Discussion Papers
Department of Economics
University of Copenhagen

No. 15-04

Dealing with Dynamic Agency

Benjamin Falkeborg

Øster Farimagsgade 5, Building 26, DK-1353 Copenhagen K., Denmark
Tel.: +45 35 32 30 01 – Fax: +45 35 32 30 00
http://www.econ.ku.dk

ISSN: 1601-2461 (E)

Dealing with Dynamic Agency*

Benjamin Falkeborg†
Department of Economics
University of Copenhagen

February 27, 2015

Abstract

I study the implications of agency frictions for the pricing policy of institutional market makers. In a setting where a market maker cannot observe the actions of an employed trader, I derive the optimal compensation structure and pricing policy. The theory demonstrates that incentive contracting and the price for immediacy are inherently linked. When the trader's compensation is optimally deferred according to order flow, market making efficiency is improved and the quoted spreads are minimized. In other words, optimizing trader compensation leads to a liquidity gain.

Keywords: Market Making, Hedging, Dynamic Moral Hazard, Recursive Contracts, Liquidity Provision.

JEL-Classifications: D81, D86, G12, J33

*This paper is a revised version of a chapter of my Ph.D. dissertation. Part of the research was conducted while visiting Massachusetts Institute of Technology, Department of Economics, whose hospitality is gratefully acknowledged. I would like to express my deepest gratitude to my advisor Peter Norman Sørensen, for his guidance and the many hours he spent discussing this paper. I would also like to thank Alexander Sebald, Bengt Holmström, Marco Ottaviani, Alexander Sokol for discussions and useful comments. Financial support from The University of Copenhagen, Faculty of Social Sciences is gratefully acknowledged. All errors are mine.

†Email: benjamin.falkeborg@econ.ku.dk

1 Introduction

The recent trading losses incurred by dealer banks and the implementation of the "Volcker Rule", has led to a renewed interest in the pricing and risk management policies of institutional market makers. Despite the important role played by these financial institutions, the internal organization of market maker firms has been largely ignored by the literature. The theoretical market microstructure literature pioneered by Garman (1976) has so far focused on four sources of frictions: inventory effects, fixed costs, asymmetric information and more recently search and bargaining.[1] The present paper studies an important yet unexplored source of frictions: internal agency problems.

An institutional market maker provides immediacy by absorbing a customer's demand and supply of assets into its own inventory. In order to make a profit on the bid-offer spread the market maker must successfully manage the risk associated with the transactions. Traders manage the risk by analyzing the trade to identify what instruments may be used as a hedge, and dynamically adjust the hedge until the risk is laid off. Managing residual risk from market making therefore requires consistent effort from traders. While profits and losses are perfectly observable, the underlying trading effort driving profits and losses is to a large extent unobservable. This creates a moral hazard problem.

This paper studies how an institutional market maker must design trader compensation in order to mitigate agency problems. The paper addresses the following questions: What is the best way to design incentives for traders in market maker firms? How is incentive compensation connected with the pricing and risk management policy of the market maker? What are the implications of optimal trader compensation for market liquidity?

In order to speak to these issues, I build a continuous-time model of a market maker concerned with managing risk from transactions in the presence of internal agency frictions. While it is intuitive that agency conflicts can be mitigated by optimizing the compensation practice, the effect of incentive compensation on the market maker's pricing policy is less obvious. It turns out that the optimal contract has striking implications for the market maker's pricing of risk. In particular, the model shows that deferring the trader's compensation optimally

[1] See O'Hara (1995) for a survey of several theoretical models. Duffie, Gârleanu, and Pedersen (2005) were the first to model trading frictions using search theory.

according to order flow, minimizes the spreads quoted by the market maker. Thus, optimizing trader compensation leads to a liquidity gain.

I consider the following setup: A risk neutral market maker takes positions dictated by customers who seek liquidity. The order flow is represented by doubly stochastic Poisson processes, with intensities that depend on the market maker's pricing policy. A trader is hired by the market maker to manage the residual risk in the *dealer book*[2]. The profit and loss on holding risk is determined by the trader's unobservable effort giving rise to a principal-agent problem. The agency model matches a standard setting in which the trader affects the mean rate of return on the risk in the dealer book and builds on the continuous time techniques introduced in Sannikov (2008) and Demarzo and Sannikov (2006). The profits and losses per unit of risk follows a Brownian motion with a drift that depends positively on the hidden effort.

A market making policy consists of a compensation policy, a pricing policy and a stochastic termination time. The optimal market making policy maximizes the expected value that the market maker derives from making markets with an incentive compatible effort process. It is characterized using two state variables: the risk in the dealer book and the trader's continuation utility. These two state variables summarize the relevant history of the market maker. Under the optimal contract the trader finds it optimal to exert full hedging effort at all times. When the trader's continuation utility hits zero for the first time, the contract is terminated and the market maker experiences a loss due to costly liquidation.

In order to motivate the trader, the optimal contract relies on the promise of payments for hedging effort. As is the case in the standard setting introduced by Sannikov (2008), compensation needs to be deferred and exposed to profits and losses in order to provide the trader with incentives for exerting hedging effort. Specifically, the sensitivity of continuation utility to profits and losses needs to be greater than the private benefits that the trader gains from shirking. In addition to the standard sensitivity to the output driven by the trader's action, the martingale representation theorem produces a sensitivity of continuation utility to customer flow. In the optimal contract the market maker will adjust the flow sensitivity to optimize market making efficiency. The crux of the analysis is that the optimal spreads quoted by the market maker are linked to the way in which

[2]The "dealer book" or simply "the books" is the term used by institutional market makers to describe the system that keeps track of open trades and risk from customer transactions.

the trader's continuation utility is adjusted for order flow. This is due to the fact that compensation and pricing are two complementary tools for managing the risk of costly termination of employment. Without proper incentive compensation, the market maker must manage the risk of costly termination by keeping risk in check through pricing primarily. However, when compensation is properly deferred and adjusted optimally according to order flow, the risk of termination is minimized and the market maker's pricing strategy becomes less conservative. Hence, letting trader compensation be optimally sensitive to fluctuations in risk will benefit customers by minimizing spreads.

The model also shows that the market maker adjusts prices in response to changes in the trader's continuation utility. Since continuation utility mirrors past profitability this essentially means that the market maker will change prices in response to profits and losses. In a survey of the market microstructure literature, Biais, Glosten, and Spatt (2005) emphasize[3] that the first generation inventory models do not fully explain why financial institutions employing dealers would be averse to diversifiable risk:

> *....While individuals are indeed likely to exhibit risk aversion, it is less obvious why the banks, securities houses and other financial institutions employing dealers would be averse to diversifiable risk.*

In the present paper institutional risk aversion is driven by the threat of costly termination of employment. When the market maker experiences a loss on residual risk from market making, the trader's continuation utility must be adjusted accordingly in order to induce full hedging effort. This increases the likelihood of terminating employment with unhedged risk in the dealer book and it is therefore optimal to adjust customer arrival intensities in order to rebalance risk. Losses on holding inventory can therefore lead to a decrease in the ask-price as well as a decrease in the bid-price. Since the market maker trades off the risk of costly liquidation against the expected upside of retaining risk, prices and order flow intensities will depend on the level of risk in the dealer book. How much prices and order flow intensities will change after a profit or loss or a transaction generally depends on the risk of termination and thus the level of the trader's continuation utility.

[3]See Biais et al. (2005), page 222.

Biais et al. (2005) then go on to suggest an analysis of an institutional market maker in spirit of the model in this paper:

> *.... To speak to this issue it could be fruitful to analyze theoretically the internal organization of these financial institutions. For example, suppose the dealers need to exert costly but unobservable effort to be efficient and take profitable inventory positions. To incentivize them to exert effort, it is necessary to compensate them based on the profits they make. In this context, even if diversifiable risk does not enter the objective function of the financial institution, it plays a role in the objective function of an individual dealer quoting bid and ask prices.*

Something similar happens in this model: even though both players are risk neutral, the residual risk from market making enters the market maker's value function and therefore plays a role in the optimal contract. The main difference between the analysis suggested above and the approach taken in this paper concerns the way prices are quoted. In practice the trader is responsible for quoting prices. However, in the analysis that follows, the market maker dictates the pricing policy rather than letting the trader quote the prices. This assumption rests on the fact that in most financial institutions, the quoted prices will be publicly known within the firm which makes it possible for senior management to monitor and control the firm's pricing policy.[4]

The market microstructure theory can be roughly divided into three generations. The first generation market microstructure models that followed the work of Garman (1976) dealt with inventories and order handling costs. Liquidity provision with inventory considerations has been studied by Ho and Stoll (1981), Ho and Stoll (1983) and Amihud and Mendelson (1980), Grossman and Miller (1988) and Mildenstein and Schleef (1983). Roll (1984) studies a model where market makers incur a fixed cost trading shares. The second generation market microstructure models is concerned with asymmetric information. Specifically, Glosten and Milgrom (1985) and Kyle (1985) study the implications of adverse selection on the bid and ask prices quoted by the market maker. The recent third

[4]If the trader holds private and useful information about market conditions it might be suboptimal for the market maker to dictate a pricing policy. In this case the market maker is facing a delegation problem. An analysis of this situation would take the analysis in an entirely different direction and is beyond the scope of this paper.

generation models pioneered by Duffie et al. (2005) studies the implications of search frictions for market liquidity. Building on this search-theoretic framework Lagos and Rocheteau (2009) and Weill (2007) study the distribution of asset holdings and market making under selling pressure respectively. The present paper contributes to the market microstructure literature by considering dynamic agency as a source of market making frictions. To my knowledge, this work is the first to integrate financial intermediation with dynamic incentive contracting.[5]

The paper is also related to the growing literature on dynamic moral hazard that uses recursive techniques to characterize the optimal contract. This literature began with Green (1987), Abreu, Pearce, and Stacchetti (1990) and Spear and Srivastava (1987). By characterizing incentive compatibility using martingale techniques Sannikov (2008) and Demarzo and Sannikov (2006) were the first to provide a tractable dynamic principal-agent model, in which it is possible to explicitly characterize the optimal contract using a differential equation. Using the same martingale technique Biais, Mariotti, Rochet, and Villeneuve (2010) study a continuous-time agency model in which an agent with limited liability must exert unobservable effort to reduce the likelihood of losses. Biais, Mariotti, and Rochet (2007) study agency problems and the optimal design of securities. In a distantly related paper Demarzo, Fishman, He, and Wang (2012) study the implications of agency problems for a firm's investment decision. The main contribution of the present paper relative to the dynamic agency literature is to analyze the interplay between incentive compensation and liquidity provision.

The remainder of this paper is organized as follows. Section 2 specifies the model and characterizes incentive compatibility. I derive the market maker's value function and solve for the optimal market making policy in section 3. Section 4 studies the implications of the optimal contract for prices quoted in the optimal policy. Section 5 analyzes an example of inventory liquidation which is the simplest special case of the general model. Section 6 studies dealing with inventory as a special case of the general model. The empirical implications of the model are discussed in section 7 and section 8 concludes. Proofs are included in the Appendix.

[5]This paper is a revised version of a chapter in Falkeborg (2013). Another chapter studies an agency model of an over-the-counter market where the agent is generating order flow.

2 The Model

This section presents a model of financial intermediation where a market maker faces an agency problem. There are two players: a market maker and a trader. The market maker dictates the firms pricing policy and the trader is responsible for managing risk from order flow by exerting unobservable effort.

