = \frac{\pi^f_{1,t} + \pi^r_{1,t}}{\pi^f_{0,t} + \pi^r_{0,t}}$$

to

$$\frac{\pi_{1,t+}}{\pi_{0,t+}} = \frac{\pi^f_{1,t+} + \pi^r_{1,t+}}{\pi^f_{0,t+} + \pi^r_{0,t+}} = \frac{(1-\Delta_f)\pi^f_{1,t} + (1-\Delta_r)\pi^r_{1,t}}{(1-\Delta_f)\pi^f_{0,t} + (1-\Delta_r)\pi^r_{0,t}} \ .$$

If $\pi^f_{1,t}/\pi^f_{0,t} = \pi^r_{1,t}/\pi^r_{0,t}$, then

$$\frac{\pi_{1,t}}{\pi_{0,t}} = \frac{\pi^f_{1,t} + \pi^r_{1,t}}{\pi^f_{0,t} + \pi^r_{0,t}} = \frac{\pi^f_{1,t}}{\pi^f_{0,t}} = \frac{\pi^r_{1,t}}{\pi^r_{0,t}}$$

and

$$\frac{\pi_{1,t+}}{\pi_{0,t+}} = \frac{(1-\Delta_f)\pi^f_{1,t} + (1-\Delta_r)\pi^r_{1,t}}{(1-\Delta_f)\pi^f_{0,t} + (1-\Delta_r)\pi^r_{0,t}} = \frac{(1-\Delta_f)\pi^f_{1,t}}{(1-\Delta_f)\pi^f_{0,t}} = \frac{(1-\Delta_r)\pi^r_{1,t}}{(1-\Delta_r)\pi^r_{0,t}} \ .$$

Thus, $\pi_{1,t}/\pi_{0,t} = \pi_{1,t+}/\pi_{0,t+}$. This implies $p_{t+} = p_t$ and $L_{t+} = L_t$.

Generally,

$$\begin{aligned}
\frac{\pi_{1,t}}{\pi_{0,t}} > \frac{\pi_{1,t+}}{\pi_{0,t+}} &\iff \frac{\pi^f_{1,t} + \pi^r_{1,t}}{\pi^f_{0,t} + \pi^r_{0,t}} > \frac{(1-\Delta_f)\pi^f_{1,t} + (1-\Delta_r)\pi^r_{1,t}}{(1-\Delta_f)\pi^f_{0,t} + (1-\Delta_r)\pi^r_{0,t}} \ , \\
&\iff (1-\Delta_f)[\pi^r_{1,t}\pi^f_{0,t} - \pi^r_{0,t}\pi^f_{1,t}] > (1-\Delta_r)[\pi^f_{0,t}\pi^r_{1,t} - \pi^f_{1,t}\pi^r_{0,t}] \ , \\
&\iff (\Delta_r - \Delta_f)(\pi^r_{1,t}\pi^f_{0,t} - \pi^r_{0,t}\pi^f_{1,t}) > 0 \ , \\
&\iff (\Delta_r - \Delta_f)(\frac{\pi^r_{1,t}}{\pi^r_{0,t}} - \frac{\pi^f_{1,t}}{\pi^f_{0,t}}) > 0 \ .
\end{aligned}$$

Suppose $\pi^f_{1,t}/\pi^f_{0,t} > \pi^r_{1,t}/\pi^r_{0,t}$. Then $\pi_{1,t}/\pi_{0,t} > \pi_{1,t+}/\pi_{0,t+}$ if and only if $\Delta_f > \Delta_r$,

. This implies $p_t > p_{t+}$. Moreover if $max\{p_{t+}, p_t\} \leq \frac{1}{2}$, $\pi^f_{1,t}/\pi^f_{0,t}$ is closer to 1 if and only if $\Delta_f > \Delta_r$, thus $L_{t+} > L_t$. On the other hand, $p_t < p_{t+}$ if and only if $\Delta_f > \Delta_r$. Thus, if $min\{p_{t+}, p_t\} \geq \frac{1}{2}$, $\pi^f_{1,t}/\pi^f_{0,t}$ is closer to 1 and $L_{t+} > L_t$. The argument for $\pi^f_{1,t}/\pi^f_{0,t} < \pi^r_{1,t}/\pi^r_{0,t}$ is similar. □

C.2.6 Proof of Proposition 26

Proof. For ease of notation, we set $\hat{p} = 0$. The proof for $\hat{p} \in (0, 1)$ is similar. We will also use π and θ, instead of π_1 and θ_1, to refer to the probability that the insider exists and her trading rate, respectively.

Note that $\lambda_1 = \frac{\pi(\theta - \pi\theta)}{\sigma^2}$ and $\phi = \pi\theta$. The belief updating process is $d\pi_t = \lambda_{1,t} \cdot [(\theta_t - \phi_t)dt + \sigma dZ_t]$. Thus, the HJB equation is (using π as the state variable):

$$rV = max_\theta (1 - \pi)\theta + \lambda_1(\theta - \phi)V^{'} + \frac{1}{2}\lambda_1^2\sigma^2 V^{''}. \tag{C.28}$$

Since θ is unbounded and the right hand side is linear in θ, to achieve a finite maximum, we obtain

$$1 - \pi + \lambda_1 V^{'} = 0.$$

Then, the equilibrium should be pinned down by the following equations:

$$rV = -\lambda_1\phi V^{'} + \frac{1}{2}\lambda_1^2\sigma^2 V^{''}, \tag{C.29}$$

$$1 - \pi + \lambda_1 V^{'} = 0\ , \tag{C.30}$$

$$\lambda_1 = \frac{\pi(1-\pi)\theta}{\sigma^2} \ , \tag{C.31}$$

and

$$\phi = \pi\theta \ . \tag{C.32}$$

From equations (C.30), (C.31) and (C.32), we obtain

$$\lambda_1 = -\frac{1-\pi}{V'}$$

and

$$\phi = \frac{\sigma^2\lambda}{1-\pi} = -\frac{\sigma^2}{V'} \ .$$

Plugging the expressions for λ_1 and ϕ into equation (C.29), we obtain

$$rV = -\frac{(1-\pi)\sigma^2}{V'} + \frac{1}{2}\frac{(1-\pi)^2\sigma^2}{(V')^2}V'' \tag{C.33}$$

with boundary conditions $V(0) = \infty$ and $V(1) = 0$.

To simplify the notation, we set $r = \sigma^2 = 1$. This has no effect on the equilibrium existence result. Let $S(t) = V(1-t)$. Then equation (C.33) becomes

$$S = \frac{t}{S'} + \frac{1}{2}\frac{t^2}{(S')^2}S'' \tag{C.34}$$

for $t \in (0,1)$ with boundary conditions $S(0) = 0$ and $S(1) = \infty$. Note that the insider can guarantee herself a non-negative expected payoff and hence $S(t) \geq 0$ for $t \in (0,1)$. Then, if we can find a $k \geq 0$ such that initial conditions $S(0) = 0$ and $S'(0) = k$ induce $S(1) = \infty$, we obtain an equilibrium, and the insider's strategy can

be pinned down by equations (C.29) to (C.32). Transform equation (C.34) into

$$\frac{1}{2}t^2 \cdot S^{''} = S \cdot (S^{'})^2 - t \cdot S^{'} \ . \tag{C.35}$$

First note that $S(t) = t$ is a solution of this ODE with initial condition $S(0) = 0, S^{'}(0) = 1$. But the boundary condition $S(1) = \infty$ does not hold. Next consider the case of $S'(0) = k \in (0,1)$. Then there exists a $t_0 \in (0,1)$, s.t. $S'(t) \in (0,1)$ for $t \in (0,t_0)$. Hence, $S(t) < t$ and $S(t) \cdot S'(t) < t \cdot 1 = t$ for $t \in (0,t_0)$. This implies $S^{''}(t) < 0$ for $t \in (0,t_0)$, which further implies $S(t_0) < t_0$ and $S^{'}(t_0) < 1$. Repeatedly using this argument, we find that $S(t) < t$ for all $t \in (0,1)$. Thus, the boundary condition $S(1) = \infty$ is violated, and $S'(0) = k \in (0,1)$ cannot be the right initial condition. Then consider the case of $S'(0) = k > 1$. Then there exists a $t_0 \in (0,1)$, such that $S'(t) > 1$ for $t \in (0,t_0)$. This implies $S(t) > t$ and $(S^{'})^2 > S^{'}$ in $(0,t_0)$. Then we have

$$\frac{1}{2}t^2 S^{''}(t) > tS^{'}(t)\left[S^{'}(t) - 1\right] > tk(k-1)$$

for $t \in (0,t_0)$, where the second inequality holds because $S^{'}(t) > k$, which stems from $S^{''}(t) > 0$ as a result of the first inequality. Let $C = 2k(k-1)$, then the above inequalities imply that

$$S^{''}(t) > \frac{C}{t}$$

for $t \in (0,t_0)$. Taking the integral of both sides of this inequality with respect to t, we see that that $S(t)$ explodes at 0_+. Finally, consider the case of $S^{'}(0) = 0$. Then by continuity, there exists a $t_0 \in (0,1)$, s.t. $0 < S'(t) < 1$ for $t \in (0,t_0)$. This implies $S(t) < t$ for $t \in (0,t_0)$. Following the same argument for the case of

$S'(0) = k \in (0,1)$ we can see that this ODE cannot explode at $t = 1$, and the boundary condition $S(1) = \infty$ is violated. Therefore, the game does not admit an equilibrium. □

C.2.7 Proof of Proposition 27

Proof. We prove by contradiction. Without loss of generality, consider the value function of the type-1 insider. From the proof of Proposition 26, the expected payoff for the type-1 insider is infinity on the interval $[0,1)$. Moreover, if the market maker thinks that the type-1 insider exists with probability 1, then the type-1 insider's payoff is 0. This discontinuity along the boundary precludes a continuous value function over Π^o.

Consider ball B with center $(0,1)$ and radius δ. By continuity, for a small enough δ, for any $(\pi_0, \pi_1) \in \Pi^o \bigcap B$, $V_1(\pi_0, \pi_1) < \epsilon$. However, for any point $(0, \pi_1^{'}) \in \partial\Pi \bigcap B$, by continuity, we can draw ball $B^{'}$ centered at $(0, \pi_1^{'})$ with a small enough radius such that for any $(\pi_0, \pi_1) \in \Pi^o \bigcap B^{'}$, V_1 is arbitrarily large. This is a contradiction. □

Appendix D

Appendix to Chapter 5

D.1 Proofs

D.1.1 Proofs in Section 5.2

Proof of Lemma 5

Proof. The proof is a standard application of the martingale representation theorem. For any given contract $X = (\alpha, I, \tau)$ and effort process a, define

$$M_t^{1,a} = Y_t^1 - \mu \int_0^t \alpha_s a_s ds$$

and

$$M_t^{0,a} = Y_t^0 - \mu \int_0^t (1-\alpha_s)(1-a_s) ds \ .$$

If the agent follows the effort process a, his lifetime expected payoff, conditional on information at time t, is

$$U_t = \int_0^{t\wedge\tau} e^{-\rho s}(dI_s + \lambda a_s ds) + e^{-\rho t} W_t \ .$$

Let $\tilde{a}$ be an arbitrary effort process. Let $\tilde{U}_t$ denote the agent's lifetime expected payoff conditional on information at time t if he follows $\tilde{a}$ until time t and then reverts to a. Then by the martingale representation theorem, U_t can be written as

$$U_t = U_0 - \int_0^{t\wedge\tau} e^{-\rho s}\beta_{1,s} dM_s^{1,a} + \int_0^{t\wedge\tau} e^{-\rho s}\beta_{0,s} dM_s^{0,a}$$

For each $t \geq 0$,

$$\begin{aligned}
\tilde{U}_t =& U_t + \int_0^{t\wedge\tau} e^{-\rho s}\lambda(\tilde{a}_s - a_s)ds \\
=& U_0 - \int_0^{t\wedge\tau} e^{-\rho s}\beta_{1,s} dM_s^{1,a} + \int_0^{t\wedge\tau} e^{-\rho s}\beta_{0,s} dM_s^{0,a} + \int_0^{t\wedge\tau} e^{-\rho s}\lambda(\tilde{a}_s - a_s)ds \\
=& U_0 - \int_0^{t\wedge\tau} e^{-\rho s}\beta_{1,s} dM_s^{1,\tilde{a}} + \int_0^{t\wedge\tau} e^{-\rho s}\beta_{0,s} dM_s^{0,\tilde{a}} + \int_0^{t\wedge\tau} e^{-\rho s}\lambda(\tilde{a}_s - a_s)ds \\
& - \int_0^{t\wedge\tau} e^{-\rho s}\mu\alpha_s\beta_{1,s}(\tilde{a}_s - a_s)ds - \int_0^{t\wedge\tau} e^{-\rho s}\mu(1-\alpha_s)\beta_{0,s}(\tilde{a}_s - a_s)ds
\end{aligned}$$

Hence, $a_t = 0$ for all t is incentive compatible if and only if the drift term of the above expression is non-positive for any effort process $\tilde{a} \neq 0$; i.e.,

$$\lambda \leq \mu\alpha_t\beta_{1,t} + \mu(1-\alpha_t)\beta_{0,t}$$

for all t before termination. □

D.1.2 Proofs in Section 5.3

Here we provide proofs for Property 2 and Theorem 7 here. Those of all the other properties are straightforward from the text and are therefore omitted.