2.1 Environment

Fix a probability space $(\Omega, \mathcal{F}, P)$. Time is continuous and the market maker determines the price for immediacy by continuously proposing prices for taking on risk from liquidity seekers. Transactions occur according to a family of point processes $\boldsymbol{N}_t = (N_{it})_{i\in\mathcal{I}}$ where $\mathcal{I}$ is the set of possible transactions. The market maker determines the price for immediacy at any given time. Specifically, the market maker proposes spreads $\boldsymbol{S}_t = (S_{it})_{i\in\mathcal{I}}$ where S_{it} is the spread charged for transaction $i \in \mathcal{I}$.[6] The intensities of the customer arrival times depend on the prices set by the market maker. Formally, assume that for a function $\boldsymbol{\Lambda} = (\Lambda_{it})_{i\in\mathcal{I}} : [0,\infty)^{\mathcal{I}} \to [0,\infty)^{\mathcal{I}}$ a martingale M_i is defined by

$$M_{it} = N_{it} - \int_0^t \Lambda_i(S_{iu})du \tag{1}$$

for each $i \in \mathcal{I}$. Customer arrival intensities are decreasing in the spread proposed by the market maker. Moreover, I will assume throughout the paper that the following condition is satisfied.

Condition 1. *The function $\boldsymbol{\Lambda} = (\Lambda_i)_{i\in\mathcal{I}}$ is bounded and differentiable with $\Lambda_i(S_i) \equiv 0$ for S_i greater than some value S_{max}. Furthermore $\Lambda_i'(S_i) < 0$ and $\Lambda_i''(S_i) \leq 0$ for $S_i < S_{max}$.*

Assuming that $\Lambda_i(S_i) \equiv 0$ for large enough S_i captures the idea that if the spread proposed by the market maker is large enough, no customer will arrive to trade. In a market with some degree of competition this can be expected. In general, a

[6] For instance, if the market maker is dealing a single asset by proposing bid and ask prices we may take $\mathcal{I} = \{a, b\}$. In this case N_a represents the number of transactions on the ask side and N_b represents the number on transactions on the bid side. In this case, we can let P denote the fundamental exogenous value of the asset as in Ho and Stoll (1981). The market maker will then propose bid and ask prices $P + S_{at}$ and $P - S_{bt}$. Assuming assets enter the dealer book with value P the change in value for the market maker becomes $S_{at}dN_{at} + S_{bt}dN_{bt}$.

higher degree of competition, means that the customer arrival intensities will be more sensitive to price changes.

2.2 The Dealer Book

Whenever a customer arrives to trade and $dN_{it} > 0$, the trade is recorded and enters the *dealer book*. The dealer book reflects the inevitable residual risk from market making and must be actively managed by the trader.[7] The profits and losses from holding risk and therefore the value of the dealer book B_t is driven by the trader's action as well as a standard Brownian motion Z on $(\Omega, \mathcal{F}, P)$. Specifically, letting $A \in \{0, 1\}$ denote the trader's effort level, the profits and losses in the dealer book evolves according to

$$dB_t(\boldsymbol{N}_t, A_t) = K(\boldsymbol{N}_t)\Big(\mu(\boldsymbol{N}_t, A_t) + \sigma dZ_t\Big) \tag{2}$$

where $K : \mathbb{R}^{\mathcal{I}} \to (0, \infty)$ with $K(0) = 0$ is a function that reflects the size of the book and $\mu : \mathbb{R}^{\mathcal{I}} \to \mathbb{R}$ is a bounded function that measures the difficulty of managing the risk in the book. Thus, the value of the dealer book at any given time is a function of all previous trading activity as well as the trader's hedging effort up until that time. For a given pricing strategy $\mathbf{S_t} = (\mathbf{S_{it}})_{\mathbf{i}\in\mathcal{I}}$ the total profit and loss X_t from market making evolves as

$$dX_t = \sum_{i\in\mathcal{I}} S_{it} dN_{it} + dB_t(\boldsymbol{N}_t, A_t). \tag{3}$$

The trader possesses the unique and necessary skills for hedging and managing risk in the dealer book. Managing risk is costly for the trader which is modeled as a private benefit from shirking: if the trader shirks, i.e. $A_t = 0$, in the time interval $(t, t+dt)$ he obtains a private benefit of φdt where $\varphi > 0$; on the other hand, if the trader exerts a hedging effort $A_t = 1$, he obtains no private benefit. The trader's hedging effort has a positive impact on the drift $\Delta\mu_t \equiv \mu(\boldsymbol{N}_t, 1) - \mu(\boldsymbol{N}_t, 0) > 0$. Since $K(\boldsymbol{N}_t) \in [0, \infty)$, hedging effort therefore increases the expected return to holding securities in the dealer book.

[7] The probability of finding two opposing trades simultaneously is zero. For instance, if $\mathcal{I} = \{a, b\}$ the event that $dN_{as} > 0$ and $dN_{bs} > 0$ for some $s > 0$ is a null event.

2.3 The Market Maker's Problem

Hedging activities undertaken by the trader are unobservable by the market maker. The trader is protected by limited liability, and hence can only receive positive transfers from the market maker. By contrast, the market maker has unlimited liability and must therefore cover all potential losses from market making. Both players are risk neutral. The market maker discounts the future at rate $r > 0$ and the trader discounts at rate $\gamma > r$. This introduces a wedge between the valuation of future transfers by the market maker and the trader, and rules out indefinitely postponement of payments in the optimal contract.

I assume that customer transactions $\boldsymbol{N}$ as well as changes in the value of the dealer book B are observable and contractible. The market maker will offer the trader a contract that specifies a compensation policy and a termination time written on the observable market making history. A contract $\Gamma = (L, \tau)$, specifies a compensation policy L_t and an endogenously determined random time τ when the contract is terminated and employment ends. The process L_t describes cumulative transfers from the market maker to the trader and is non-negative and increasing due to the limited liability constraint. The timing of events in the time interval $(t, t + dt)$ is as follows:

1. The market maker sets the spreads S_{it} for all $i \in \mathcal{I}$.
2. The trader decides on effort $A_t \in \{0, 1\}$.
3. With probability $\Lambda_i(S_{it})dt$ transaction $i \in \mathcal{I}$ occurs.
4. The trader receives a non-negative transfer dL_t from the market maker.
5. Employment ends or continues.

With this timing of events the information setup can be described as[8]

$$\begin{aligned} \mathcal{G}_t &= \boldsymbol{\sigma}\Big\{\boldsymbol{N}_s; 0 \leq s \leq t\Big\}, \\ \mathcal{H}_t &= \boldsymbol{\sigma}\Big\{Z_s; 0 \leq s \leq t\Big\}, \\ \mathcal{F}_t &= \mathcal{G}_t \vee \mathcal{H}_t. \end{aligned}$$

[8]Here I use the boldface symbol $\boldsymbol{\sigma}$ to denote a sigma-algebra generated by a given stochastic process and $\vee$ to denote the join of to sigma-algebras.

The pricing decision of the market maker is taken before knowing the realization of customer arrival processes while payments are made after knowing whether there has been a transaction. Hence, the pricing process $\boldsymbol{S}_t$ is $\mathcal{F}_t$-predictable and the compensation process L_t is $\mathcal{F}_t$-adapted. Given a contract $\Gamma = (L, \tau)$ and an effort process A_t the trader's expected discounted payoff is

$$W_0 = \mathbb{E}^A\Big[\int_0^\tau e^{-\gamma t}\Big(dL_t + \varphi(1 - A_t)dt\Big)\Big] \tag{4}$$

and the expected discounted profits for the market maker is

$$\mathbb{E}^A\Big[\int_0^\tau e^{-rt}\Big(dX_t - dL_t\Big)\Big] \tag{5}$$

where the expectation $\mathbb{E}^A$ is taken with respect to the measure P^A generated by the trader's action A. In the following $\mathbb{E}$ is the expectation operator under the measure induced by the trader's action, unless otherwise stated.

2.4 Termination

Assume that $K(\boldsymbol{N}_0) > 0$ so that the trader's action is needed from the outset. Employment ends when the trader is fired at time τ_f, or when the dealer book reaches the empty state $K(\boldsymbol{N}) = 0$ at time τ_e when the trader's hedging effort is no longer needed. Both stopping times are determined endogenously in the model. Hence, the contract is terminated at time τ given by

$$\tau = \tau_f \wedge \tau_e.$$

Terminating the trader's employment whenever $K(\boldsymbol{N}) \neq 0$ is costly for the market maker: without someone running the dealer book the market maker is sitting on unhedged risk. If employment ends before the risk is liquidated, i.e. if $K(\boldsymbol{N}_{\tau_f}) \neq 0$, the market maker realizes a liquidation loss $\ell(\boldsymbol{N}) < 0$. I will let $\ell(\boldsymbol{N})$ be exogenous and unspecified but one could assume that in case of termination, the market maker must liquidate the position without the traders hedging effort.[9]

[9] Thus, after employment ends the market maker experiences a period of costly liquidation with a payoff of

$$\ell = \mathbb{E}\int_0^T e^{-rt}(\mu(\boldsymbol{N}_t, 0)dt + \sigma dZ_t)$$

2.5 Incentive Compatibility

An effort process A is incentive compatible with respect to a contract Γ if it maximizes the trader's expected payoff given Γ. In order to solve for the optimal contract and pricing policy we need to characterize incentive compatibility first. The optimal contract can be written recursively with the trader's continuation utility as a state variable and will be derived using dynamic programming. For a contract Γ and a effort process A_t, denote by $W_t(\Gamma, A)$ the trader's continuation utility defined as

$$W_t(\Gamma, A) \equiv \mathbb{E}\Big[\int_t^\tau e^{-\gamma(u-t)}\Big(dL_u + \varphi(1-A_u)du\Big)|\mathcal{F}_t\Big]. \tag{6}$$

The continuation utility is the trader's expected discounted payoff from time t and onwards, given that he will follow the strategy A. In other words, the continuation utility is the level of *deferred* compensation. Note that the continuation utility $W_t(\Gamma, A)$ is an $\mathcal{F}_t$-adapted stochastic process by construction. The martingale representation theorem therefore yields that the trader's continuation utility evolves in response to the stochastic changes in the dealer book.

Proposition 1. *There exists predictable processes Y_t and $\boldsymbol{H}_t = (H_{it})_{i\in\mathcal{I}}$ such that the trader's continuation utility $W_t = W_t(\Gamma, A)$ evolves according to*

$$dW_t = \Big(\gamma W_t - \varphi(1-A_t)\Big)dt - dL_t + Y_t\sigma dZ_t. + \sum_{i\in\mathcal{I}} H_{it}dM_{it} \tag{7}$$

where M_i is the martingale

$$M_{it} = N_{it} - \int_0^t \Lambda_i(S_u)du.$$

Furthermore, effort $A_t = 1$ is incentive compatible if and only if

$$Y_t \geq \frac{\varphi}{\Delta\mu_t}. \tag{8}$$

The representation (7) implies that the continuation utility of the trader evolves in response to the jumps of the compensated transaction process $\boldsymbol{M}_t$. Furthermore, to ensure that the trader chooses $A_t = 1$ for all $t > 0$, the market

where $T = \inf\{t > \tau : K(\boldsymbol{N}) = 0\}$. The cost of liquidation will then depend on drift in the dealer book when there is no hedging being done.

maker must expose the trader to the risk dZ_t. Intuitively, incentive compatibility requires that the trader is sufficiently exposed to the realized return of holding risk in the dealer book. It will follow in the next section that the incentive compatibility constraint (8) binds. This due to the fact that the trader has limited liability and thus termination is required in the optimal contract whenever $W_t = 0$. The market maker will therefore set $Y_t = \frac{\varphi}{\Delta\mu_t}$ since unlucky realizations of dZ_t will increase the probability of termination.

One of the main insights of Sannikov (2008) is that there is a one-to-one correspondence between a strategy that is incentive compatible with respect to a contract Γ and controlled processes of the form (7). Hence, even though the martingale representation theorem is not constructive, this insight allows us to solve for the optimal contract by solving a stochastic control problem with controls L, Y and $\boldsymbol{H}$. While Y_t is used to control the trader's time-t-incentives for hedging, the sensitivities $\boldsymbol{H}_t = (H_{it})_{i\in\mathcal{I}}$ will be chosen to optimize market making efficiency. It will follow from the analysis in the next section, that the optimal choice of H_{it} is inherently linked to the optimal choice of S_{it}.

3 The Optimal Market Making Policy

The market maker's problem is to find the contract and price process that maximize expected discounted profits (5) subject to incentive compatibility and subject to delivering the trader a required utility level W_0. The solution to this problem is the optimal market making policy. In the following we shall focus on implementing hedging effort. That is, we will solve for the optimal contract that induces $A_t = 1$ at all times. The optimal contract that delivers maximal trading effort can be derived by dynamic programming. Define the left limit of the trader's continuation utility by $W_{t-} = \lim_{s\nearrow t} W_s$ and similarly for the history of customer transactions $\boldsymbol{N}_{t-} = \lim_{s\nearrow t} \boldsymbol{N}_s$. These two variables will serve as state variables. Define the optimal value function $F(W_{t-}, \boldsymbol{N}_{t-})$ for the market maker as the highest expected payoff the market maker can obtain from a contract that provides the trader with the payoff W_{t-} given that the current transaction history is $\boldsymbol{N}_{t-}$. Assume for now that the value function $W \to F(W, \boldsymbol{N})$ is globally concave. This will follow formally from Theorem 1.

The market maker can always compensate the trader by paying a transfer

$dL > 0$ at a marginal cost of -1 and moving to the optimal contract with payoff $W - dL$. However, delaying compensation might be valuable for the market maker and therefore it must be the case that

$$F(W, \boldsymbol{N}) \geq F(W - dL, \boldsymbol{N}) - dL.$$

Optimizing with respect to dL shows that $dL > 0$ if and only if $F_W(W, \boldsymbol{N}) < -1$. That is, it is optimal to delay payments as long as they are more costly than utility promises. For any $\boldsymbol{N}$ with $K(\boldsymbol{N}) \neq 0$ let

$$W_{\boldsymbol{N}} \equiv \inf\{W > 0 : F_W(W, \boldsymbol{N}) = -1\}.$$

Concavity of the market maker's value function implies the following standard property: payments to the trader are made only if his continuation utility is at least $W_{\boldsymbol{N}}$ and the payment must bring the continuation utility back to the reflection point

$$dL = \max\{W - W_{\boldsymbol{N}}, 0\}.$$

The payments to the trader must be delayed so as to provide incentives to hedge so the trader's continuation utility is kept in the range $[0, W_{\boldsymbol{N}}]$.

By Itô's lemma we get that the market maker's value function over the interval $[0, W_{\boldsymbol{N}}]$ is given by the following HJB-equation

$$\begin{aligned} rF(W, \boldsymbol{N}) &= K(\boldsymbol{N})\mu(\boldsymbol{N}, A) + \max_{\boldsymbol{S}, \boldsymbol{H}, Y} \Big\{ \sum_{i \in \mathcal{I}} \Lambda_i(S_i) S_i + \gamma W F_W(W, \boldsymbol{N}) \\ &+ \frac{1}{2} Y^2 \sigma^2 F_{WW}(W, \boldsymbol{N}) - \sum_{i \in \mathcal{I}} \Lambda_i(S_i) G^i(W, \boldsymbol{N}, H_i) \Big\} \end{aligned} \tag{9}$$

where

$$G^i(W, H_i, \boldsymbol{N}) \equiv F(W, \boldsymbol{N}) + H_i F_W(W, \boldsymbol{N}) - F(W + H_i, \boldsymbol{N} + e_i) \tag{10}$$

and e_i denotes the unit vector in $\mathbb{R}^{\mathcal{I}}$ with the i'th entry equal to 1. Given the HJB-equation, we can now derive the optimal pricing policy and the evolution of W in the optimal contract. Since the function $W \to F(W, \boldsymbol{N})$ is strictly concave, it is optimal for the market maker to set $Y_t = \frac{\varphi}{\Delta \mu_t}$. The optimal choice of sensitivity to changes in the dealer book H_i is determined by the first order

condition

$$F_W(W + H_i, \boldsymbol{N} + e_i) = F_W(W, \boldsymbol{N}). \tag{11}$$

Similarly, the optimal spread S_i is given by the first order condition

$$\Lambda_i'(S_i)\Big(S_i - G^i(W, H_i, \boldsymbol{N})\Big) + \Lambda_i(S_i) = 0. \tag{12}$$

Condition 1 now ensures that spreads that satisfy (12) are indeed optimal.

Lemma 1. *The first order condition (12) is also a sufficient condition for optimality.*

Proof. Differentiating the market maker's value function twice with respect to S_i yields

$$\Lambda_i''(S_i)\Big(S_i - G^i(W, H_i, \boldsymbol{N})\Big) + 2\Lambda_i'(S_i). \tag{13}$$

From condition 1 we know that $\Lambda_i'(S_i) < 0$ for $S_i < S_{max}$ and hence that if S_i satisfy (12) we must have

$$S_i - G^i(W, H_i, \boldsymbol{N}) > 0.$$

Since $\Lambda_i''(S_i) \leq 0$ the expression in (13) is therefore negative. QED

To pin down a solution to the HJB-system we need the smooth pasting and the super contact conditions.[10] These conditions take the form

$$F_W(W_{\boldsymbol{N}}, \boldsymbol{N}) = -1 \qquad F_{WW}(W_{\boldsymbol{N}}, \boldsymbol{N}) = 0 \tag{14}$$

and the system of conditions at the boundary $W = 0$ which is determined by the cost of liquidation $\ell(\boldsymbol{N})$ is given by

$$F(0, \boldsymbol{N}) = \ell(\boldsymbol{N}). \tag{15}$$

The following result formalizes our findings. A formal proof of the result is given in the appendix.

Theorem 1. *The optimal market making policy that induces hedging effort for all $0 < t < \tau$ involves two state variables: the beginning-of-period history of customer*

[10] See, for example, Dixit (1993).

transactions $\boldsymbol{N}_{t-}$ *and the beginning-of-period level of the trader's continuation utility* W_{t-}*. The value of the market maker's policy is* $F(W_{t-}, \boldsymbol{N}_{t-})$ *where the value function* $F(W, \boldsymbol{N})$ *solves the system of ODE's*

$$\begin{aligned} rF(W,\boldsymbol{N}) &= K(\boldsymbol{N})\mu(\boldsymbol{N},A) + \sum_{i\in\mathcal{I}} \Lambda_i(S_i)\Big(S_i - G^i(W,\boldsymbol{N},H_i)\Big) \\ &+ \gamma W F_W(W,\boldsymbol{N}) + \frac{1}{2}(\frac{\varphi}{\Delta\mu})^2\sigma^2 F_{WW}(W,\boldsymbol{N}) \end{aligned} \tag{16}$$

for $W \in [0, W_{\boldsymbol{N}}]$ *with boundary conditions* (14) *and* (15)*, where* H_i *and* S_i *solve* (11) *and* (12) *respectively. For* $W > W_{\boldsymbol{N}}$ *the value function is given by* $F(W,\boldsymbol{N}) = F(W_{\boldsymbol{N}},\boldsymbol{N}) - (W - W_{\boldsymbol{N}})$*. The function* $W \to F(W,\boldsymbol{N})$ *is globally concave and strictly so in the region* $[0, W_{\boldsymbol{N}}]$*. The trader's continuation utility evolves according to*

$$dW_t = \gamma W_t dt - dL_t + \sum_{i\in\mathcal{I}} H_{it} dM_{it} + \frac{\varphi}{\Delta\mu_t}\sigma dZ_t. \tag{17}$$

Payments to the trader are made whenever the continuation utility W *exceeds the payment threshold* $W_{\boldsymbol{N}}$*, in which case the payment amounts to* $dL = \max\{W - W_{\boldsymbol{N}}, 0\}$*. Employment ends and the contract is terminated at time* τ_f *when* $W_t = 0$ *or at time* τ_e *when* $K(\boldsymbol{N}) = 0$*.*

Let us briefly discuss the optimal before we move on to study the implications. The evolution of the trader's continuation utility consists of several terms. The first two term is due to promise keeping and shows that the trader's continuation utility has to grow at the rate γ but decrease with direct payments to the trader dL_t. The second term captures the effect of a customer transaction on the trader's continuation utility. The last term provides the trader with incentives to exert hedging effort at all times. Because of inefficiencies resulting from terminating employment when the dealer book is non-empty, minimizing the risk in the trader's continuation utility while still maintaining incentive compatibility is optimal. The concavity of the market maker's value function reflects the risk aversion that the agency problem induces. The market maker is averse to fluctuations in the trader's continuation utility because of the risk of costly termination. While a higher W decreases the risk of termination, the benefit declines as deferring compensation is costly since $\gamma > r$. Overall these two effects yield a concave value function.

4 Implications for Prices

Having established the main properties of the value function we can analyze the implications for the optimal pricing policy. It followed from Theorem 1 that optimal spreads and the sensitivities to flow are functions of transactions and the trader's continuation utility. Let in the following $S_i(W, \boldsymbol{N}, H_i)$ and $H_i(W, \boldsymbol{N})$ denote the optimal spread and sensitivity respectively for transaction $i \in \mathcal{I}$.

4.1 Risk Premium

In the optimal contract, prices are set at the level that maximizes expected profits and involves the term $G^i(W, \boldsymbol{N}, H_i)$. This term can be thought of as a risk premium which is a function of transactions as well as the trader's continuation utility and reflects the possible trading upside as well as the threat of termination. We can rewrite $G^i(W, \boldsymbol{N}, H_i)$ as

$$G^i(W, \boldsymbol{N}, H_i) = \underbrace{v_i(W, \boldsymbol{N}, H_i)}_{\text{Value adjustment}} + \underbrace{\kappa_i(W, \boldsymbol{N}, H_i)}_{\text{Convexity adjustment}}$$

where

$$v_i(W, \boldsymbol{N}, H_i) \equiv F(W + H_i, \boldsymbol{N}) - F(W + H_i, \boldsymbol{N} + e_i) \tag{18}$$

and

$$\begin{aligned} \kappa_i(W, \boldsymbol{N}, H_i) &\equiv F(W, \boldsymbol{N}) + H_i F_W(W, \boldsymbol{N}) - F(W + H_i, \boldsymbol{N}) \\ &= \int_W^{W+H_i} -F_{WW}(\tilde{W}, \boldsymbol{N})(W + H_i - \tilde{W}) d\tilde{W}. \end{aligned} \tag{19}$$

From this we see that the function risk premium function driving the changes in the prices and order flow can be decomposed into two terms. One is a value adjustment term v_i, representing the change in the market maker's value due to a customer transaction. The difference in the value functions are evaluated at $W + H_i$ since this is the where continuation utility will be after the transaction $i \in \mathcal{I}$. The other term κ_i, adjusts for the change in the market makers value due to changes in the trader's continuation utility. I refer to this term as the convexity adjustment since $-F(\cdot, \boldsymbol{N})$ is convex.[11] A more concave value function due to

[11] Note that the convexity adjustment corresponds to the remainder term of the first order Taylor approximation of the market maker's value function at the point W.

higher risk and higher payoff leads to a higher convexity adjustment.