Proof of Property 2

Proof. Note that the joint value function V must be nondecreasing in continuation value w. This is because in any region where V is strictly decreasing in w, the principal can benefit from paying out to the agent, contradicting the optimality of V. Let $A \subset \mathbb{R}_+$ denote the region of continuation values in which V is strictly increasing. Then the principal does not make any payment when $w \in A$ and $\mathbb{R}_+ \backslash A$ is the payout region. Since $\rho > r$, deferring payment becomes infinitely costly as $w \to +\infty$. Thus the payout region $\mathbb{R}_+ \backslash A$ is nonempty and there exists a $\bar{w} = \inf(\mathbb{R}_+ \backslash A)$.

By construction, V is strictly increasing for $w \in [0, \bar{w}]$ and is constant in a right neighborhood of $\bar{w}$, $(\bar{w}, \bar{w} + \Delta)$. Then, if $V'(\bar{w})$ exists, it must be zero. If $V'(\bar{w})$ does not exist, i.e., the left and the right derivatives are not equal, (5.9) is not defined at $w = \bar{w}$ and the coefficient in front of $V'(w)$ must be zero at $\bar{w}$. Notice that this coefficient is the drift of the continuation value. Hence, when $V'(\bar{w})$ does not exist, $\bar{w}$ is an absorbing payout boundary. As a result, no matter whether $V'(\bar{w})$ exists or not, the second and the third terms on the right-hand side of (5.9) must be zero when $w = \bar{w}$, leading to $V(\bar{w}) = \frac{z}{r} - \frac{\rho - r}{r}\bar{w}$.

By definition, the payout region is a subset of $(\bar{w}, +\infty)$. Actually, the payout region is $(\bar{w}, +\infty)$. Otherwise, there exists an interval $(\bar{w}' - \Delta, \bar{w}') \subset (\bar{w}, +\infty)$ such that V is strictly increasing on $[\bar{w}' - \Delta, \bar{w}']$ and is constant in a right neighborhood of $\bar{w}'$. It must be the case that $\bar{w}' < \infty$, since $\rho > r$ and deferring payment is infinitely costly as $w \to +\infty$. Then a similar argument regarding the existence of $V'(\bar{w})$ also applies here: no matter whether $V'(\bar{w}')$ exists or not, the second and the third terms on the right-hand side of (5.9) must be zero when $w = \bar{w}'$, and thus $V(\bar{w}') = \frac{z}{r} - \frac{\rho - r}{r}\bar{w}' < \frac{z}{r} - \frac{\rho - r}{r}\bar{w} = V(\bar{w})$, a contradiction to the non-decreasing property of V. Hence, the above defined $\bar{w}$ is the payout boundary and the payout region is $(\bar{w}, +\infty)$. As an immediate implication, the optimal payment is $dI^* = (w - \bar{w})^+$ and for $w \in [\bar{w}, +\infty)$, $V(w) = V(\bar{w})$.

The above proof has already shown that either $V'(\bar{w}) = 0$, or $V'(\bar{w})$ does not exist and $\rho\bar{w} - \mu(1 - \alpha^*(\bar{w}))\beta_0^*(\bar{w})$, the drift at $w = \bar{w}$, is 0.

The proof for $\bar{w} \leq \frac{\lambda}{\rho + \mu\bar{\alpha}}$ is straightforward from the text. □

Proof of Theorem 7

Lemma 12. *For any $\bar{w} \in (0, \frac{\lambda}{\rho + \mu\bar{\alpha}}]$, let $\bar{V} \equiv \frac{z}{r} - \frac{\rho - r}{r}\bar{w}$. Then the ODE*

$$rV(w) = \max_{\alpha \in [0, \bar{\alpha}]} z - (\rho - r)w + \rho w V'(w) + (1 - \alpha)\mu[\bar{V} - V(w)] - (\lambda - \mu\alpha w)V'(w) \quad \text{(D.1)}$$

with boundary condition $V(0) = 0$ has a unique solution on $[0, \bar{w}]$.

Proof. For any $w < \frac{\lambda}{\rho+\bar{\alpha}\mu}$, since $\lambda - \mu\alpha w - \rho w > 0$, we can rearrange (D.1) to obtain

$$V^{'} = \max_{\alpha\in[0,\bar{\alpha}]} \frac{z-(\rho-r)w+(1-\alpha)\mu[\bar{V}-V]-rV}{\lambda-\mu\alpha w-\rho w} \ .$$

Let

$$F(w,V) = \max_{\alpha\in[0,\bar{\alpha}]} \frac{z-(\rho-r)w+(1-\alpha)\mu[\bar{V}-V]-rV}{\lambda-\mu\alpha w-\rho w} \ .$$

For any fixed $\epsilon > 0$, for any $(w_1, V_1), (w_2, V_2) \in [0, \frac{\lambda}{\rho+\bar{\alpha}\mu} - \epsilon] \times [0, \bar{V}]$, there exists an M such that $|F(w_1, V_1) - F(w_2, V_2)| \leq M|V_1 - V_2|$. Then, by the Cauchy-Lipschitz theorem, the initial value problem has a unique solution over $[0, \frac{\lambda}{\rho+\bar{\alpha}\mu} - \epsilon]$. Further, notice that V is increasing and upper bounded, and therefore V does not explode as $w \to \bar{w}$. Then the maximum interval of existence reaches the boundary $\bar{w}$ for all $\bar{w} \leq \frac{\lambda}{\rho+\bar{\alpha}\mu}$. When $\bar{w} = \frac{\lambda}{\rho+\bar{\alpha}\mu}$, taking $\epsilon \to 0$, we can extend the solution over $\left[0, \frac{\lambda}{\rho+\bar{\alpha}\mu}\right]$. □

Proposition 41. *Consider two ODEs*

$$rV_1 = max_{\alpha\in[0,\bar{\alpha}]} z-(\rho-r)w+\rho w V_1^{'}+(1-\alpha)\mu[\bar{V}_1-V_1]-(\lambda-\mu\alpha w)V_1^{'}$$

and

$$rV_2 = max_{\alpha\in[0,\bar{\alpha}]} z-(\rho-r)w+\rho w V_2^{'}+(1-\alpha)\mu[\bar{V}_2-V_2]-(\lambda-\mu\alpha w)V_2^{'} \ ,$$

where $\bar{V}_1 = \frac{z}{r} - \frac{\rho-r}{r}\bar{w}_1$, $\bar{V}_2 = \frac{z}{r} - \frac{\rho-r}{r}\bar{w}_2$, $\bar{w}_1 < \bar{w}_2 \leq \frac{\lambda}{\rho+\bar{\alpha}\mu}$; *and* $V_1(0) = V_2(0) = 0$. *Then* $V_1 > V_2$ *for* $w \in (0, \bar{w}_1)$.

Proof. Suppose the opposite holds. Note that $V_1^{'}(0) > V_2^{'}(0)$. Then, there exists a $w \in (0, \bar{w}_1)$ such that $V_1(w) = V_2(w)$. Define $\tilde{w} = \inf\{w \in (0, \bar{w}_1) : V_1(w) = V_2(w)\}$. By the continuity of V_1 and V_2, we have $V_1(\tilde{w}) = V_2(\tilde{w})$. Let α_2 be the α that solves the maximization problem for V_2 at $\tilde{w}$. Taking the difference between the two ODEs at $w = \tilde{w}$, we obtain

$$(\rho\tilde{w} + \mu\alpha_2\tilde{w} - \lambda) \cdot (V_1 - V_2)^{'} + (1 - \alpha_2)\mu(\bar{V}_1 - \bar{V}_2) \leq 0 \ .$$

Since $\alpha_2 < 1$ and $\bar{V}_1 - \bar{V}_2 > 0$,

$$(\rho\tilde{w} + \mu\alpha_2\tilde{w} - \lambda) \cdot (V_1 - V_2)^{'} < 0 \ .$$

Since $\bar{w}_1 < \frac{\lambda}{\rho+\bar{\alpha}\mu}$, $\rho\tilde{w} + \mu\alpha_2\tilde{w} - \lambda < 0$. Thus, $V_1^{'}(\tilde{w}) - V_2^{'}(\tilde{w}) > 0$. Note that this inequality holds whenever $V_1 = V_2$. Since $V_1(w) - V_2(w)$ is continuous and the inequality is strict, it also holds for w close to $\tilde{w}$; i.e., $V_1^{'}(w) - V_2^{'}(w) > 0$ in $(\tilde{w} - \delta, \tilde{w})$ for some $\delta > 0$. By the definition of $\tilde{w}$, $V_1(w) - V_2(w) > 0$ for $w \in (\tilde{w} - \delta, \tilde{w})$. Then, it is impossible to have $V_1(\tilde{w}) = V_2(\tilde{w})$, a contradiction. □

According to the above results, the candidate for the optimal payout boundary is the smallest $\bar{w} \in (0, \frac{\lambda}{\rho+\bar{\alpha}\mu}]$ such that the solution of ODE (5.15) satisfies $V(\bar{w}) = \frac{z}{r} - \frac{\rho-r}{r}\bar{w}$. The existence of such $\bar{w}$ is guaranteed by the continuity of V. Now we are ready to prove Theorem 7.

Proof. Here, we show that the contract that we derive is optimal among all contracts that always implement $a = 0$. The Online Appendix further establishes that such

implementation is indeed optimal for the principal.

Let τ denote the first time that w_t hits zero. We first verify that the principal's value function can be induced by the proposed control processes in Property 5 and the proposed payment process $dI_t = (\beta_0 + w - \bar{w})^+ dY_t^0$. Note that by Property 3, $\beta_0 + w > \bar{w}$, so that $dI_t = (\beta_0 + w - \bar{w}) dY_t^0$. By Ito's Formula for jump processes,

$$\begin{aligned} e^{-r(t\wedge\tau)} B(w_{t\wedge\tau}) =& B(w_0) + \int_0^{t\wedge\tau} e^{-rs}[(\rho w_s - \beta_{0,s}\mu(1-\alpha_s))B^{'}(w_s) - rB(w_s)]ds \\ &+ \int_0^{t\wedge\tau} e^{-rs}[B(\bar{w}) - B(w_s)]dY_s^0 \ . \end{aligned}$$

Under the optimal control processes, the HJB equation becomes

$$rB(w) = z + (\rho w - \beta_0\mu(1-\alpha))B^{'}(w) + (1-\alpha)\mu[B(\bar{w}) - B(w) - (w + \beta_0 - \bar{w})] \ .$$

Thus,

$$\begin{aligned} B(w_0) =& \int_0^{t\wedge\tau} e^{-rs}[z + (1-\alpha_s)\mu(B(\bar{w}) - B(w_s) - (w_s + \beta_{0,s} - \bar{w}))]ds \\ &- \int_0^{t\wedge\tau} e^{-rs}[B(\bar{w}) - B(w_s)]dY_s^0 - e^{-r(t\wedge\tau)} B(w_{t\wedge\tau}) \ . \end{aligned}$$

Due to the fact that $Y_s^0 - (1-\alpha_s)\mu s$ is a martingale and $w_\tau = 0$, letting $t \to \infty$ and taking expectation on the right hand side of the above equation, we obtain

$$B(w_0) = \mathbb{E}(\int_0^\tau e^{-rs}[zds - (w_s + \beta_{0,s} - \bar{w})dY_s^0]) \ ,$$

which verifies that the principal's expected payoff given by (5.1) is indeed achieved

with the proposed control and payment processes.

We then verify that the proposed contract is optimal. Since the cumulative payment process is increasing in time, without loss of generality, we write a general payment process as

$$I_t = I_t^c + I_t^d \ ,$$

where I_t^c is a continuous increasing process and I_t^d includes discrete upward jumps. By Ito's Formula for jump processes,

$$\begin{aligned} e^{-r(t\wedge\tau)}B(w_{t\wedge\tau}) =& B(w_0) + \int_0^{t\wedge\tau} e^{-rs}[(\rho w_s - \beta_{0,s}\mu(1-\alpha_s))B^{'}(w_s) - rB(w_s)]ds \\ & - \int_0^{t\wedge\tau} e^{-rs}B^{'}(w_s)dI_s^c + \int_0^{t\wedge\tau} e^{-rs}[B(w_s+\beta_{0,s}) - B(w_s)]dY_s^0 \\ & + \sum_{s\in[0,t\wedge\tau]} e^{-rs}[B(w_s+\beta_{0,s}\Delta Y_s^0 - \Delta I_s^d) - B(w_s+\beta_{0,s}\Delta Y_s^0)] \ , \end{aligned}$$

where $\Delta Y_s^0 \equiv Y_s^0 - Y_{s^-}^0$. We then rearrange the terms to get

$$\begin{aligned} B(w_0) =& e^{-r(t\wedge\tau)}B(w_{t\wedge\tau}) \\ & + \int_0^{t\wedge\tau} e^{-rs}\{rB(w_s) - (\rho w_s - \beta_{0,s}\mu(1-\alpha_s))B^{'}(w_s) \\ & \qquad - (1-\alpha_s)\mu[B(w+\beta_{0,s}) - B(w)]\}ds \\ & + \int_0^{t\wedge\tau} B^{'}(w_s)e^{-rs}dI_s^c + \int_0^{t\wedge\tau}[B(w+\beta_{0,s}) - B(w)][(1-\alpha_s)\mu ds - dY_s^0] \\ & - \sum_{s\in[0,t\wedge\tau]} e^{-rs}[B(w_s+\beta_{0,s}\Delta Y_s^0 - \Delta I_t^d) - B(w_s+\beta_{0,s}\Delta Y_s^0)] \ . \end{aligned}$$

Taking expectation on both sides and using the fact that $Y_t^0 - \int_0^s (1-\alpha_s)\mu ds$ is a

martingale, we obtain

$$\begin{aligned} B(w_0) =& \mathbb{E}(e^{-r(t\wedge\tau)}B(w_{t\wedge\tau})) + \mathbb{E}\left(\begin{array}{c} \int_0^{t\wedge\tau} e^{-rs}\{rB(w_s) - (\rho w_s - \beta_{0,s}\mu(1-\alpha_s))B^{'}(w_s) \\ -(1-\alpha_s)\mu[B(w+\beta_{0,s}) - B(w)]\}ds \end{array} \right) \\ &+ \mathbb{E}(\int_0^{t\wedge\tau} B^{'}(w_s)e^{-rs}dI_s^c) \\ &- \mathbb{E}(\sum_{s\in[0,t\wedge\tau]} e^{-rs}[B(w_s+\beta_{0,s}\Delta Y_s^0 - \Delta I_t^d) - B(w_s+\beta_{0,s}\Delta Y_s^0)]) \ . \end{aligned}$$

Notice that

$$rB(w) \geq z + (\rho w - \beta_0\mu(1-\alpha))B^{'}(w) + (1-\alpha)\mu[B(\bar{w}) - B(w) - (w+\beta_0-\bar{w})]$$

and for any incentive compatible contract,

$$B(w+\beta_{0,s}) = B(\bar{w}) - (w+\beta_{0,s}-\bar{w}) \ .$$

Moreover, since $B^{'}(w) \geq -1$,

$$B\left(w_0\right) \geq \mathbb{E}(e^{-r(t\wedge\tau)}B(w_{t\wedge\tau})) + \mathbb{E}(\int_0^{t\wedge\tau} ze^{-rs}ds - \int_0^{t\wedge\tau} e^{-rs}dI_s^c) - \mathbb{E}(\sum_{s\in[0,t\wedge\tau]} e^{-rs}\Delta I_t^d) \ .$$

Letting $t\to\infty$ and using the fact that $B(w)$ is bounded, we obtain

$$B(w_0) \geq \mathbb{E}(\int_0^{\tau} e^{-rs}(zds - dI_s)) \ .$$

Therefore, any function satisfying all these conjectured properties is indeed the value

function for the principal. □

D.1.3 Proofs in Section 5.4

Proof of Proposition 28

Proof. Recall from Property 2 that $\bar{w} \leq \frac{\lambda}{\rho+\mu\bar{\alpha}}$. If $w = \bar{w} = \frac{\lambda}{\rho+\mu\bar{\alpha}}$, (5.18) is exactly (5.10), so $\alpha^*(w) = \bar{\alpha}$.

For $w \in \left(0, \frac{\lambda}{\rho+\mu\bar{\alpha}}\right)$, (5.15) is equivalent to

$$V'(w) = \max_{\alpha\in[0,\bar{\alpha}]} \frac{z-(\rho-r)w+(1-\alpha)\mu[V(\bar{w})-V(w)]-rV(w)}{\lambda-\rho w-\mu\alpha w}. \tag{D.2}$$

Let $G(\alpha; w) = \frac{z-(\rho-r)w+(1-\alpha)\mu[V(\bar{w})-V(w)]-rV(w)}{\lambda-\rho w-\mu\alpha w}$, which is obviously continuous in both α and w. Property 5 establishes that the maximizer of the right-hand side (RHS) of (D.2) must be 0 or $\bar{\alpha}$. So to figure out $\alpha^*(w)$, it suffices to compare $G(0; w)$ with $G(\bar{\alpha}; w)$, taking as given $V(0) = 0$ and $V(\bar{w})$.

For $w \in \left(0, \frac{\lambda}{\rho+\mu\bar{\alpha}}\right)$, $G(\bar{\alpha}; w) \geq G(0; w)$ is equivalent to

$$w\left[z-(\rho-r)w-rV\right] \geq \left[\lambda-(\mu+\rho)w\right]\left(V(\bar{w})-V(w)\right). \tag{D.3}$$

Let $\hat{w}_0 = \min\left\{\bar{w}, \frac{\lambda}{\rho+\mu}\right\}$. For any $w \in (0, \hat{w}_0)$, the left-hand side of (D.3) is negative, but its right-hand side is positive. So it fails, establishing the optimality of $\alpha(w) = 0$ in this range.

Now we establish the optimality of $\alpha(w) = \bar{\alpha}$ for w in the vicinity of $\bar{w}$. Note

that by (5.10), (D.3) is equivalent to

$$w\left(\rho-r\right)\left(\bar{w}-w\right)\geq\left[\lambda-\left(\mu+\rho+r\right)w\right]\left(V(\bar{w})-V\left(w\right)\right), \tag{D.4}$$

which holds for all $w\geq\frac{\lambda}{\rho+\mu+r}$. So if $\bar{w}\in(\frac{\lambda}{\rho+\mu+r},\frac{\lambda}{\rho+\mu\bar{\alpha}}]$, $\alpha^{*}\left(w\right)=\bar{\alpha}$ for all $w\in(\frac{\lambda}{\rho+\mu+r},\bar{w}]$.

Note that (D.4) is equivalent to $\frac{\bar{w}-w}{V(\bar{w})-V(w)}\geq\frac{\lambda-(\mu+\rho+r)w}{(\rho-r)w}$. If $\bar{w}\leq\frac{\lambda}{\rho+\mu+r}<\frac{\lambda}{\rho+\mu\bar{\alpha}}$, by Lemma 6 (whose proof does not require Proposition 28), $\bar{w}$ is reflective so that $V'\left(\bar{w}\right)=0$. Then by L'Hôpital's rule, $\lim_{w\to\bar{w}^{-}}\frac{\bar{w}-w}{V(\bar{w})-V(w)}=\lim_{w\to\bar{w}^{-}}\frac{1}{V'(w)}=+\infty$, while $\lim_{w\to\bar{w}^{-}}\frac{\lambda-(\mu+\rho+r)w}{w(\rho-r)}=\frac{\lambda-(\mu+\rho+r)\bar{w}}{(\rho-r)\bar{w}}<+\infty$. Hence, there also exists a $\hat{w}_{\bar{\alpha}}<\bar{w}$, such that $\alpha\left(w\right)=\bar{\alpha}$ for all $w\in(\hat{w}_{\bar{\alpha}},\bar{w}]$.

$\beta_{0}^{*}\left(w\right)$ for $w\in(0,\hat{w}_{0})\cup(\hat{w}_{\bar{\alpha}},\bar{w}]$ results from (5.14). □

Here we provide the closed-form solutions to (5.17) and 5.18). As a first-order linear ODE, (5.17) has general solutions

$$V(w)=\frac{\rho-r}{r+\mu-\rho}(\frac{\lambda}{\rho}-w)+\frac{\mu V(\bar{w})+z-(\rho-r)\frac{\lambda}{\rho}}{r+\mu}+K(\frac{\lambda}{\rho}-w)^{\frac{r+\mu}{\rho}}\ , \tag{D.5}$$

which are all strictly concave in $(0,\bar{w})$. From $V\left(0\right)=0$, we can pin down for $w\in(0,\hat{w}_{0})$ that $K=-\frac{\rho(\rho-r)}{(r+\mu)(r+\mu-\rho)}\cdot\frac{\lambda}{\rho}^{-\frac{r+\mu-\rho}{\rho}}-\frac{\mu V(\bar{w})+z}{r+\mu}\cdot\frac{\lambda}{\rho}^{-\frac{r+\mu}{\rho}}$.