Since the value function is concave, we see that the convexity adjustment will have the same sign as the sensitivity H_i. If a transaction is accompanied by a positive bump in continuation utility the convexity adjustment will be positive since deferring compensation is costly for the market maker. The sign of the value adjustments depends on the gain from moving the dealer book from $\boldsymbol{N}$ to $\boldsymbol{N} + e_i$ relative to the associated change in liquidation cost. If the potential downside is sufficiently high compared to the expected upside, the value adjustment will be positive. These two effects will add up to determine the overall risk premium embedded in the quoted prices.

4.2 Compensation and Liquidity

An important implication of the optimal contract, is that optimal spreads will be affected by the market maker's compensation policy. To see this, note that for a given sensitivity H_i, the optimal spread will be a function of W, $\boldsymbol{N}$ and H_i since the market maker will choose a pricing policy $S_i(W, \boldsymbol{N}, H_i)$ such that

$$S_i(W, \boldsymbol{N}, H_i) = \arg\max_{S_i} \ \Lambda_i(S_i)\Big(S_i - G^i(W, \boldsymbol{N}, H_i)\Big). \tag{20}$$

A concave value function and the fact that spreads are increasing in risk premium yields the following fundamental result.

Theorem 2. *Optimally adjusting the trader's continuation utility according to transactions leads to lower spreads. Specifically, for each $i \in \mathcal{I}$ the optimal sensitivity $H_i(W, \boldsymbol{N})$ is chosen such that*

$$H_i(W, \boldsymbol{N}) = \arg\min_{H_i} \ S_i(W, \boldsymbol{N}, H_i). \tag{21}$$

where $S_i(W, \boldsymbol{N}, H_i)$ is given by (20).

Theorem 2 shows that letting trader compensation be optimally sensitive to fluctuations in the dealer book will minimize the quoted spreads. The intuition for this result is that pricing and compensation are complimentary tools for reducing the risk of liquidation. Without proper incentive compensation the market maker must control the risk in the dealer book through the pricing primarily. However by optimally shifting the trader's compensation according to transactions, the

agency cost is mitigated. This allows the market maker to enjoy a stronger flow through sharper pricing. The efficiency gain from proper incentive compensation will therefore benefit the market maker's customers.

4.3 Profits and Losses

When the market maker experiences a profit or loss, the trader's continuation utility must be adjusted accordingly. This changes the likelihood of terminating employment and it is therefore optimal to adjust the arrival intensity according to the perceived risk. A consequence of (11) is that the market maker will adjust prices and therefore intensities in response to changes in continuation utility according to the flow adjustments H_i. We can directly derive the effect of changes in the trader's continuation utility on the risk premium term. Differentiating $G^i(W, \boldsymbol{N}, H_i(W, \boldsymbol{N}))$ shows how the risk premium term changes with continuation utility.

Theorem 3. *The optimized risk premium $G^i(W, \boldsymbol{N}, H_i(W, \boldsymbol{N}))$ changes in response to changes in continuation utility and*

$$\frac{\partial}{\partial W} G^i(W, \boldsymbol{N}, H_i(W, \boldsymbol{N})) = H_i(W, \boldsymbol{N}) F_{WW}(W, \boldsymbol{N}). \tag{22}$$

Consequently, when $H_i > 0$ ($H_i < 0$), the spread $S_i(W, \boldsymbol{N}, H_i(W, \boldsymbol{N}))$ is a decreasing (increasing) function of W.

Equation (22) shows that when a transaction is accompanied by an adjustment in continuation utility, the risk premium term will be either decreasing or increasing in continuation utility depending on the sign of H_i since the value function is concave. To see why price changes are linked to adjustments in continuation utility, note that the sensitivities H_i, if interior, are set such that a marginal change in the market maker's expected value with respect to the trader's continuation utility is equalized before and after a transaction. If the market maker bumps the trader's continuation utility in responds to transaction $i \in \mathcal{I}$, it must be because $F_W(W, \boldsymbol{N} + e_i) \neq F_W(W, \boldsymbol{N})$, which means that the risk profile in the dealer book changes with that transaction. If a transaction $i \in \mathcal{I}$ is accompanied by a positive bump in continuation utility, i.e. $H_i > 0$, it means that the the market maker finds it optimal to decrease the likelihood of costly termination

of employment after that transaction. In other words, it is more costly for the market maker to terminate employment holding $\boldsymbol{N} + e_i$ than simply holding $\boldsymbol{N}$. When this is the case, the market maker will optimally control the likelihood of the transaction through spreads in such a way that the probability increases as continuation utility increases.

The first generation market microstructure models generally assume that market makers are risk averse or face capital constraints. These models predict that bid-ask prices are set according to inventory levels in order to keep the inventory within a preferred range. In this model, a form of risk aversion arises endogenously. Although the market maker is risk neutral and maximizes expected lifetime wealth, risk from market making is kept in check through the adjustment of prices. Price changes are driven by the threat of ex-post inefficient termination, which is needed to maintain incentive compatibility, as the trader is protected by limited liability. From (2) and (17) we see that a profit or loss $dB(\boldsymbol{N}_t, A_t)$ from holding risk in the dealer book will induce a change in continuation utility Δ_W given by

$$\Delta_W \equiv \frac{\varphi}{\Delta\mu_t}\left(\frac{1}{K(\boldsymbol{N_t})}dB(\boldsymbol{N}_t, A_t) - \mu(\boldsymbol{N}_t, A_t)\right).$$

From (22) it then follows that for a given level of continuation utility W, a profit or loss of $dB(\boldsymbol{N}_t, A_t)$ will induce an adjustment of the risk premium Δ_{G_i} that amounts to

$$\Delta_{G_i} \approx H_i(W, \boldsymbol{N})F_{WW}(W, \boldsymbol{N})\Delta_W.$$

Thus, in contrast to the model analyzed in Ho and Stoll (1981), risk aversion arises endogenously in this setup and the risk aversion is reflected in the risk premium through the concavity of the market maker's value function.

The next two sections explores the implications of the model based on numerical examples.

5 Example: Liquidating an Asset Position

We first consider the simplest special case of the model. In this example, the market maker starts out with one unit of the asset, and must determine how and when to offload the asset. The market maker must continuously determine the ask price S_t, i.e. the price for offloading the asset and the intensity of the

customer arrival time depends on the prices set by the market maker. Specifically, a customer arrives to buy the asset when $dN_t > 0$ where N is a doubly stochastic Poisson process with intensity $\Lambda(S)$ given by

$$\Lambda(S_t) \;=\; \eta - \lambda S_t \tag{23}$$

where $\eta, \lambda > 0$. The market maker can choose any spread $S_t \in [0, \eta/\lambda]$.

Assume that the observable incremental profit and loss from holding the asset over time interval dt is given by $d\Pi_t$, where Π is the process that represents the value gained or lost from trading activities. For a given spread S_t provided by the market maker, the revenue evolves as

$$dX_t = S_t dN_t + d\Pi_t. \tag{24}$$

The profit and losses on from holding the asset is determined by the trader's unobservable hedging action $A_t \in \{0, 1\}$. The trader's action determines the expected change of inventory value, so that

$$d\Pi_t = -(1 - A_t)\mu_l dt + A_t \mu_g dt + \sigma dZ_t \tag{25}$$

where $\mu_l, \mu_g \geq 0$. This setting corresponds to letting $\mathcal{I}$ contain a single asset with $K(\boldsymbol{N}_t) = 1 - N_t \geq 0$ and $dB_t(\boldsymbol{N}_t, A_t) = (1 - N_t)d\Pi_t$ in the general model.

The market maker's problem leads to the following HJB equation:

$$\begin{aligned} rF(W) \;&=\; \mu_g + \max_{S,H,Y} \Big\{ S\Lambda(S) + \big[\gamma W - H\Lambda(S)\big] F_W(W) \\ &+\; \frac{1}{2} Y^2 \sigma^2 F_{WW}(W) \\ &+\; \Lambda(S)\big[-(W + H) - F(W)\big] \Big\} \end{aligned}$$

where H and Y are the sensitivities produced by the martingale representation theorem. Given the concavity of $F(W)$, setting $Y = \varphi/(\mu_g + \mu_l)$ is optimal for the market maker. From the HJB we see that the optimal choice of sensitivity to the transaction H maximizes

$$H\Lambda(S)(-F_W(W) - 1).$$

Since $F_W(W) \geq -1$ for all W, whenever $\Lambda(S) > 0$, it is optimal to set $H = 0$. Elementary calculations show that the optimal price as a function of the trader's continuation utility is

$$S(W) = \frac{\eta - \lambda(-W - F(W))}{2\lambda}$$

whenever

$$-\frac{\eta}{\lambda} \leq -W - F(W) \leq \frac{\eta}{\lambda}.$$

To ease notation it will useful to define the following function

$$\Psi(x) = \max(\min(x, \frac{\eta}{\lambda}), -\frac{\eta}{\lambda}) \tag{26}$$

and let

$$G(W) \equiv \Psi(W + F(W)). \tag{27}$$

The value function is then the solution to the ODE

$$\begin{aligned} rF(W) &= \mu_g + \frac{\lambda}{4}\Big(\frac{\eta}{\lambda} - G(W))\Big)^2 \\ &+ \gamma W F_W(W) + \frac{1}{2}\varphi^2/(\mu_g + \mu_l)^2 F_{WW}(W). \end{aligned}$$

The value function is illustrated in Figure 1. The dotted line represents the reflection point W_1 where $F_W(W_1) = -1$ and $F_{WW}(W_1) = 0$. The optimal spread and transaction intensity as a function of continuation utility are illustrated in Figure 2. When W is low, costly termination becomes more likely and it is optimal for the market maker to decrease the ask price and thus increase the transaction intensity in order to offload the risk before the trader's employment is terminated. Note that price and transaction intensity are less sensitive to changes in continuation utility near the threshold W_1, due to the super contact condition $F_{WW}(W_1) = 0$. In this example the sensitivity to customer flow H is set at zero, since employment ends when the asset position is liquidated. In the next example the sensitivity can be set to optimize liquidity.

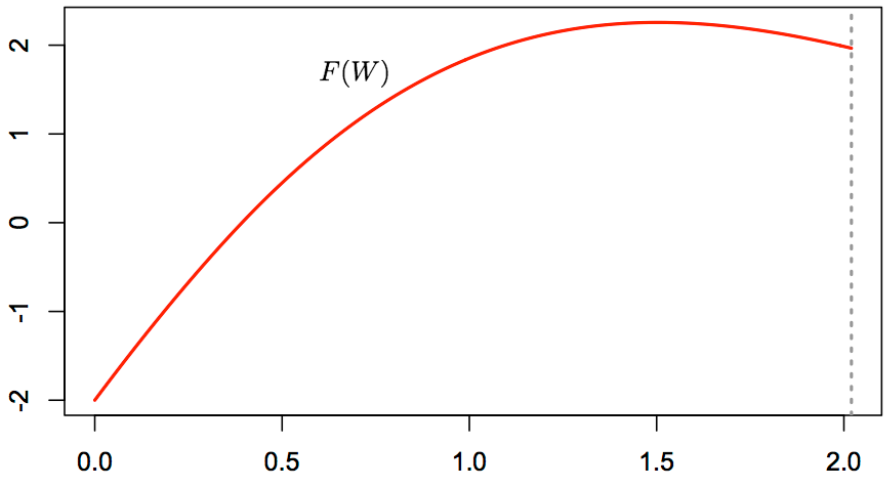

Figure 1: The market maker's value function. Parameters are $r = 0.1$, $\gamma = 0.3$, $\sigma = 0.8$, $\eta = 1$, $\lambda = 2$, $F(0) = -2$, $\mu_l = 0$ and $\mu_g = 0.8$.