Also as a first-order linear ODE, (5.18) has general solutions

$$\begin{aligned}V(w) &= \frac{\rho-r}{r+(1-\bar{\alpha})\mu-(\rho+\bar{\alpha}\mu)}(\frac{\lambda}{\rho+\mu\bar{\alpha}}-w)+\frac{(1-\bar{\alpha})\mu V(\bar{w})+z-(\rho-r)\frac{\lambda}{\rho+\bar{\alpha}\mu}}{r+\mu(1-\bar{\alpha})}\\ &\quad +K(\frac{\lambda}{\rho+\mu\bar{\alpha}}-w)^{\frac{r+(1-\bar{\alpha})\mu}{\rho+\mu\bar{\alpha}}}\end{aligned} \tag{D.6}$$

if $r+(1-\bar{\alpha})\mu \neq \rho+\bar{\alpha}\mu$, and

$$\begin{aligned} V(w) &= -\frac{\rho-r}{\rho+\mu\bar{\alpha}}(\frac{\lambda}{\rho+\mu\bar{\alpha}}-w)\ln(\frac{\lambda}{\rho+\mu\bar{\alpha}}-w)+\frac{(1-\bar{\alpha})\mu V(\bar{w})+z-(\rho-r)\frac{\lambda}{\rho+\bar{\alpha}\mu}}{r+\mu(1-\bar{\alpha})} \\ &\quad +K(\frac{\lambda}{\rho+\mu\bar{\alpha}}-w) \end{aligned} \tag{D.7}$$

if $r+(1-\bar{\alpha})\mu = \rho+\bar{\alpha}\mu$. It is shown later in the proof of Proposition 29 that the solutions that are increasing in $(0,\bar{w})$ are strictly convex in $(0,\bar{w})$ if $r+(1-\bar{\alpha})\mu < \rho+\bar{\alpha}\mu$ and $K<0$, linear if $r+(1-\bar{\alpha})\mu < \rho+\bar{\alpha}\mu$ and $K=0$, and strictly concave in $(0,\bar{w})$ otherwise.

With the closed-form solutions and their concavity properties discussed above, we show the following proposition:

Proposition 42. *If* $\bar{w} \geq \frac{\lambda}{\rho+\mu+r}$, *then* $\hat{w}_0 = \hat{w}_{\bar{\alpha}}$.

To prove Proposition 42, we first prove Lemma 13, which articulates that the optimal α takes values in $\{0,\overline{\alpha}\}$ almost surely.

Lemma 13. *There does not exist an interval* (w_1,w_2) *such that* $w\cdot V^{'}(w) = V(\bar{w}) - V(w)$ *for all* $w\in(w_1,w_2)$.

Proof. Suppose the contrary. Then $w\cdot V^{'}(w) = V(\bar{w}) - V(w)$ implies

$$V(w) = \frac{c}{w} + V(\bar{w}) \tag{D.8}$$

in (w_1,w_2) for some constant c. Plugging $w\cdot V^{'}(w) = V(\bar{w}) - V(w)$ into the HJB

equation (5.9) we obtain

$$V(w) = \frac{z - (\rho - r)\, w + (\rho + \mu - \lambda/w)\, V(\bar{w})}{r + \rho + \mu - \lambda/w} \ . \tag{D.9}$$

It is straightforward to verify that (D.8) and (D.9) cannot both be satisfied in any interval. □

Lemma 14 shows that the convexity of V in an interval below the payout boundary $\bar{w}$ is "contagion" up to $\bar{w}$.

Lemma 14. *If there exists an interval* $[w_1, w_2) \subset (0, \bar{w})$ *such that* $w_1 \cdot V'(w_1) \geq V(\bar{w}) - V(w_1)$ *and* V *is convex in* (w_1, w_2)*, then* $\alpha^*(w) = \bar{\alpha}$ *for all* $w \in (w_1, \bar{w}]$ *and* V *is convex in* $[w_1, \bar{w}]$*.*

Proof. If V is convex in (w_1, w_2), since V is continuously differentiable in $(0, \bar{w})$, $w \cdot V'(w) + V(w)$ is strictly increasing in $[w_1, w_2)$. Given that $w_1 \cdot V'(w_1) \geq V(\bar{w}) - V(w_1)$, we have $w \cdot V'(w) > V(\bar{w}) - V(w)$ for all $w \in (w_1, w_2)$. So there exists $w_3 \in (w_2, \bar{w})$ such that $w \cdot V'(w) > V(\bar{w}) - V(w)$ for all $w \in (w_1, w_3)$. Iteration of this argument yields $w \cdot V'(w) > V(\bar{w}) - V(w)$ and thus $\alpha^*(w) = \bar{\alpha}$ for all $w \in (w_1, \bar{w})$. By Proposition 28, $\alpha^*(\bar{w}) = \bar{\alpha}$ as well.

Given that $\alpha^*(w) = \bar{\alpha}$ for all $w \in (w_1, \bar{w}]$, the specific solution to (5.18) that matches the value function V in $[w_1, w_2)$ must also match V in $[w_1, \bar{w}]$. Since V is convex in $[w_1, w_2)$, that specific solution must be given by (D.6) with $K \leq 0$ and $r + (1 - \bar{\alpha})\mu < \rho + \bar{\alpha}\mu$. This proves the convexity of V in $[w_1, \bar{w}]$. □

With Lemmas 13 and 14, we can now prove Proposition 42.

Proof. Let $\hat{W} \equiv \left\{w \in (0, \bar{w}) : w \cdot V^{'}(w) = V(\bar{w}) - V(w)\right\}$. We are to show that $\hat{W}$ is a singleton if $\bar{w} \geq \frac{\lambda}{\rho+\mu+r}$. By Proposition 28, $\hat{W}$ is non-empty and has a maximum. Without loss of generality, assume $\hat{w}_{\bar{\alpha}} = \max \hat{W}$. Then V must be strictly concave in $(0, \hat{w}_{\bar{\alpha}}]$. To see this, Lemma 13 and the properties of the general solutions to (5.17) and (5.18) imply that V must be piecewise concave or convex in $(0, \hat{w}_{\bar{\alpha}}]$. If there is an interval $(w_1, w_2) \subset (0, \hat{w}_{\bar{\alpha}}]$ such that V is convex in it, then by Lemma 14, $\alpha^*(w) = \bar{\alpha}$ for all $w \in (w_1, \bar{w}]$, contradicting the fact that $\hat{w}_{\bar{\alpha}} = \max \hat{W}$.

Note that (D.9) holds for $w = \hat{w}_{\bar{\alpha}}$. Plug it into $V^{'}(\hat{w}_{\bar{\alpha}}) = \frac{V(\bar{w})-V(\hat{w}_{\bar{\alpha}})}{\hat{w}_{\bar{\alpha}}}$, we have $V^{'}(\hat{w}_{\bar{\alpha}}) = \frac{\rho-r}{r+\rho+\mu}(1 - \frac{\frac{\lambda}{r+\mu+\rho}-\bar{w}}{\frac{\lambda}{r+\mu+\rho}-\hat{w}_{\bar{\alpha}}})$. Similarly, if there exists $\hat{w}' \in \hat{W}$ such that $\hat{w}' < \hat{w}_{\bar{\alpha}}$, then $V^{'}(\hat{w}') = \frac{\rho-r}{r+\rho+\mu}(1 - \frac{\frac{\lambda}{r+\mu+\rho}-\bar{w}}{\frac{\lambda}{r+\mu+\rho}-\hat{w}'})$. If $\bar{w} \geq \frac{\lambda}{\rho+\mu+r}$, then we have $V'(\hat{w}') \leq V'(\hat{w}_{\bar{\alpha}})$, contradicting the concavity of V in $(0, \hat{w}_{\bar{\alpha}}]$. □

Proof of Proposition 29

Proof. By Proposition 28, $\alpha^*(w) = \bar{\alpha}$ if $w \in (\hat{w}_{\bar{\alpha}}, \bar{w}]$, so here we focus on the solutions to (5.18) when studying the property of the payout boundary $\bar{w}$. Let V_K be the solution with constant K in (D.6) or (D.7).

Case 1: If $r + (1-\bar{\alpha})\mu > \rho + \bar{\alpha}\mu$, we must have $\bar{w} < \frac{\lambda}{\rho+\mu\bar{\alpha}}$, and thus $\bar{w}$ is reflective by Lemma 6. To see this, (D.6) yields

$$V_K^{'} = -\frac{\rho - r}{r + (1-\bar{\alpha})\mu - (\rho + \bar{\alpha}\mu)} - K\frac{r + (1-\bar{\alpha})\mu}{\rho + \mu\bar{\alpha}}(\frac{\lambda}{\rho + \mu\bar{\alpha}} - w)^{\frac{r+(1-\bar{\alpha})\mu}{\rho+\mu\bar{\alpha}}-1}. \quad \text{(D.10)}$$

Since the first term of the right-hand side of (D.10) is negative, K must be negative, otherwise $V_K^{'} < 0$ for all $w \leq \frac{\lambda}{\rho+\mu\bar{\alpha}}$, contradicting the optimality of $\alpha^*(w) = \bar{\alpha}$ for

$w \in (\hat{w}_{\bar{\alpha}}, \bar{w}]$. Since $\frac{r+(1-\bar{\alpha})\mu}{\rho+\mu\bar{\alpha}} - 1 > 0$, V_K is concave. Moreover, as $w \to \frac{\lambda}{\rho+\mu\bar{\alpha}}$, $V_K^{'} \to -\frac{\rho-r}{r+(1-\bar{\alpha})\mu-(\rho+\bar{\alpha}\mu)} < 0$, contradicting the optimality of $\alpha^*(w) = \bar{\alpha}$ for $w \in (\hat{w}_{\bar{\alpha}}, \bar{w}]$ if $\bar{w} = \frac{\lambda}{\rho+\mu\bar{\alpha}}$.

Case 2: If $r + (1-\bar{\alpha})\mu = \rho + \bar{\alpha}\mu$, we have

$$V_K^{'} = -K + \frac{\rho - r}{\rho + \mu\bar{\alpha}} + \frac{\rho - r}{\rho + \mu\bar{\alpha}} \ln(\frac{\lambda}{\rho + \mu\bar{\alpha}} - w) \ .$$

Regardless of the value of K, V_K is concave and as $w \to \frac{\lambda}{\rho+\mu\bar{\alpha}}$, $V_K^{'} \to -\infty$. Thus, analogous to the previous case, it must be that $\bar{w} < \frac{\lambda}{\rho+\mu\bar{\alpha}}$ and $\bar{w}$ is reflective.

Case 3: If $r + (1-\bar{\alpha})\mu < \rho + \bar{\alpha}\mu$, since $-\frac{\rho-r}{r+(1-\bar{\alpha})\mu-(\rho+\bar{\alpha}\mu)} > 0$, K in (D.6) can be either positive or negative. (D.10) yields

$$V_K^{''} = K\frac{r + (1-\bar{\alpha})\mu}{\rho + \mu\bar{\alpha}}(\frac{r + (1-\bar{\alpha})\mu}{\rho + \mu\bar{\alpha}} - 1)(\frac{\lambda}{\rho + \mu\bar{\alpha}} - w)^{\frac{r+(1-\bar{\alpha})\mu}{\rho+\mu\bar{\alpha}}-2} \ .$$

If $K > 0$, since $\frac{r+(1-\bar{\alpha})\mu}{\rho+\mu\bar{\alpha}} - 1 < 0$, $V_K^{''} < 0$ so that V_K is concave. Moreover, as $w \to \frac{\lambda}{\rho+\mu\bar{\alpha}}$, $V_{\bar{\alpha}}^{'} \to -\infty$. Again, it must be that $\bar{w} < \frac{\lambda}{\rho+\mu\bar{\alpha}}$, and $\bar{w}$ is reflective.

If $K = 0$, then $V_K^{'}(w) = -\frac{\rho-r}{r+(1-\bar{\alpha})\mu-(\rho+\bar{\alpha}\mu)} > 0$ for all $w \in (\hat{w}_{\bar{\alpha}}, \bar{w}]$. Thus we must have $\bar{w} = \frac{\lambda}{\rho+\mu\bar{\alpha}}$ as an absorbing state.

If $K < 0$, $V_K^{''} > 0$ so that V_K is strictly convex. Thus, the value function V satisfies $V^{'} > 0$ for all $w < \frac{\lambda}{\rho+\mu\bar{\alpha}}$. This implies that $\bar{w} = \frac{\lambda}{\rho+\mu\bar{\alpha}}$, and $\bar{w}$ is absorbing by Lemma 6.

To summarize all the cases above, we have $\bar{w}$ is absorbing (i.e., $\bar{w} = \frac{\lambda}{\rho+\mu\bar{\alpha}}$) if and only if $r + (1-\bar{\alpha})\mu < \rho + \bar{\alpha}\mu$ and $K \leq 0$; i.e., if and only if V is (weakly) convex in $(\hat{w}_{\bar{\alpha}}, \bar{w})$.

For the "if" claim, it is sufficient to show that when $\bar{\alpha} > \frac{r-\rho+\mu}{2\mu}$ and z is large enough, it must be the case that $K \leq 0$. Suppose the contrary that K is strictly positive for an arbitrarily large z. Let $\hat{w}$ denote the smallest w at which the principal switches from $\alpha = 0$ to $\alpha = \bar{\alpha}$. Then (D.6) implies

$$V(\hat{w}) \leq \frac{\rho - r}{r + (1-\bar{\alpha})\mu - (\rho + \bar{\alpha}\mu)} \frac{\lambda}{\rho + \mu\bar{\alpha}} + \frac{(1-\bar{\alpha})\mu\bar{V} + z - (\rho - r)\frac{\lambda}{\rho+\bar{\alpha}\mu}}{r + \mu(1-\bar{\alpha})} \ . \quad \text{(D.11)}$$

Further notice that from the solution to the ODE with $\alpha = 0$ control,

$$V(\hat{w}) = [1 - (1 - \frac{\rho}{\lambda}\hat{w})^{\frac{r+\mu}{\rho}}][\frac{\rho - r}{(r+\mu)(r+\mu-\rho)}\lambda + \frac{\mu\bar{V} + z}{r+\mu}] - \frac{\rho - r}{r+\mu-\rho}\hat{w} \ . \quad \text{(D.12)}$$

Since $\hat{w} < \bar{w} \leq \frac{\lambda}{\rho+\bar{\alpha}\mu}$, the coefficient in front of $\frac{z}{r}$ on the right-hand side of (D.12) is smaller then 1.[1] On the other hand, the coefficient in front of $\frac{z}{r}$ on the right-hand side of (D.11) is 1. Thus, when z is large enough, (D.11) cannot be satisfied, which is a contradiction. As a result, when z is sufficently large, V is (weakly) convex in $(\hat{w}_{\bar{\alpha}}, \bar{w})$ and $\bar{w}$ is absorbing.

Now we prove the "only if" claim; i.e., $\bar{w}$ is absorbing only if $\bar{\alpha} > \frac{r-\rho+\mu}{2\mu}$ and $z/\lambda \geq \theta(r, \rho, \mu, \bar{\alpha})$, where $\theta(r, \rho, \mu, \bar{\alpha})$ is defined by (D.1.3).

From (D.5), we have

$$V'(\hat{w}) = \frac{\rho - r}{r+\mu-\rho}(1 - \frac{\rho}{\lambda}\hat{w})^{\frac{r+\mu}{\rho}-1} + \frac{\mu V(\bar{w}) + z}{\lambda}(1 - \frac{\rho}{\lambda}\hat{w})^{\frac{r+\mu}{\rho}-1} - \frac{\rho - r}{r+\mu-\rho} \ . \quad \text{(D.13)}$$

On the other hand, by Property 5, we have $V'(\hat{w}) = \frac{V(\bar{w})-V}{\hat{w}}$. Plugging this into

[1] Note that $\bar{V} = \frac{z}{r} - \frac{\rho-r}{r}\bar{w}$.

(5.17), we have

$$V^{'}(\hat{w}) = \frac{\rho - r}{r + \rho + \mu}(1 - \frac{\frac{\lambda}{r+\mu+\rho} - \bar{w}}{\frac{\lambda}{r+\mu+\rho} - \hat{w}}) \ . \tag{D.14}$$

As $\hat{w}$ increases from 0 to $\frac{\lambda}{r+\mu+\rho}$, the right-hand side of (D.13) is decreasing from $\frac{\mu V(\bar{w})+z}{\lambda}$, and that of (D.14) is increasing from $\frac{z-V(\bar{w})}{\lambda}$ to $+\infty$. Thus, there exists a unique $\hat{w} \in (0, \frac{\lambda}{r+\mu+\rho})$ such that both equations hold simultaneously.

Next, we show that V_K is convex if and only if $\hat{w} \geq \frac{\lambda}{2(\rho+\mu\bar{\alpha})}$. Observe that $V(\hat{w})$ should also satisfy (D.6), and thus

$$V^{'}(\hat{w}) = -\frac{\rho - r}{r + (1-\bar{\alpha})\mu - (\rho + \bar{\alpha}\mu)} - K\frac{r + (1-\bar{\alpha})\mu}{\rho + \mu\bar{\alpha}}(\frac{\lambda}{\rho + \mu\bar{\alpha}} - \hat{w})^{\frac{r+(1-\bar{\alpha})\mu}{\rho+\mu\bar{\alpha}}-1} \ . \tag{D.15}$$

We have shown that V_K is convex if and only if $r + (1-\bar{\alpha})\mu < \rho + \bar{\alpha}\mu$ and $K \leq 0$. From (D.14) and (D.15),

$$K \leq 0 \implies \frac{\rho - r}{r + \rho + \mu}(1 - \frac{\frac{\lambda}{r+\mu+\rho} - \bar{w}}{\frac{\lambda}{r+\mu+\rho} - \hat{w}}) \geq -\frac{\rho - r}{r + (1-\bar{\alpha})\mu - (\rho + \bar{\alpha}\mu)},$$

where $\bar{w} = \frac{\lambda}{\rho+\bar{\alpha}\mu}$. This reduces to $\hat{w} \geq \frac{\lambda}{2(\rho+\mu\bar{\alpha})}$. Notice that if $r + (1-\bar{\alpha})\mu < \rho + \bar{\alpha}\mu$, $\frac{\lambda}{2(\rho+\mu\bar{\alpha})} < \frac{\lambda}{r+\mu+\rho}$.

Therefore, V_K is convex only if $r + (1-\bar{\alpha})\mu < \rho + \bar{\alpha}\mu$ (i.e., $\bar{\alpha} > \frac{r-\rho+\mu}{2\mu}$) and the right-hand sides of (D.13) and (D.14) intersect at some $\hat{w} \in [\frac{\lambda}{2(\rho+\mu\bar{\alpha})}, \frac{\lambda}{r+\mu+\rho})$. The

second condition holds if and only if

$$\begin{aligned}&\left(\frac{\rho-r}{r+\mu-\rho}+\frac{\mu\bar{V}+z}{\lambda}\right)(1-\frac{\rho}{\lambda}\cdot\frac{\lambda}{2(\rho+\mu\bar{\alpha})})^{\frac{r+\mu}{\rho}-1}-\frac{\rho-r}{r+\mu-\rho}\\ \geq&\frac{\rho-r}{r+\rho+\mu}(1-\frac{\frac{\lambda}{r+\mu+\rho}-\frac{\lambda}{\rho+\mu\bar{\alpha}}}{\frac{\lambda}{r+\mu+\rho}-\frac{\lambda}{2(\rho+\mu\bar{\alpha})}})\ ,\end{aligned}$$

which is equivalent to

$$\frac{z}{\lambda}\geq\theta\left(r,\rho,\mu,\bar{\alpha}\right)\ .$$

Finally, if $\bar{w}$ is absorbing, $\bar{w}=\frac{\lambda}{\rho+\mu\bar{\alpha}}>\frac{\lambda}{\rho+\mu+r}$, so we have $\hat{w}_0=\hat{w}_{\bar{\alpha}}$ by Proposition 42. (We write $\hat{w}=\hat{w}_0=\hat{w}_{\bar{\alpha}}$ in this case.) □

Appendix E

Appendix to Chapter 6

E.1 Proofs

E.1.1 Proof of Lemma 7

Proof. For $i^n = (\underbrace{i, ..., i}_{n}), s^n = (\underbrace{s, ..., s}_{n})$, by Axiom 3, $A^n(i^n, s^n) \subseteq A^1(i, s)$. By Axiom 4, if aspect $S_j \in A^{n-1}(i^{n-1}, s^{n-1})$, then $S_j \in A^n(i^n, s^n)$. For any $S_j \in A^1(i, s)$, we can apply Axiom 4 $n-1$ times to get $S_j \in A^n(i^n, s^n)$. Thus, $A^1(i, s) \subseteq A^n(i^n, s^n)$. In sum, $A^1(i, s) = A^n(i^n, s^n)$. □

E.1.2 Proof of Theorem 8

Proof. First consider awareness structure $\{\bar{A}^n\}$. It is straightforward to check that given A^1 and $\bar{A}^2$ to $\bar{A}^{n-1}$, $\bar{A}^n$ satisfies all three axioms. Note that given A^1 and any awareness awareness structure, by Axiom 3, $A^n(i^n, s^n) \subseteq A^1(i_1, s_1)$. Thus, $A^n(i^n, s^n) \subseteq \bar{A}^n(i^n, s^n)$.

Then consider awareness structure $\{\underline{A}^n\}$. Easy to see that given A^1, $\underline{A}^2$ satisfies three Axioms. Then check that given A^1 and $\underline{A}^2$ to $\underline{A}^{n-1}$, $\underline{A}^n$ satisfies three Axioms.

Axiom 2: Fix i^n. Let $s_1^n = (s_1, ..., s_n)$, $s_2^n = (s_1, ..., s_n^{'})$, $s^{n-1} = s_1^n \setminus s_n = s_2^n \setminus s_n^{'}$. If

$$P_{A^{n-1}(i^{n-1}, s^{n-1})}(s_n) = P_{A^{n-1}(i^{n-1}, s^{n-1})}(s_n^{'}) \ ,$$

then $s \sim s_n$ if and only if $s \sim s_n^{'}$. This implies $\underline{A}^n(i^n, s_1^n) = \underline{A}^n(i^n, s_2^n)$.

Axiom 3: $\underline{A}^n(i^n, s^n) \subseteq \underline{A}^{n-1}(i^n \setminus i_n, s^n \setminus s_n)$ by definition.

Axiom 4: We first fix some notations. For any $i_1, .., i_{k+1} \in I$, $s_1, .., s_{k+1} \in S$ and $k \geq 2$ let

$$i^{k-1} = \{i_1, ..., i_{k-1}\}, \ s^{k-1} = \{s_1, ..., s_{k-1}\} \ ;$$

$$i_1^k = \{i_1, ..., i_{k-1}, i_k\}, \ s_1^k = \{s_1, ..., s_{k-1}, s_k\} \ ;$$

$$i_2^k = \{i_1, ..., i_{k-1}, i_{k+1}\}, \ s_1^k = \{s_1, ..., s_{k-1}, s_{k+1}\} \ ;$$

$$i^{k+1} = \{i_1, ..., i_k, i_{k+1}\}, \ s^{k+1} = \{s_1, ..., s_k, s_{k+1}\} \ .$$

Fix an arbitrary aspect S_j, suppose $S_j \in \underline{A}^k(i_1^k, s_1^k)$ and $S_j \in \underline{A}^k(i_2^k, s_2^k)$. We need to show that $S_j \in \underline{A}^{k+1}(i^{k+1}, s^{k+1})$. By the definition of $\underline{A}$, $S_j \in \underline{A}^{k-1}(i^{k-1}, s^{k-1})$ and there exists $s \sim s_{k+1}$ such that $S_j \in A^1(i_{k+1}, s)$ where the equivalence relation is induced by $\underline{A}^{k-1}(i^{k-1}, s^{k-1})$. Moreover, since

$$\underline{A}^k(i_1^k, s_1^k) \subseteq \underline{A}^{k-1}(i^{k-1}, s^{k-1}) \ ,$$

$s \sim s_{k+1}$ when the equivalence relation is induced by $\underline{A}^k(i_1^k, s_1^k)$. Since $S_j \in A^1(i_{k+1}, s)$,

we have $S_j \in \underline{A}^{k+1}(i^{k+1}, s^{k+1})$.