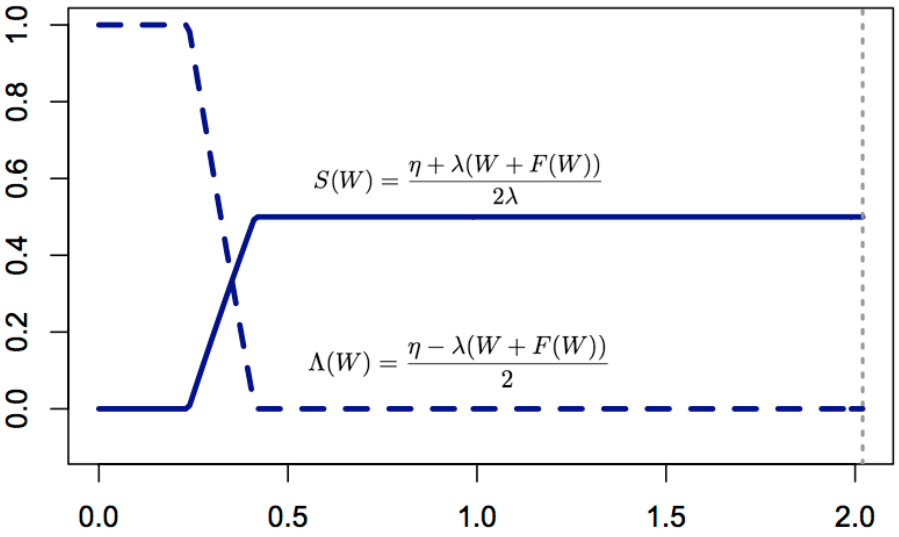

Figure 2: The optimal spread (solid line) and the optimal transaction intensity (dashed line) as a function of the trader's continuation utility.

6 Example: Dealing with Inventory

In order to illustrate the effect of properly adjusting deferred compensation according to flow, we will now turn to look at a problem where the market maker is restricted by a given level of inventory. To avoid computational difficulties, I assume that the market maker must keep inventory levels below 3 units at all times. This assumption ensures that we only have to solve a boundary value problem with 3 coupled ordinary differential equations in order to find the optimal market making policy. The arrival of clients is now modeled as a set of doubly stochastic Poisson processes $\boldsymbol{N}_t = (N_{it})_{i\in\{a,b\}}$. That is, $dN_a > 0$ when clients arrive on the ask-side and similarly $dN_b > 0$ represents the arrival of clients on the bid-side. The market maker starts out with inventory $I_0 > 0$ and inventory levels I_t evolves as

$$dI_t = dN_{bt} - dN_{at}.$$

When $I_t = 0$ the contract is terminated and market making ends. Since inventory is restricted to $I_t \leq 3$, the market maker is forced to lay off risk whenever $I_t = 3$. The market maker determines the price for immediacy by continuously proposing bid and ask prices. Let P denote the fundamental value of the asset. The fundamental value P is exogenous and represents the markets opinion of the true price.[12] The market maker will propose bid and ask spreads denoted by S_{at} and S_{bt} around P. For a given pair of spreads (S_{at}, S_{bt}) provided by the market maker the revenue evolves as

$$(S_{at} + P)dN_{at} - (P - S_{bt})dN_{bt}. \tag{28}$$

Assets are marked at mid price and enter holdings with a value of P. The market maker faces return uncertainty described by the stochastic process Π given by (25) which is driven by the trader's hedging action. The profits and losses from holding inventory therefore evolves as

$$PdN_{bt} - PdN_{at} + I_t d\Pi_t. \tag{29}$$

[12] We can think of P as being the mid price of the asset.

The evolution of the total profits X_t from market making is found by adding (28) and (29) and is given by

$$dX_t = S_{at}dN_{at} + S_{bt}dN_{bt} + I_t d\Pi_t. \tag{30}$$

Note that this corresponds to letting $\mathcal{I} = \{a, b\}$ with $K(\boldsymbol{N}_t) = N_b - N_a \geq 0$ and therefore $dB_t(\boldsymbol{N}_t, A_t) = I_t d\Pi_t$ in the general model. As in the previous section, the intensities of the doubly stochastic Poisson processes depend on the prices set by the market maker. For a given pair of prices S_{at} and S_{bt} the intensities are

$$\begin{aligned} \Lambda_a(S_a) &= \eta - \lambda S_{at} \\ \Lambda_b(S_b) &= \eta - \lambda S_{bt} \end{aligned} \tag{31}$$

where $\eta, \lambda > 0$. To keep things simple, the parameters in the intensity specification is the same on the bid and ask side, i.e. demand is symmetric.

The Value Function. In this example the relevant state variables are W and I. Since I can only take 3 different values let F_1, F_2 and F_3 denote the corresponding value functions. Setting up the HJB and optimizing over spreads and sensitivities we find that the optimal value function under effort $A_t = 1$ is given by the solution to the following system of coupled ODE's:

$$\begin{aligned} rF_1(W) &= \frac{\lambda}{4}\Big(\frac{\eta}{\lambda} + \Psi\Big(W + F_1(W)\Big)\Big)^2 \\ &+ \frac{\lambda}{4}\Big(\frac{\eta}{\lambda} + \Psi\Big(F_2(W + H_1) - F_1(W) + H_1 F_1'(W)\Big)\Big)^2 + \mu_g \\ &+ \gamma W F_1'(W) + \frac{1}{2}\varphi^2/(\mu_l + \mu_g)^2 F_1''(W), \end{aligned} \tag{32}$$

$$\begin{aligned} rF_2(W) &= \frac{\lambda}{4}\Big(\frac{\eta}{\lambda} + \Psi\Big(F_1(W + H_{2a}) - F_2(W) + H_{2a}F_2'(W)\Big)\Big)^2 \\ &+ \frac{\lambda}{4}\Big(\frac{\eta}{\lambda} + \Psi\Big(F_3(W + H_{2b}) - F_2(W) + H_{2b}F_2'(W)\Big)\Big)^2 + 2\mu_g \\ &+ \gamma W F_2'(W) + \frac{1}{2}\varphi^2/(\mu_l + \mu_g)^2 F_2''(W), \end{aligned} \tag{33}$$

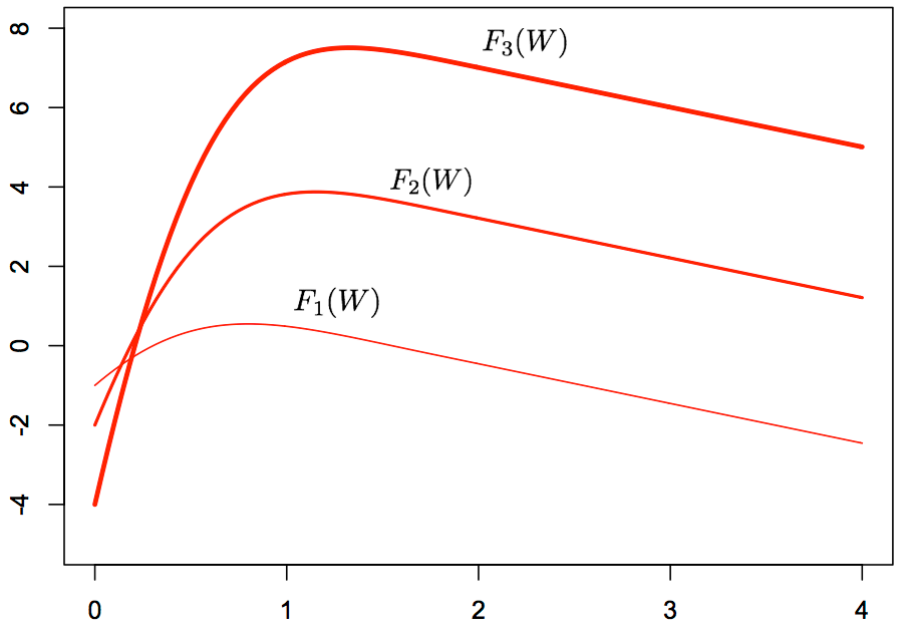

Figure 3: The market maker's value functions for $r = 0.1$, $\gamma = 0.25$, $\sigma = 0.2$, $\eta = 1$, $\lambda = 8$, $(F_1(0), F_2(0), F_3(0)) = (-1, -2, -4)$ and $\mu_l = 0$, $\mu_g = 0.4$.

$$\begin{aligned} rF_3(W) &= \frac{\lambda}{4}\Big(\frac{\eta}{\lambda} + \Psi\Big(F_2(W + H_3) - F_3(W) + H_3 F_3'(W)\Big)\Big)^2 + 3\mu_g \\ &+ \gamma W F_3'(W) + \frac{1}{2}\varphi^2/(\mu_l + \mu_g)^2 F_3''(W), \end{aligned} \tag{34}$$

where the function Ψ is defined by (26) and the sensitivities H_{1b}, H_{2a}, H_{2b} and H_{3a} are set to maximize the right-hand side of (32)-(34). Note that as in the previous example, $H_{a1} = 0$ as employment ends when inventory reaches zero. Since the market maker must lay off risk when $I = 3$, the drift on the bid side in this state must be set to zero. The boundary conditions are

$$F_\theta'(W_\theta) = -1 \qquad F_\theta''(W_\theta) = 0 \qquad F_\theta(0) = \ell_\theta, \qquad \theta \in \{1, 2, 3\}. \tag{35}$$

Figure 3 depicts the optimal value functions for the three levels of inventory. The value functions cross in this example reflecting the fact that a higher level of inventory leads to more downside risk as well as more upside potential. When continuation utility is low, the market maker prefers to keep inventory low to avoid a costly liquidation. As continuation utility increases the risk of liquidation

decreases and the market maker prefers to benefit from the positive drift. The optimal pricing and compensation policy reflects this trade-off. Since inventory is restricted to be less than 3 units, the interesting state is $I = 2$ when the market maker can freely choose between taking on more risk or instead offloading risk. The optimal ask spread in the state $I = 2$ is given by

$$S_a(W, 2, H_a) = \frac{\eta - \lambda(F_1(W + H_{2a}) - F_2(W) + H_{2a}F_2'(W))}{2\lambda} \tag{36}$$

and the optimal bid spread in the same state is given by

$$S_b(W, 2, H_b) = \frac{\eta - \lambda(F_3(W + H_{2b}) - F_2(W) + H_{2b}F_2'(W))}{2\lambda}. \tag{37}$$

Liquidity Gain from Flow-based Compensation. In this example, the implication of Theorem 2 follows directly from equation (36) and (37): when the sensitivities H_{a2} and H_{b2} are chosen to maximize the market maker's value function, the bid and ask spreads are both minimized. Figure 4 illustrates the optimal bid and ask spreads when compensation is not adjusted for order flow, i.e. when the spread is $S_\theta(W, 0)$. Figure 5 shows the compensation-optimized spreads $S_\theta(W, 2, H_\theta(W, 2)) = \arg\min_{H_\theta} S_\theta(W, \boldsymbol{N}, H_\theta)$. Comparing the two figures we clearly see that both the bid-spread and the ask-spread are decreased when the trader's continuation utility is optimally adjusted for order flow.