Next we check that for any awareness structure satisfying Axioms 2-4, $\underline{A}^n(i^n, s^n) \subseteq A^n(i^n, s^n)$. First consider the case when $n = 2$. If $\underline{A}^2(i^2, s^2) \not\subseteq A^2(i^2, s^2)$, there exists an aspect $S_j \in \mathcal{S}$ and a state $s \sim s_2$ with the equivalence relation induced from $A^1(i_1, s_1)$ such that

$$S_j \in A^1(i_1, s_1) \text{ and } S_j \in A^1((i_2, s)) \ ,$$

but

$$S_j \notin A^2(i^2, s^2) \ .$$

Consider another sequence of states $s_1^2 = (s_1, s)$. Since $s \sim s_2$ with the equivalence relation induced from $A^1(i_1, s_1)$, by Axiom 2, $S_j \notin A^2(i^2, s^2) = A^2(i^2, s_1^2)$. However, by Axiom 4, $S_j \in A^1(i_1, s_1)$ and $S_j \in A^1(i_2, s)$ implies $S_j \in A^2(i^2, s_1^2)$. Contradiction.

Suppose this statement holds for $n = 2, ..., k$. Consider the situation where $n = k + 1$. If $\underline{A}^{k+1}(i^{k+1}, s^{k+1}) \not\subseteq A^{k+1}(i^{k+1}, s^{k+1})$, there exists an aspect $S_j \in \mathcal{S}$ and a state $s \sim s_{k+1}$ with the equivalence relation induced from $\underline{A}^k(i_1^k, s_1^k)$ such that

$$S_j \in \underline{A}^k(i_1^k, s_1^k) \text{ and } S_j \in A^1(i_{k+1}, s) \ ,$$

but

$$S_j \notin A^{k+1}(i^{k+1}, s^{k+1}) \ .$$

Since $\underline{A}^k(i_1^k, s_1^k) \subseteq A^k(i_1^k, s_1^k)$ by assumption,

$$S_j \in A^k(i_1^k, s_1^k) \ .$$

Then by Axiom 4, we must have

$$S_j \notin A^k(i_2^k, s_2^k) \ .$$

Since $\underline{A}^k(i_2^k, s_2^k) \subseteq A^k(i_2^k, s_2^k)$,

$$S_j \notin \underline{A}^k(i_2^k, s_2^k) \ .$$

Since $S_j \in \underline{A}^k(i_1^k, s_1^k)$, by definition, $S_j \in \underline{A}^{k-1}(i^{k-1}, s^{k-1})$. Then by the definition of $\underline{A}$ it must be that for all $s^{'} \sim s_{n+1}$ where the equivalence is induced by $\underline{A}^{k-1}(i^{k-1}, s^{k-1})$, $S_j \notin A^1(i_{k+1}, s^{'})$. Since $\underline{A}^k(i_1^k, s_1^k) \subseteq \underline{A}^{k-1}(i^{k-1}, s^{k-1})$, this contradicts $S_j \in A^1(i_{k+1}, s)$. Thus, it must be that

$$\underline{A}^{k+1}(i^{k+1}, s^{k+1}) \subseteq A^{k+1}(i^{k+1}, s^{k+1}) \ .$$

□

E.1.3 Proof of Proposition 31

Proof. Suppose state $s^{'}$ is excluded by agent i in state s under an awareness structure $\{A^n\}$. Then there exists $n \geq 2$ and an aspect $S_j \in \mathcal{S}$ such that

$$S_j \notin A^n(i^n, s^n) \text{ and } S_j \in A^1(i, s) \ ,$$

where

$$i^n = (\underbrace{i, ..., i}_{n}), s^n = (s, \underbrace{s^{'}, ..., s^{'}}_{n-1}) \ .$$

By Theorem 8, $\underline{A}^n(i^n, s^n) \subseteq A^n(i^n, s^n)$. Thus, state s' would be excluded by agent i in state s under the sophisticated awareness structure $\{\underline{A}^n\}$. □

E.1.4 Proof of Proposition 32

Proof. For $k = 1, ..., j+1$, define i^{n+j-k} and i^{n+j-k} iteratively by

$$i^{n+j-k} = i^{n+j-(k-1)} \setminus i_l \ , \ s^{n+j-k} = s^{n+j-(k-1)} \setminus s_l \ .$$

For the ease of notation, let $P_{A^{n+j}}$ be an abbreviation for $P_{A^{n+j}(i^{n+j}, s^{n+j})}$. By Axiom 3,

$$A^{n+j}(i^{n+j}, s^{n+j}) \subseteq A^{n+j-1}(i^{n+j-1}, s^{n+j-1}) \subseteq A^{n+j-1}(i^{n+j-2}, s^{n+j-2}) \ .$$

This implies

$$\{s | P_{A^{n+j-1}}(s) = P_{A^{n+j-1}}(s_n)\} \supseteq \{s | P_{A^{n+j-2}}(s) = P_{A^{n+j-2}}(s_n)\} \ .$$

Thus,

$$\bigcup_{s: P_{A^{n+j-1}}(s) = P_{A^{n+j-1}}(s_n)} A^1((i_n, s)) \supseteq \bigcup_{s: P_{A^{n+j-2}}(s) = P_{A^{n+j-2}}(s_n)} A^1((i_n, s)) \ . \tag{E.1}$$

Consider $\underline{A}^{n+j}$ and $\underline{A}^{n+j-1}$. By the definition of $\underline{A}^n$,

$$\underline{A}^{n+j}(i^{n+j}, s^{n+j}) = \underline{A}^{n+j-1}(i^{n+j-1}, s^{n+j-1}) \bigcap \bigcup_{s: s \sim s_n} A^1((i_n, s)) \tag{E.2}$$

where the equivalence relation is induced by $P_{\underline{A}^{n+j-1}}$. Moreover,

$$\underline{A}^{n+j-1}(i^{n+j-1}, s^{n+j-1}) = \underline{A}^{n+j-2}(i^{n+j-2}, s^{n+j-2}) \bigcap \bigcup_{s:s\sim s_n} A^1((i_n, s)) \tag{E.3}$$

where the equivalence relation is induced by $P_{\underline{A}^{n+j-2}}$.

Plug equation (E.3) into equation (E.2) and using (E.1), we get

$$\underline{A}^{n+j}(i^{n+j}, s^{n+j}) = \underline{A}^{n+j-1}(i^{n+j-1}, s^{n+j-1}) \ .$$

Using the same argument all the way back to $k = j + 1$, we get

$$\underline{A}^{n+j}(i^{n+j}, s^{n+j}) = \underline{A}^{n}(i^{n}, s^{n}) \ .$$

□

E.1.5 Proof of Corollary 14

Proof. From the first order awareness principle, agent i in state s would exclude state $s^{'}$ if

$$\underline{A}^n(i^n, s^n) \neq A^1(i, s), \text{ where } i^n = (\underbrace{i, ..., i}_{n}), s^n = (s, \underbrace{s^{'}, ..., s^{'}}_{n-1})$$

for any $n \geq 2$ under the sophisticated awareness structure. By Proposition 32,

$$\underline{A}^n(i^n, s^n) = \underline{A}^2(i^2, s^2) \text{ where } i^2 = (i, i), s^2 = (s, s^{'}) \ .$$

□

E.1.6 Proof of Corollary 15

Proof. We prove by induction. First consider the case when $n = 2$. For any $s, s^{'}, s_2 \in S$, let $i^2 = (i, i_2)$, $s_1^2 = (s, s_2)$, $s_2^2 = (s^{'}, s_2)$. Then

$$\underline{A}^2(i^2, s_1^2) = A^1(i, s) \bigcap \bigcup_{\tilde{s} \sim s_2} A^1((i_2, \tilde{s}))$$

where the equivalence relation is induced by $A^1(i, s)$;

$$A^2(i^2, s_2^2) = A^1(i, s^{'}) \bigcap \bigcup_{\tilde{s} \sim s_2} A^1((i_2, \tilde{s}))$$

where the equivalence relation is induced by $A^1(i, s^{'})$. Since $A^1(i, s) = A^1(i, s^{'})$, $\underline{A}^2(i^2, s_1^2) = \underline{A}^2(i^2, s_2^2)$. Take $i_2 = i$, this implies $L_i(s) = L_i(s^{'})$.

Suppose this statement holds for $n = k - 1$. For $n = k$, let $i^k = (i, ..., i_k)$, $s_1^k = (s, ..., s_n)$, $s_2^2 = (s^{'}, ..., s_n)$. Then,

$$\underline{A}^k(i^k, s_1^k) = \underline{A}^{k-1}(i^{k-1}, s_1^{k-1}) \bigcap \bigcup_{\tilde{s}: \tilde{s} \sim s_k} A^1(i_k, \tilde{s})$$

where the equivalence relation is induced by $\underline{A}^{k-1}(i^{k-1}, s_1^{k-1})$;

$$\underline{A}^k(i^k, s_2^k) = \underline{A}^{k-1}(i^{k-1}, s_2^{k-1}) \bigcap \bigcup_{\tilde{s}: \tilde{s} \sim s_k} A^1(i_k, \tilde{s})$$

where the equivalence relation is induced by $\underline{A}^{k-1}(i^{k-1}, s_2^{k-1})$. By the same argument,

$\underline{A}^k(i^k, s_1^k) = \underline{A}^k(i^k, s_2^k)$. Thus, by induction, the statement holds for all n. □

E.1.7 Proof of Theorem 9

By proposition 32, if $A_m^n = \underline{A}_m^n$ for all n, m and permutations of ϕ, considering only $\underline{A}_m^2$ and A_m^1 is sufficient in the first order awareness inference. This can be formalized as follows:

Lemma 15. *Suppose $A_m^n = \underline{A}_m^n$ for all n, m and permutations of ϕ. For any $s \in S$, agent i will exclude state $s^{'} \in S$ at state s if and only if there exists a permutation of $\{\phi_1, ..., \phi_k\}$ and $m = 0, ..., k$ such that*

$$A_m^1(i, s) \neq \underline{A}_m^2(i^2, s^2) \ , \ i^2 = (i, i), s^2 = (s, s^{'}) \ .$$

The proof of lemma 15 since it directly follows from Proposition 32.

Proof. Fix $I^2 = (i, i), S^2 = (s, s^{'})$. Suppose

$$A^1(i, s) = A^1(i, s) \bigcap \bigcup_{\tilde{s}: \tilde{s} \sim s^{'}} A^1(i, \tilde{s}) \ ,$$

where the equivalence relation is induced by $A_k^1(i, s)$. Take an arbitrary permutation of $\{\phi_1, ..., \phi_k\}$. For any m, $A_m^1(i, s) \subseteq A_k^1(i, s)$. Thus,

$$A^1(i, s) = A^1(i, s) \bigcap \bigcup_{\tilde{s}: \tilde{s} \sim s^{'}} A^1(i, \tilde{s}) \ , \tag{E.4}$$

where the equivalence relation is induced by $A_m^1(i, s)$. Take union of both sides of

equation (E.4) with $\bigcup_{d=1}^{m} \phi_d$. By the definition of $\underline{A}_m^2$, we get

$$A_m^1(i,s) = \underline{A}_m^2(i^2, s^2) \ .$$

Since the permutation is arbitrary, we have shown that agent i cannot rule out state $s^{'}$.

For the reverse direction, suppose

$$A^1(i,s) \neq A^1(i,s) \bigcap \bigcup_{\tilde{s}:\tilde{s}\sim s^{'}} A^1(i,\tilde{s}) \tag{E.5}$$

where the equivalence relation is induced by $A_k^1(i,s)$ and agent i cannot exclude state $s^{'}$. By lemma 15, we must have

$$A_k^1(i,s) = \underline{A}_k^2(i^2, s^2) = A_k^1(i,s) \bigcap \bigcup_{\tilde{s}:\tilde{s}\sim s^{'}} A_k^1(i,\tilde{s}) \tag{E.6}$$

where the equivalence relation is also induced by $A_k^1(i,s)$. Notice that $A_k^1(i,s) = A^1(i,s) \bigcup \phi$. Since inequality (E.5) and equation (E.6) are satisfied at the same time, there must exist an aspect $\phi_l \in \phi$ such that

$$\phi_l \in A^1(i,s) \text{ and } \phi_l \notin A^1(i,\tilde{s}) \ , \forall \tilde{s} \sim s^{'}$$

where the equivalence relation is induced by $A_k^1(i,s)$.

Consider a permutation of ϕ where ϕ_l is the last element in the sequence. By

lemma 15,

$$A^1_{k-1}(i,s) = \underline{A}^2_{k-1}(i^2,s^2) \ .$$

Since $\phi_l \in A^1(i,s)$, $\phi_l \in A^1_{k-1}(i,s)$. Thus,

$$A^1_{k-1}(i,s) = A^1_{k-1}(i,s) \bigcup \{\phi_l\} = A^1_k(i,s) \ .$$

Since

$$A^1_{k-1}(i,s) = \underline{A}^2_{k-1}(i^2,s^2) \ ,$$

for any aspect $S_j \in A^1_{k-1}(i,s)$, there exists $\tilde{s} \sim s^{'}$ where the equivalence relation is induced by $A^1_{k-1}(i,s) = A^1_k(i,s)$ such that

$$S_j \in A^1_{k-1}(i,\tilde{s}) \ .$$

Let $S_j = \phi_l$. Since $\phi_l \notin \bigcup_{j=1}^{k-1}\{\phi_j\}$,

$$\phi_l \in A^1_{k-1}(i,\tilde{s}) \implies \phi_l \in A^1(i,\tilde{s}) \ .$$

Thus, there exists $\tilde{s} \sim s^{'}$ where the equivalence relation is induced by $A^1_k(i,s)$ such that $\phi_l \in A^1(i,\tilde{s})$. Contradiction. □

E.1.8 Proof of Corollary 16

Proof. By Proposition 31, we may focus on the situation where $\{A^n_m\} = \{\underline{A}^n_m\}$ for all m and all permutation of elements in ϕ. Fix state s, by Theorem 9, an

awareness signal enables agent i to exclude the maximum number of states when $A^1(i,s)\bigcup\phi=\mathcal{S}$. Since the equivalence relation induced by $\mathcal{S}$ is trivial, agent i can exclude state $s^{'}$ if and only if

$$A^1(i,s)\neq A^1(i,s)\bigcap A^1(i,s^{'})\ ;$$

i.e.,

$$A^1(i,s)\not\subseteq A^1(i,s^{'})\ .$$

In other words, if $A^1(i,s)\subseteq A^1(i,s^{'})$, agent i cannot exclude $s^{'}$ at state s under the sophisticated awareness structure. Then agent i cannot exclude $s^{'}$ at state s under any awareness structure. This also prove the reverse direction of the statement since there always exists an awareness signal ϕ such that $A^1(i,s)\bigcup\phi=\mathcal{S}$. □

Bibliography

Admati, A. R. and Pfleiderer, P. (1991). Sunshine trading and financial market equilibrium. *The Review of Financial Studies*, 4(3):443–481.

Ait-Sahalia, Y. and Saglam, M. (2017a). High frequency market making: Implications for liquidity. *Available at SSRN 2908438*.

Ait-Sahalia, Y. and Saglam, M. (2017b). High frequency market making: Optimal quoting. *Available at SSRN 2331613*.

Allen, F. and Gale, D. (1992). Stock-price manipulation. *The Review of Financial Studies*, 5(3):503–529.

Alonso, R. and Câmara, O. (2018). On the value of persuasion by experts. *Journal of Economic Theory*, 174:103–123.

Anderson, A. and Smith, L. (2013). Dynamic deception. *The American Economic Review*, 103(7):2811–2847.

Aumann, R. J., Maschler, M., and Stearns, R. E. (1995). *Repeated games with incomplete information*. MIT press.

Avery, C. and Zemsky, P. (1998). Multidimensional uncertainty and herd behavior in financial markets. *American economic review*, pages 724–748.

Back, K. and Baruch, S. (2004). Information in securities markets: Kyle meets glosten and milgrom. *Econometrica*, 72(2):433–465.

Back, K. and Baruch, S. (2013). Strategic liquidity provision in limit order markets. *Econometrica*, 81(1):363–392.

Back, K., Crotty, K., and Li, T. (2017). Identifying information asymmetry in securities markets. *The Review of Financial Studies*, 31(6):2277–2325.

Baldauf, M. and Mollner, J. (2019). High-frequency trading and market performance. *Journal of Finance*.

Banerjee, S. and Breon-Drish, B. (2020a). Strategic trading and unobservable information acquisition. *Journal of Financial Economics*, 138(2):458–482.

Banerjee, S. and Breon-Drish, B. M. (2020b). Dynamics of research and strategic trading. *Available at SSRN 2846059*.

Baron, M., Brogaard, J., Hagströmer, B., and Kirilenko, A. (2018). Risk and return in high-frequency trading. *Journal of financial and quantitative analysis*, pages 1–32.

Baruch, S. and Glosten, L. R. (2019). Tail expectation and imperfect competition in limit order book markets. *Journal of Economic Theory*, 183:661–697.

Basak, S. and Chabakauri, G. (2010). Dynamic mean-variance asset allocation. *The Review of Financial Studies*, 23(8):2970–3016.

Battalio, R., Ellul, A., and Jennings, R. (2007). Reputation effects in trading on the new york stock exchange. *Journal of Finance*, 62(3):1243–1271.

Bell, H. and Searles, H. (2014). An analysis of global hft regulation: Motivations, market failures, and alternative outcomes.

Bertsimas, D. and Lo, A. W. (1998). Optimal control of execution costs. *Journal of Financial Markets*, 1(1):1–50.

Bessembinder, H., Hao, J., and Zheng, K. (2019). Liquidity provision contracts and market quality: Evidence from the new york stock exchange. *The Review of Financial Studies*.

Bessembinder, H., Jacobsen, S., Maxwell, W., and Venkataraman, K. (2018). Capital commitment and illiquidity in corporate bonds. *The Journal of Finance*, 73(4):1615–1661.

Biais, B., Foucault, T., and Moinas, S. (2015). Equilibrium fast trading. *Journal of Financial economics*, 116(2):292–313.

Biais, B., Mariotti, T., Plantin, G., and Rochet, J.-C. (2007). Dynamic security design: Convergence to continuous time and asset pricing implications. *The Review of Economic Studies*, 74(2):345–390.

Biais, B., Mariotti, T., Rochet, J.-C., and Villeneuve, S. (2010). Large risks, limited liability, and dynamic moral hazard. *Econometrica*, 78(1):73–118.

Biais, B., Martimort, D., and Rochet, J.-C. (2000). Competing mechanisms in a common value environment. *Econometrica*, 68(4):799–837.

Black, F. (1995). Equilibrium exchanges. *Financial Analysts Journal*, 51(3):23–29.

Boehmer, E., Fong, K., and Wu, J. (2018). International evidence on algorithmic trading.

Brogaard, J. and Garriott, C. (2019). High-frequency trading competition. *Journal of Financial and Quantitative Analysis*, 54(4):1469–1497.

Brogaard, J., Hagströmer, B., Nordén, L., and Riordan, R. (2015). Trading fast and slow: Colocation and liquidity. *The Review of Financial Studies*, 28(12):3407–3443.

Brogaard, J., Hendershott, T., and Riordan, R. (2014). High-frequency trading and price discovery. *The Review of Financial Studies*, 27(8):2267–2306.

Brunnermeier, M. K. and Pedersen, L. H. (2005). Predatory trading. *The Journal of Finance*, 60(4):1825–1863.

Brunnermeier, M. K. and Pedersen, L. H. (2008). Market liquidity and funding liquidity. *The review of financial studies*, 22(6):2201–2238.

Budish, E., Cramton, P., and Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, 130(4):1547–1621.

Budish, E., Lee, R., and Shim, J. (2019). A theory of stock exchange competition and innovation: Will the market fix the market? *NBER Working Paper*, 25855.

Carlin, B. I., Lobo, M. S., and Viswanathan, S. (2007). Episodic liquidity crises: Cooperative and predatory trading. *The Journal of Finance*, 62(5):2235–2274.

Chakraborty, A. and Yilmaz, B. (2004a). Informed manipulation. *Journal of Economic Theory*, 114(1):132 – 152.

Chakraborty, A. and Yilmaz, B. (2004b). Manipulation in market order models. *Journal of Financial Markets*, 7(2):187 – 206.

Chakravarty, S. and Holden, C. W. (1995). An integrated model of market and limit orders. *Journal of Financial Intermediation*, 4(3):213–241.

Che, Y.-K. and Mierendorff, K. (2019). Optimal dynamic allocation of attention. *American Economic Review*, 109(8):2993–3029.

Chen, M., Sun, P., and Xiao, Y. (2020). Optimal monitoring schedule in dynamic contracts. *Operations Research*, 68(5):1285–1314.

Chiyachantana, C. N., Jiang, C. X., Taechapiroontong, N., and Wood, R. A. (2004). The impact of regulation fair disclosure on information asymmetry and trading: An intraday analysis. *Financial Review*, 39(4):549–577.

Chordia, T., Roll, R., and Subrahmanyam, A. (2011). Recent trends in trading activity and market quality. *Journal of Financial Economics*, 101(2):243–263.

Cipriani, M. and Guarino, A. (2014). Estimating a structural model of herd behavior in financial markets. *American Economic Review*, 104(1):224–51.

Clark-Joseph, A. D., Ye, M., and Zi, C. (2017). Designated market makers still matter: Evidence from two natural experiments. *Journal of Financial Economics*, 126(3):652–667.

Comerton-Forde, C., Hendershott, T., Jones, C. M., Moulton, P. C., and Seasholes, M. S. (2010). Time variation in liquidity: The role of market-maker inventories and revenues. *The Journal of Finance*, 65(1):295–331.

Comerton-Forde, C., Putniņš, T. J., and Tang, K. M. (2011). Why do traders choose to trade anonymously? *Journal of Financial and Quantitative Analysis*, pages 1025–1049.

Conrad, J., Wahal, S., and Xiang, J. (2015). High-frequency quoting, trading, and the efficiency of prices. *Journal of Financial Economics*, 116(2):271–291.

Conrad, J. S. and Wahal, S. (2018). The term structure of liquidity provision. *Available at SSRN 2837111*.

Crippen, A. (2014). Cnbc transcript: Warren buffett, charlie munger and bill gates.

Dai, L., Wang, Y., and Yang, M. (2020a). Dynamic contracting with flexible monitoring. *Available at SSRN 3496785*.

Dai, L., Wang, Y., and Yang, M. (2020b). Insider trading when there may not be an insider. *Available at SSRN 2720736*.

Dekel, E., Lipman, B. L., and Rustichini, A. (1998). Standard state-space models preclude unawareness. *Econometrica*, 66(1):159–173.

DeMarzo, P. M. and Fishman, M. J. (2007). Optimal long-term financial contracting. *The Review of Financial Studies*, 20(6):2079–2128.

DeMarzo, P. M. and Sannikov, Y. (2006). Optimal security design and dynamic capital structure in a continuous-time agency model. *The Journal of Finance*, 61(6):2681–2724.

Dennis, P. J. and Sandås, P. (2020). Does trading anonymously enhance liquidity? *Journal of Financial and Quantitative Analysis*, 55(7):2372–2396.

Du, S. and Zhu, H. (2017). What is the optimal trading frequency in financial markets? *The Review of Economic Studies*, 84(4):1606–1651.

Dunning, D. (2011). The dunning–kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology*, volume 44, pages 247–296. Elsevier.