With the parameters chosen in this example, the market maker quotes conservatively on the bid-side for small W when compensation is not adjusted for flow. On the other hand, when compensation is optimized the market maker quotes aggressively on the bid-side for all levels of continuation utility. Comparing the ask-spread in Figure 4 with the ask-spread in Figure 5 we see, perhaps surprisingly, that the market maker is more aggressive on the ask-side in the compensation-optimized case. Hence, optimal compensation makes the market maker more prudent when it comes to offloading risk. Intuitively, when compensation is optimized the market maker finds it more attractive to reduce inventory due to the value gained from decreasing costly deferred compensation (since $H_{2a} < 0$).

In this example the bid-ask spread is positive for the chosen set of parameters and the non-optimized spread has a peak around 0.2. To understand why the non-optimized spread peaks in this example, note from Figure 4 that the market maker will increase the ask-spread before decreasing the bid-spread as continuation

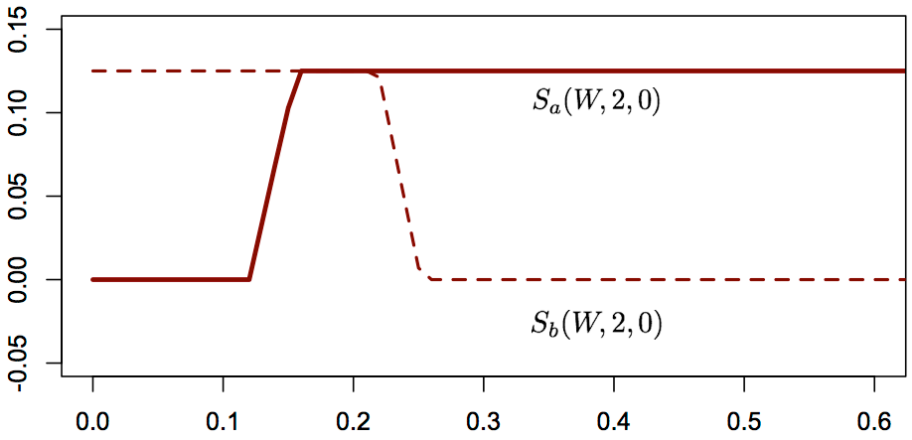

Figure 4: The market maker's bid and ask spread for $I = 2$ as a function of continuation utility when continuation utility is not adjusted for order flow. The dashed line shows the bid spread and the solid line shows the ask spread. Parameters are the same as in Figure 3.

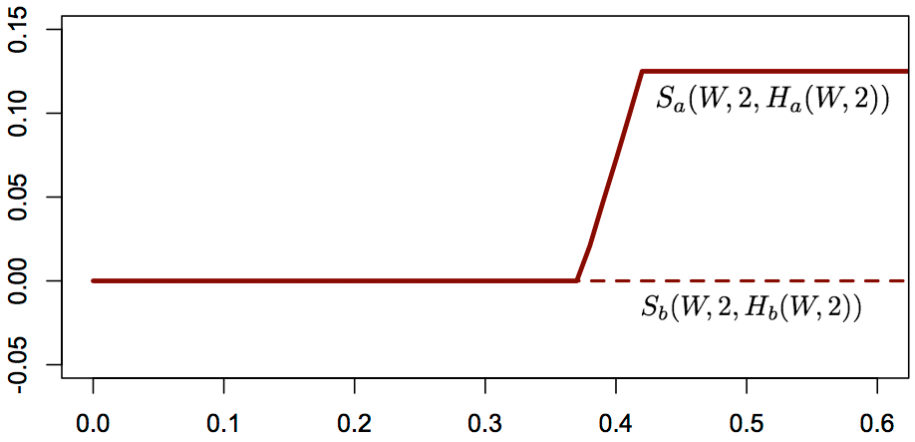

Figure 5: The market maker's compensation-optimized bid and ask spread as a function of continuation utility when $I = 2$. The dashed line shows the compensation-optimized bid spread and the solid line shows the compensation-optimized ask spread. Parameters are the same as in Figure 3.

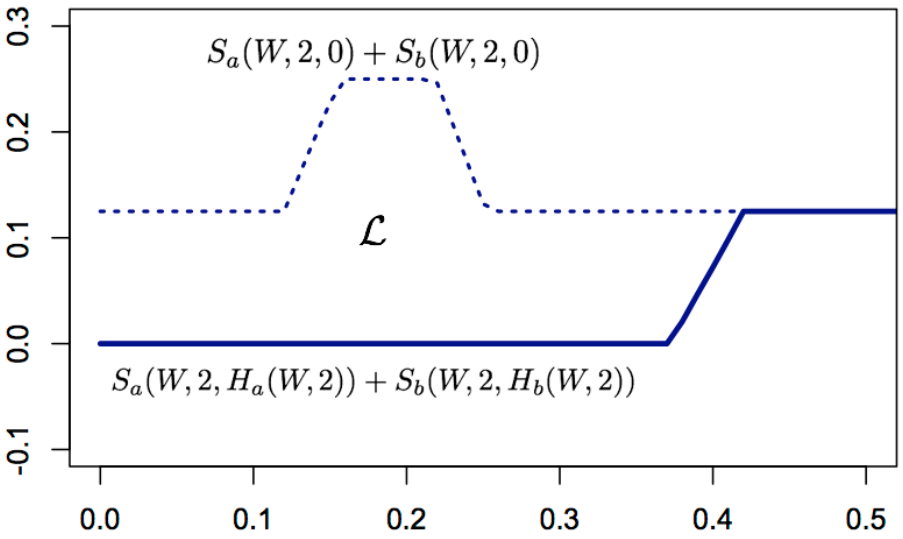

Figure 6: The market maker's total spread as a function of continuation utility when $I = 2$. The dotted line shows the case when continuation utility is not adjusted for order flow and the solid line shows the optimized spread. The area $\mathcal{L}$ between the two graphs represents the liquidity gain from optimizing trader compensation. Parameters are the same as in Figure 3.

utility increases from zero. When the sensitivities are optimized, we see from Figure 5 that both spreads are lowered. One way to measure the liquidity gained from adjusting the trader's continuation utility according to flow, is to calculate the area $\mathcal{L}$ between the two graphs in Figure 6:

$$\begin{aligned} \mathcal{L} &= \int_0^{W_2} S_a(W,2,0) + S_b(W,2,0) \\ &- S_a(W,2,H_a(W,2)) - S_b(W,2,H_b(W,2))dW. \end{aligned} \tag{38}$$

Strict concavity of the value function on $[0, W_2]$ implies that $\mathcal{L} > 0$. The size of the liquidity gain will generally depend on the parameters η and λ determining how much flow will change when prices are adjusted. A higher market depth measured by η/λ relative to the drift μ_g means that market making is more attractive than holding on to inventory. In this case the gain from properly adjusting continuation utility is high.

7 Discussion and Extensions

This section discusses the main predictions of the model as well as some possible extensions.

Speculating with Inventory. The Volcker Rule prohibits a financial institution from engaging in proprietary trading not related to the provision of immediacy.[13] While some forms of proprietary trading can be identified, it is generally more difficult to determine whether inventory risk has been retained in order to profit from price movements or because the retention is necessary to provide efficient intermediation. The model analyzed in this paper shows that efficient institutional market making involves "speculating with inventory" in the sense that the price for taking on a trade is set according to the profit potential from warehousing the risk from the trade. This is perhaps best illustrated in Figure 5 where it is evident that the market maker holds on to inventory by raising the ask-spread when continuation utility is high enough in order to benefit from the positive expected return to holding risk.

Flow-based Compensation. One of the key predictions of the model is that letting trader compensation be optimally sensitive to fluctuations in the dealer book will increase liquidity by minimizing the spread. This is due to the fact that the pricing and the compensation are complimentary tools for reducing the risk of liquidation. This result rests on the assumption the transaction intensities $\Lambda_i(S_i)$ are exogenously specified as a function of price and hence that sensitivities can be set freely set to optimize market making. A more realistic assumption would be that transactions have to be generated by effort. Specifically, instead of letting λ and η determine demand we could let

$$\Lambda_i = \Lambda_i(S_i, A_i) \tag{39}$$

where $A_i \in \mathcal{A}$ is an action taken by the agent that increases the probability that transaction $i \in \mathcal{I}$ occurs. We are then be faced with a multitask agency problem where $\{0,1\} \times \mathcal{A}^{|\mathcal{I}|}$ is the action space. In this case the sensitivity H_i controls the agent's incentives for increasing the probability of a transaction. A full analysis of

[13] For a discussion of the Volcker Rule and institutional market making see Duffie (2012).

endogenous flow is beyond the scope of this paper, but in order to induce effort, the sensitivity of continuation utility to the transaction N_i has to exceed the private benefits from shirking. Hence, the sensitivity H_i cannot be set freely to optimize liquidity in this case.[14]

Optimal Effort. So far, the analysis has focused on the optimal contract under maximal trading effort and also relies on the assumption that the agency problem is orthogonal to the market making problem in the sense that private benefits do not depend on the risk in the dealer book. Both these assumptions can be relaxed by analyzing the optimal contract under *optimal* effort when private benefits are risk dependent. Letting effort $A \in [0,1]$ the private benefits from shirking could be specified by function

$$\Phi : \mathbb{N}^{\mathcal{I}} \times [0,1] \to (0,\infty).$$

In this case we know from Sannikov (2008), that in order to induce a given level of effort A, the sensitivity to profit and losses is set at the minimum level that induces A, which is given by

$$H(\boldsymbol{N}, A) = \min\{h \in [0,\infty) : A \in \arg\max_{A' \in [0,1]} h\mu(\boldsymbol{N}_t, A'_t) + \Phi(\boldsymbol{N}, A')\}.$$

In this setting the HJB equation is given by

$$\begin{aligned} rF(W,\boldsymbol{N}) &= \max_{A\in[0,1]} \Big\{ K(\boldsymbol{N})\mu(\boldsymbol{N},A) + \sum_{i\in\mathcal{I}} \Lambda_i(S_i)\Big(S_i - G^i(W,\boldsymbol{N},H_i)\Big) \\ &+ (\gamma W - \Phi(\boldsymbol{N},A))F_W(W,\boldsymbol{N}) + \frac{1}{2}H(\boldsymbol{N},A)^2\sigma^2 F_{WW}(W,\boldsymbol{N})\Big\}. \end{aligned} \tag{40}$$

where S_i and H_i are determined by the same first order conditions.

8 Conclusion

This paper introduces dynamic agency into a continuous-time model of market making. Profits and losses on residual risk from customer flow is determined

[14] Falkeborg (2013) analyzes an agency model of market making where the agent induces order flow and contains some initial results in this direction.

by a trader's unobservable effort. Using a recursive contracting methodology, I characterize the impact of dynamic agency on quoted spreads.

The market maker's pricing policy is driven by the fear of inefficient termination of the trader's employment as well as the expected upside from retaining risk in the dealer book. Optimal contracting implies that agency costs are reflected in the pricing policy of the market maker which is a function of the trader's continuation utility. Since continuation utility mirrors past profitability, the market maker will keep risk in check by adjusting spreads in response to profits and losses. The analysis highlights the importance of deferring compensation according to transactions and shows that this will minimize the spreads set by the market maker.