Eleswarapu, V. R., Thompson, R., and Venkataraman, K. (2004). The impact of regulation fair disclosure: Trading costs and information asymmetry. *Journal of Financial and Quantitative Analysis*, 39(2):209–225.

Eurex. (2016). Hft act: Amendments to the calculation of excessive system usage fee.

Eurex. (2019). Enhancement of the excessive system usage concept: Introduction of a new limit type.

Fagin, R. and Halpern, J. Y. (1988). Belief, awareness, and limited reasoning. *Artificial intelligence*, 34(1):39–76.

Fardeau, V. (2020). Strategic trading around anticipated supply/demand shocks. *Available at https://www.dropbox.com/s/3wn23fqsr4dcrb5/DST12.pdf?dl=0*.

Foucault, T., Hombert, J., and Roşu, I. (2016). News trading and speed. *The Journal of Finance*, 71(1):335–382.

Foucault, T., Kadan, O., and Kandel, E. (2005). Limit order book as a market for liquidity. *The review of financial studies*, 18(4):1171–1217.

Foucault, T., Moinas, S., and Theissen, E. (2007). Does anonymity matter in electronic limit order markets? *The Review of Financial Studies*, 20(5):1707–1747.

Friederich, S. and Payne, R. (2014). Trading anonymity and order anticipation. *Journal of Financial Markets*, 21:1–24.

Frino, A. and Jones, S. (2005). The impact of mandated cash flow disclosure on bid-ask spreads. *Journal of Business Finance & Accounting*, 32(7-8):1373–1396.

Galanis, S. (2015). The value of information under unawareness. *Journal of Economic Theory*, 157:384–396.

Galanis, S. (2016). The value of information in risk-sharing environments with unawareness. *Games and Economic Behavior*, 97:1–18.

Galperti, S. (2019). Persuasion: The art of changing worldviews. *American Economic Review*.

Gentzkow, M. and Kamenica, E. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Gentzkow, M. and Kamenica, E. (2016). Competition in persuasion. *The Review of Economic Studies*, 84(1):300–322.

Georgiadis, G. and Szentes, B. (2020). Optimal monitoring design. *Econometrica*, 88(5):2075–2107.

Glebkin, S., Malamud, S., and Teguia, A. (2020). Asset prices and liquidity with market power and non-gaussian payoffs. *Swiss Finance Institute Research Paper*, (20-80).

Glosten, L. R. (1989). Insider trading, liquidity, and the role of the monopolist specialist. *Journal of Business*, pages 211–235.

Glosten, L. R. (1994). Is the electronic open limit order book inevitable? *The Journal of Finance*, 49(4):1127–1161.

Goettler, R. L., Parlour, C. A., and Rajan, U. (2009). Informed traders and limit order markets. *Journal of Financial Economics*, 93(1):67–87.

Grant, J. and Demos, T. (2011). Institutional investors air hft concerns institutional investors air hft concerns.

Gromb, D. and Vayanos, D. (2002). Equilibrium and welfare in markets with financially constrained arbitrageurs. *Journal of financial Economics*, 66(2-3):361–407.

Hachmeister, A. and Schiereck, D. (2010). Dancing in the dark: post-trade anonymity, liquidity and informed trading. *Review of Quantitative Finance and Accounting*, 34(2):145–177.

Hagerman, R. L. and Healy, J. P. (1992). The impact of sec-required disclosure and insider-trading regulations on the bid/ask spreads in the over-the-counter market. *Journal of Accounting and Public Policy*, 11(3):233–243.

Hameed, A., Kang, W., and Viswanathan, S. (2010). Stock market declines and liquidity. *The Journal of Finance*, 65(1):257–293.

Han, J., Khapko, M., and Kyle, A. (2014). Liquidity with high-frequency market making.

Hasbrouck, J. and Saar, G. (2013). Low-latency trading. *Journal of Financial Markets*, 16(4):646–679.

He, Z. (2012). Dynamic compensation contracts with private savings. *The Review of Financial Studies*, 25(5):1494–1549.

Heifetz, A., Meier, M., and Schipper, B. C. (2006). Interactive unawareness. *Journal of economic theory*, 130(1):78–94.

Heifetz, A., Meier, M., and Schipper, B. C. (2013). Unawareness, beliefs, and speculative trade. *Games and Economic Behavior*, 77(1):100–121.

Hendershott, T., Jones, C. M., and Menkveld, A. J. (2011a). Does algorithmic trading improve liquidity? *The Journal of Finance*, 66(1):1–33.

Hendershott, T., Jones, C. M., and Menkveld, A. J. (2011b). Does algorithmic trading improve liquidity? *The Journal of Finance*, 66(1):1–33.

Hendershott, T. and Riordan, R. (2013). Algorithmic trading and the market for liquidity. *Journal of Financial and Quantitative Analysis*, 48(4):1001–1024.

Hirschey, N. (2018). Do high-frequency traders anticipate buying and selling pressure? *Available at SSRN 2238516*.

Hirschey, N. (2020). Do high-frequency traders anticipate buying and selling pressure? *Working Paper*.

Hu, E. (2019). Intentional access delays, market quality, and price discovery: Evidence from iex becoming an exchange. *Available at SSRN 3195001*.

Kacperczyk, M. and Pagnotta, E. S. (2019). Chasing private information. *The Review of Financial Studies*.

Karni, E. and Vierø, M.-L. (2013). "reverse bayesianism": A choice-based theory of growing awareness. *The American Economic Review*, 103(7):2790–2810.

Karni, E. and Vierø, M.-L. (2017). Awareness of unawareness: a theory of decision making in the face of ignorance. *Journal of Economic Theory*, 168:301–328.

Kirilenko, A., Kyle, A. S., Samadi, M., and Tuzun, T. (2017). The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998.

Korajczyk, R. A. and Murphy, D. (2019a). Do high-frequency traders improve your implementation shortfall? *Journal of Investment Management*, 18:18–33.

Korajczyk, R. A. and Murphy, D. (2019b). High-frequency market making to large institutional trades. *The Review of Financial Studies*, 32(3):1034–1067.

Kreps, D. M. and Scheinkman, J. A. (1983). Quantity precommitment and bertrand competition yield cournot outcomes. *The Bell Journal of Economics*, pages 326–337.

Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121.

Kuvalekar, A. and Ravi, N. (2019). Rewarding failure. *Available at SSRN 3281644*.

Kyle, A. S. (1985a). Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, pages 1315–1335.

Kyle, A. S. (1985b). Continuous auctions and insider trading. *Econometrica*, 53(6):1315–35.

Kyle, A. S. (1989). Informed speculation with imperfect competition. *The Review of Economic Studies*, 56(3):317–355.

Kyle, A. S. and Obizhaeva, A. A. (2016). Market microstructure invariance: Empirical hypotheses. *Econometrica*, 84(4):1345–1404.

Kyle, A. S. and Viswanathan, S. (2008). How to define illegal price manipulation. *American Economic Review*, 98(2):274–79.

Kyle, A. S. and Xiong, W. (2001). Contagion as a wealth effect. *The Journal of Finance*, 56(4):1401–1440.

Li, A. and Yang, M. (2019). Optimal incentive contract with endogenous monitoring technology. *Theoretical Economics*, forthcoming.

Li, J. (2009). Information structures with unawareness. *Journal of Economic Theory*, 144(3):977–993.

Li, S., Wang, X., and Ye, M. (2020). Who provides liquidity, and when? *NBER Working Paper*, (w25972).

Mayskaya, T. (2020). Dynamic choice of information sources. *California Institute of Technology Social Science Working Paper*.

Meling, T. (2020). Anonymous trading in equities. *The Journal of Finance*, page forthcoming.

Menkveld, A. J. (2016). The economics of high-frequency trading: Taking stock. *Annual Review of Financial Economics*, 8:1–24.

Modica, S. and Rustichini, A. (1994). Awareness and partitional information structures. *Theory and decision*, 37(1):107–124.

Myerson, R. B. (2015). Moral hazard in high office and the dynamics of aristocracy. *Econometrica*, 83(6):2083–2126.

Nikandrova, A. and Pancs, R. (2018). Dynamic project selection. *Theoretical Economics*, 13(1):115–143.

Obizhaeva, A. A. and Wang, J. (2013). Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*, 16(1):1–32.

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, 116(2):257–270.

O'Hara, M., Yao, C., and Ye, M. (2014). What's not there: Odd lots and market data. *The Journal of Finance*, 69(5):2199–2236.

Orlov, D. (2018). Frequent monitoring in dynamic contracts. Technical report, working paper.

Parlour, C. A. and Seppi, D. J. (2003). Liquidity-based competition for order flow. *The Review of Financial Studies*, 16(2):301–343.

Pham, T. P., Swan, P. L., and Westerholm, P. J. (2016). Intra-day revelation of counterparty identity in the world's best-lit market. In *28th Australasian Finance and Banking Conference*.

Piskorski, T. and Westerfield, M. M. (2016). Optimal dynamic contracts with moral hazard and costly monitoring. *Journal of Economic Theory*, 166:242–281.

Pritsker, M. (2009). Large investors: Implications for equilibrium asset returns, shock absorption, and liquidity.

Rindi, B. (2008). Informed traders as liquidity providers: Anonymity, liquidity and price formation. *Review of Finance*, 12(3):497–532.

Rostek, M. and Weretka, M. (2015). Dynamic thin markets. *The Review of Financial Studies*, 28(10):2946–2992.

Roşu, I. (2009). A dynamic model of the limit order book. *The Review of Financial Studies*, 22(11):4601–4641.

Sannikov, Y. (2008). A continuous-time version of the principal-agent problem. *The Review of Economic Studies*, 75(3):957–984.

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

Shim, J. J. and Wang, Y. (2021). Do electronic markets improve execution if you cannot identify yourself? *Available at SSRN 3805830*.

Sidhu, B., Smith, T., Whaley, R. E., and Willis, R. H. (2008). Regulation fair disclosure and the cost of adverse selection. *Journal of Accounting Research*, 46(3):697–728.

Simaan, Y., Weaver, D. G., and Whitcomb, D. K. (2003). Market maker quotation behavior and pretrade transparency. *The Journal of Finance*, 58(3):1247–1267.

Smolin, A. (2017). Dynamic evaluation design. *Available at SSRN 3051703*.

Sun, P. and Tian, F. (2017). Optimal contract to induce continued effort. *Management Science*, 64(9):4193–4217.

Van Kervel, V. (2015). Competition for order flow with fast and slow traders. *The Review of Financial Studies*, 28(7):2094–2127.

Van Kervel, V. and Menkveld, A. J. (2019). High-frequency trading around large institutional orders. *The Journal of Finance*.

Varas, F., Marinovic, I., and Skrzypacz, A. (2020). Random inspections and periodic reviews: Optimal dynamic monitoring. *The Review of Economic Studies*, 87(6):2893–2937.

Vayanos, D. (2001). Strategic trading in a dynamic noisy market. *The Journal of Finance*, 56(1):131–171.

Venkataraman, K. and Waisburd, A. C. (2007). The value of the designated market maker. *Journal of Financial and Quantitative Analysis*, pages 735–758.

Viswanathan, S. and Wang, J. J. (2002). Market architecture: limit-order books versus dealership markets. *Journal of Financial Markets*, 5(2):127–167.

Wang, Y. (2021a). High-frequency trading, endogenous capital commitment and market quality. *Available at SSRN 3470187*.

Wang, Y. (2021b). Learning from awareness. *Available at SSRN 3470204*.

Weill, P.-O. (2007). Leaning against the wind. *The Review of Economic Studies*, 74(4):1329–1354.

Yang, L. and Zhu, H. (2020). Back-running: Seeking and hiding fundamental information in order flows. *The Review of Financial Studies*, 33(4):1484–1533.

Yao, C. and Ye, M. (2018). Why trading speed matters: A tale of queue rationing under price controls. *The Review of Financial Studies*, 31(6):2157–2183.