A Proofs

A.1 Stochastic Environment

Let $(\Omega, \mathcal{F}, P)$ be a probability space completed with null sets and let $\mathcal{N}$ denote the collection of sets of $\mathcal{F}$ with P-measure zero (that is, the null sets of $\mathcal{F}$). In the sequel we work with the augmented natural-filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$, i.e. the underlying probability space $(\Omega, \mathcal{F}, P)$ equipped with the augmented natural filtration $(\mathcal{F}_t)_{t\geq 0}$ where $\mathcal{F}_t = \boldsymbol{\sigma}\Big((Z_s, \boldsymbol{N}_s), 0 \leq s \leq t\Big) \bigvee \mathcal{N}$. Note that since $\boldsymbol{N}$ is a doubly stochastic Poisson process, the process

$$X_t = (\boldsymbol{N}, Z) \tag{41}$$

is a process with *conditionally* independent increments, see Jacod and Shiryaev (2002) Definition II.6.2.

A.2 Proof of Proposition 1

Proof. The trader's lifetime expected payoff from a given contract Γ when he acts according to A can be expressed as

$$V_t(\Gamma, A) = \int_0^t e^{-\gamma u}\Big(dL_u + \varphi(1 - A_u)du\Big) + e^{-\gamma t}W_t(\Gamma, A). \tag{42}$$

By construction $V_t(\Gamma, A)$ is an $\mathcal{F}_t$-martingale under the probability measure P^A. The martingale representation theorem (see Jacod and Shiryaev (2002), Theorem III.4.34) implies that there exists $\mathcal{F}_t$-predictable processes Y and $\boldsymbol{H} = (H_i)_{i\in\mathcal{I}}$ such that

$$e^{\gamma t} dV_t(\Gamma, A) = Y_t \sigma dZ_t + \sum_{i\in\mathcal{I}} H_{it} dM_{it} \tag{43}$$

where M_{it}, $i \in \mathcal{I}$ are the compensated processes. From (42) and (43) we get the representation (7).

Let A be the maximal effort process, i.e. $A_t = 1$ for all t and let A'_t be an arbitrary alternative effort process. Let V'_t denote the trader's lifetime utility when he acts according to A'_t until time t and then reverts back to A_t. Then

$$V'_t = \int_0^t e^{-\gamma u}\Big(dL_u + \varphi(1 - A'_u)du\Big) + e^{-\gamma t} W_t(\Gamma, A). \tag{44}$$

From (42) we find that

$$V'_t = V_t(\Gamma, A) + \int_0^t e^{-\gamma u}\varphi(1 - A'_u)du. \tag{45}$$

Using the representation (43) and the fact that the Brownian Motions under the two probability measures P^A and $P^{A'}$ are related by

$$\sigma dZ^A_t = \sigma dZ^{A'}_t + \mu(\boldsymbol{N}_t, A'_t) - \mu(\boldsymbol{N}_t, A_t) \tag{46}$$

we can express the evolution of the trader's lifetime expected utility under the probability measure $P^{A'}$ as

$$\begin{aligned} e^{\gamma t} dV'_t &= \Big(\varphi(1 - A'_t) - Y_t\Big(\mu(\boldsymbol{N}_t, A_t) - \mu(\boldsymbol{N}_t, A'_t)\Big)\Big)dt & (47)\\ &+ Y_t \sigma dZ^{A'}_t . + \sum_{i\in\mathcal{I}} H_{it} dM_{it}. & (48) \end{aligned}$$

From this we see that if

$$Y_t \geq \frac{\varphi}{\mu(\boldsymbol{N}_t, 1) - \mu(\boldsymbol{N}_t, 0)} = \frac{\varphi}{\Delta\mu_t}$$

then V' is a supermartingale and a martingale only if $A'_t = 1$ for all t. If $Y_t < \frac{\varphi}{\Delta\mu_t}$ on a set of positive measure then the trader can gain by shirking on that set

and maximal effort is therefore suboptimal. Hence (8) is both a necessary and sufficient condition for maximal effort $A_t = 1$ to be incentive compatible. QED

A.3 Proof of Theorem 1

Let $H_i(W, \boldsymbol{N})$ denote the value H_i that solves (11), $S_i(W, \boldsymbol{N})$ denote the optimal spread defined by (12) and $\Lambda_i(W, \boldsymbol{N}) \equiv \Lambda_i(S_i(W, \boldsymbol{N}))$.

Lemma A.1. *For the function $W \to G^i(W, \boldsymbol{N}, H_i(W, \boldsymbol{N}))$ it holds that*

$$\frac{\partial}{\partial W} G^i(W, \boldsymbol{N}, H_i(W, \boldsymbol{N})) = H_i(W, \boldsymbol{N}) F_{WW}(W, \boldsymbol{N}). \tag{49}$$

Proof. Differentiation directly yields:

$$\begin{aligned} \frac{\partial}{\partial W} G^i(W, \boldsymbol{N}, H_i(W, \boldsymbol{N})) &= -(1 + \frac{\partial}{\partial W} H_i(W, \boldsymbol{N})) F_W(W + H_i(W, \boldsymbol{N}), \boldsymbol{N} + e_i) \\ &+ F_W(W, \boldsymbol{N}) \\ &+ \frac{\partial}{\partial W} H_i(W, \boldsymbol{N}) F_W(W, \boldsymbol{N}) + H_i(W, \boldsymbol{N}) F_{WW}(W, \boldsymbol{N}) \\ &= H_i(W, \boldsymbol{N}) F_{WW}(W, \boldsymbol{N}). \end{aligned}$$

QED

Lemma A.2. *For any $\boldsymbol{N}$ with $K(\boldsymbol{N}) > 0$ there exists an $\epsilon > 0$ such that the function $W \to F(W, \boldsymbol{N})$ is strictly concave on $[W_{\boldsymbol{N}} - \epsilon, W_{\boldsymbol{N}}]$.*

Proof. Note that $S_i(W, \boldsymbol{N})$ only depends on W through $G^i(W, \boldsymbol{N}, H_i(W, \boldsymbol{N}))$. Hence, differentiating F with respect to W and evaluating the expression at $W = W_{\boldsymbol{N}}$ using Lemma A.1 and the boundary conditions (14), we find that

$$F_{WWW}(W_{\boldsymbol{N}}, \boldsymbol{N}) = 2\frac{\Delta\mu_t^2}{\varphi^2}(\gamma - r) > 0$$

since $\gamma > r$. Because of the super contact condition $F_{WW}(W_{\boldsymbol{N}}, \boldsymbol{N}) = 0$ we must have $F_{WW}(W, \boldsymbol{N}) < 0$ when $W \in [W_{\boldsymbol{N}} - \epsilon, W_{\boldsymbol{N}}]$ for some $\epsilon > 0$. QED

Proof of Theorem 1. We first show that the value function is strictly concave over the entire interval $[0, W_{\boldsymbol{N}}]$. Set $\tilde{W}_{\boldsymbol{N}} = \sup\{W < W_{\boldsymbol{N}} : F_{WW}(W, \boldsymbol{N}) = 0\}$ and suppose $\tilde{W}_{\boldsymbol{N}} < W_{\boldsymbol{N}}$ so that $F_{WW}(\tilde{W}_{\boldsymbol{N}} + \epsilon, \boldsymbol{N}) < 0$ for $\epsilon > 0$. By continuity

$F_{WWW}(\tilde{W}_{\boldsymbol{N}}, \boldsymbol{N}) < 0$. By differentiating F and evaluating at $\tilde{W}_{\boldsymbol{N}}$ it follows that

$$F_W(\tilde{W}_{\boldsymbol{N}}, \boldsymbol{N}) = \frac{-1}{2(\gamma - r)} \frac{\varphi^2}{\Delta\mu_t^2} F_{WWW}(\tilde{W}_{\boldsymbol{N}}, \boldsymbol{N}) > 0.$$

Evaluating F at $\tilde{W}_{\boldsymbol{N}}$:

$$\begin{aligned} rF(\tilde{W}_{\boldsymbol{N}}, \boldsymbol{N}) - K(\boldsymbol{N})\mu(\boldsymbol{N}, A) &= \sum_{i \in \mathcal{I}} \Lambda_i(\tilde{W}_{\boldsymbol{N}}, \boldsymbol{N}) \Big(S_i(\tilde{W}_{\boldsymbol{N}}, \boldsymbol{N}) \\ &- G^i(\tilde{W}_{\boldsymbol{N}}, \boldsymbol{N}, H_i) \Big) \\ &+ \gamma \tilde{W}_{\boldsymbol{N}} F_W(\tilde{W}_{\boldsymbol{N}}, \boldsymbol{N}). \end{aligned} \tag{50}$$

Since

$$\Lambda_i(\tilde{W}_{\boldsymbol{N}}, \boldsymbol{N}) \Big(S_i(\tilde{W}_{\boldsymbol{N}}, \boldsymbol{N}) - G^i(\tilde{W}_{\boldsymbol{N}}, \boldsymbol{N}, H_i) \Big) \geq 0$$

this contradicts $F_W(\tilde{W}_{\boldsymbol{N}}, \boldsymbol{N}) > 0$. Hence $\tilde{W}_{\boldsymbol{N}} = W_{\boldsymbol{N}}$ and the value function $F(\cdot, \boldsymbol{N})$ is therefore strictly concave over the interval $[0, W_{\boldsymbol{N}}]$.

We now proceed to show that the contract given in Theorem 1 is indeed optimal. That is, for any policy that induces maximal hedging effort for all $t < \tau$, one has

$$F(W_{0-}, \boldsymbol{N}_{0-}) \geq \mathbb{E}\Big[\int_0^\tau e^{-rt} \Big(dX_t - dL_t \Big) \Big]. \tag{51}$$

Take any incentive compatible contract-effort pair $\{\Gamma, A\}$ and define the cumulative gains process $\mathcal{G}$ as

$$\mathcal{G}_t(\Gamma, A) = \int_0^t e^{-rs} \Big(dX_s - dL_s \Big) + e^{-rt} F(W_t, \boldsymbol{N}_t). \tag{52}$$

The continuation utility W of the trader evolves according to

$$dW_t = W_t dt - dL_t + Y_t \sigma dZ_t. + \sum_{i \in \mathcal{I}} H_{it} dM_{it}$$

where Y is such that $Y_t \geq \varphi/\Delta\mu_t$. By Itô's lemma we have that

$$\begin{aligned} dF(W_{t-}, \boldsymbol{N}_{t-}) &= \Big[\Big(\gamma W_t - \sum_{i\in\mathcal{I}} \Lambda_{it} H_{it}\Big) F_W(W_{t-}, \boldsymbol{N}_{t-}) \\ &+ \frac{1}{2} Y_t^2 \sigma^2 F_{WW}(W_{t-}, \boldsymbol{N}_{t-})\Big] dt \\ &- F_W(W_{t-}, \boldsymbol{N}_{t-}) dL_t + Y_t \sigma F_W(W_{t-}, \boldsymbol{N}_{t-}) dZ_t \\ &+ \sum_{i\in\mathcal{I}} \Big[F(W_{t-} + H_{it}, \boldsymbol{N}_{t-} + e_i) - F(W_{t-}, \boldsymbol{N}_{t-})\Big] dN_{it}. \end{aligned} \tag{53}$$

By equation (52) we find that

$$e^{rt} d\mathcal{G}_t(\Gamma, A) = dX_t - dL_t - rF(W_{t-}, \boldsymbol{N}_{t-}) dt + dF(W_{t-}, \boldsymbol{N}_{t-}) \tag{54}$$

and therefore

$$\begin{aligned} e^{rt} d\mathcal{G}_t(\Gamma, A) &= \Big[\Big(\gamma W_t - \sum_{i\in\mathcal{I}} \Lambda_{it} H_{it}\Big) F_W(W_{t-}, \boldsymbol{N}_{t-}) \\ &+ \frac{1}{2} Y_t^2 \sigma^2 F_{WW}(W_{t-}, \boldsymbol{N}_{t-}) - rF(W_{t-}, \boldsymbol{N}_{t-}) \\ &+ \sum_{i\in\mathcal{I}} \Lambda_{it}\Big(S_{it} + F(W_{t-} + H_{it}, \boldsymbol{N}_{t-} + e_i) - F(W_{t-}, \boldsymbol{N}_{t-})\Big) dt \\ &+ \Big[\sigma K(\boldsymbol{N}_{t-}) + Y_t \sigma F_W(W_{t-}, \boldsymbol{N}_{t-})\Big] dZ_t \\ &- \Big[1 + F_W(W_{t-}, \boldsymbol{N}_{t-}\Big] dL_t \\ &+ \sum_{i\in\mathcal{I}} \Big[S_{it} + F(W_{t-} + H_{it}, \boldsymbol{N}_{t-} + e_i) - F(W_{t-}, \boldsymbol{N}_{t-})\Big] dM_{it}. \end{aligned} \tag{55}$$

Recall that the value of the market maker under the optimal policy given in Theorem 1 solves system of ODE's (16) subject to the boundary conditions given in (14) and (15) for $W \in [0, W_{\boldsymbol{N}}]$. Plugging $rF(W_{t-}, \boldsymbol{N}_{t-})$ into (55) we see that for $t < \tau$ the drift of (55) is non-positive because $Y_t^2 \geq \varphi^2/\Delta\mu_t^2$, $F(\cdot, \boldsymbol{N})$ is concave with $F_W \geq -1$ and $dL_t \geq 0$. From this it follows that the cumulative gains process $\mathcal{G}(\Gamma, A)$ is an $\mathcal{F}_t$-supermartingale up to time t. $\mathcal{G}(\Gamma, A)$ is an $\mathcal{F}_t$-martingale if and only if $Y_t^2 = \varphi^2/\Delta\mu_t^2$, $H_{it} = H_i(W_{t-}, \boldsymbol{N}_{t-})$, $S_{it} = S_i(W_{t-}, \boldsymbol{N}_{t-})$ and $dL_t > 0$ only when $W_{t-} > W_{\boldsymbol{N}}$.

Let the total expected payoff for the market maker under an arbitrary incentive

compatible contract (Γ, A) be denoted

$$F(\Gamma, A) = \mathbb{E}\Big[\int_0^\tau e^{-rt}(dX_t - dL_t)\Big].$$

Then $F(\Gamma, A)$ can be written as

$$\begin{aligned} F(\Gamma, A) &= \mathbb{E}\Big[\mathcal{G}_\tau\Big] \\ &= \mathbb{E}\Big[\mathcal{G}_{t\wedge\tau}\Big] \\ &+ \mathbb{E}\Big[\mathbf{1}_{\{t\le\tau\}}\Big(\int_t^\tau e^{-rs}(dX_s - dL_s) - e^{-rt}F(W_{t-}, \boldsymbol{N}_{t-})\Big)\Big] \\ &\le F(W_{0-}, \boldsymbol{N}_{0-}) \\ &+ \mathbb{E}\Big[\mathbf{1}_{\{t\le\tau\}}\Big(\int_t^\tau e^{-rs}(dX_s - dL_s) - e^{-rt}F(W_{t-}, \boldsymbol{N}_{t-})\Big)\Big] \\ &= F(W_{0-}, \boldsymbol{N}_{0-})\Big] \\ &+ e^{-rt}\mathbb{E}\Big[\mathbf{1}_{\{t\le\tau\}}\Big(\mathbb{E}\Big[\int_t^\tau e^{-r(s-t)}(dX_s - dL_s)|\mathcal{F}_t\Big] - F(W_{t-}, \boldsymbol{N}_{t-})\Big) \end{aligned} \tag{56}$$

where the inequality follows from the fact that $\mathcal{G}_{t\wedge\tau}$ is a supermartingale with $\mathcal{G}_0 = F(W_{0-}, \boldsymbol{N}_{0-})$. Note that

$$\mathbb{E}\Big[\Big(\int_t^\tau e^{-r(s-t)}(dX_s - dL_s)\Big)|\mathcal{F}_t\Big] \le \sup_{\boldsymbol{N}} \mu(\boldsymbol{N}, A) + \sum_{i\in\mathcal{I}} \sup_{S_i} S_i\Lambda(S_i) - W_{t-}$$

and thus

$$\begin{aligned} F(\Gamma, A) &\le F(W_{0-}, \boldsymbol{N}_{0-}) + e^{-rt}\mathbb{E}\Big[\mathbf{1}_{\{t\le\tau\}}\Big(\sup_{\boldsymbol{N}} \mu(\boldsymbol{N}, A) \\ &+ \sum_{i\in\mathcal{I}} \sup_{S_i} S_i\Lambda(S_i) - W_{t-} - F(W_{t-}, \boldsymbol{N}_{t-})\Big)\Big] \\ &\le F(W_{0-}, \boldsymbol{N}_{0-}) \\ &+ e^{-rt}\mathbb{E}\Big[\mathbf{1}_{\{t\le\tau\}}(\sup_{\boldsymbol{N}} \mu(\boldsymbol{N}, A) + \sum_{i\in\mathcal{I}} \sup_{S_i} S_i\Lambda(S_i) - \ell(\boldsymbol{N}))\Big] \end{aligned} \tag{57}$$

since $F_W(W_{t-}, \boldsymbol{N}_{t-}) \ge -1$ and therefore

$$-W_{t-} - F(W_{t-}, \boldsymbol{N}_{t-}) \le -F(0, \boldsymbol{N}) = -\ell(\boldsymbol{N}). \tag{58}$$

Finally, letting $t \to \infty$ then yields

$$F(\Gamma, A) \le F(W_{0-}, \boldsymbol{N}_{0-}). \tag{59}$$

The result now follows by noticing that for the optimal policy the market maker achieves $F(W_{0-}, \boldsymbol{N}_{0-})$ and thus (59) holds with equality.

QED

A.4 Proof of Theorem 2

Proof. Let $S_i(G)$ be the value S_i that solves

$$\Lambda_i'(S_i)\Big(S_i - G\Big) + \Lambda_i(S_i) = 0. \tag{60}$$

Note that if S_i solves (60) then we must have $S_i - G \geq 0$ since $\Lambda_i'(S_i) < 0$. Implicit differentiation of $S_i(G)$ yields

$$S_i'(G) = \frac{\Lambda_i'(S_i)}{\Lambda_i''(S_i)\Big(S_i - G\Big) + 2\Lambda_i'(S_i)} > 0 \tag{61}$$

since $\Lambda_i''(S_i) \leq 0$. This show that optimal spreads are increasing in risk premium. The result now follows from noting from (11) that the optimal sensitivity $H_i(W, \boldsymbol{N})$ is chosen such that

$$H_i(W, \boldsymbol{N}) = \arg\min_{H_i} \; G^i(W, \boldsymbol{N}, H_i) \tag{62}$$

since $G^i(W, \boldsymbol{N}, H_i)$ is a convex function of H_i. QED

A.5 Proof of Theorem 3

Proof. The first part of the Theorem follows from Lemma A.1. The second claim follows from noting that $S_i(W, \boldsymbol{N})$ depends on W only through G^i. By Lemma A.1, it follows that

$$\frac{\partial S_i}{\partial W} = S_i'(G^i)\frac{\partial G^i(W, \boldsymbol{N}, H_i(W, \boldsymbol{N})))}{\partial W} = S_i'(G^i)H_i(W, \boldsymbol{N})F_{WW}(W, \boldsymbol{N}) \tag{63}$$

and therefore $\frac{\partial S_i}{\partial W} = -\operatorname{sgn}(H_i)$ by (61). QED

References

Abreu, D., D. Pearce, and E. Stacchetti (1990). Toward a Theory of Discounted Repeated Games with Imperfect Monitoring. *Econometrica 58*(5), 1041–1063.

Amihud, Y. and H. Mendelson (1980, March). Dealership markets: Market-making with inventory. *Journal of Financial Economics 8*(1), 31–53.

Biais, B., L. Glosten, and C. Spatt (2005, May). Market microstructure: A survey of microfoundations, empirical results, and policy implications. *Journal of Financial Markets 8*(2), 217–264.

Biais, B., T. Mariotti, and J.-C. Rochet (2007). Dynamic Security Design: Convergence to Continuous Time and Asset Pricing Implications. *Review of Economic Studies 74 (2)*, 345–390.

Biais, B., T. Mariotti, J.-C. Rochet, and S. Villeneuve (2010). Large risks, limited liability, and dynamic moral hazard. *Econometrica 78*(1), 73–118.

Demarzo, P. M., M. J. Fishman, Z. He, and N. Wang (2012). Dynamic agency and the q theory of investment. *The Journal of Finance 67*(6), 2295–2340.

Demarzo, P. M. and Y. Sannikov (2006). Optimal security design and dynamic capital structure in a continuous-time agency model. *The Journal of Finance 61*(6), 2681–2724.

Dixit, A. (1993). *The Art of Smooth Pasting*, Volume 55. Fundamentals in Pure and Applied Economics.

Duffie, D. (2012). Market Making under the Proposed Volcker Rule. In *Rock Center for Corporate Governance at Stanford University Working Paper No. 106.*

Duffie, D., N. Gârleanu, and L. H. Pedersen (2005). Over-the-Counter Markets. *Econometrica 73*, 1815–1847.

Falkeborg, B. (2013). Essays in Financial Economics: On Trading Risk and Incentives. Ph.D. Dissertation, University of Copenhagen.

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

Glosten, L. and P. Milgrom (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics 14*, 71–100.

Green, E. J. (1987). Lending and the Smoothing of Uninsurable Income. In E. C. Prescott and N. Wallace (Eds.), *Contractual Arrangements for Intertemporal Trade*. Minneapolis: University of Minnesota Press.

Grossman, S. J. and M. H. Miller (1988). Liquidity and Market Structure. *The Journal of Finance 43*(3), 617–633.

Ho, T. and H. R. Stoll (1981). Optimal dealer pricing under transaction and return uncertainty. *Journal of Financial Economics 9*, 47–73.

Ho, T. S. Y. and H. R. Stoll (1983). The Dynamics of Dealer Markets Under Competition. *The Journal of Finance 38*(4), 1053–1074.

Jacod, J. and A. N. Shiryaev (2002). *Limit Theorems for Stochastic Processes*. Springer.

Kyle, A. S. (1985). Continuous Auctions and Insider Trading. *Econometrica 53*(6), 1315–1335.

Lagos, R. and G. Rocheteau (2009). Liquidity in asset markets with search frictions. *Econometrica 77*(2), 403–426.

Mildenstein, E. and H. Schleef (1983). The Optimal Pricing Policy of a Monopolistic Marketmaker in the Equity Market. *The Journal of Finance 38*(1), 218–231.

O'Hara, M. (1995). *Market Microstructure Theory*. Blackwell.

Roll, R. (1984). A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market. *The Journal of Finance 39*(4), 1127–1139.

Sannikov, Y. (2008). A Continuous-Time version of the Principal-Agent problem. *Review of Economic Studies 75*, 957–984.

Spear, S. and S. Srivastava (1987). On repeated moral hazard with discounting. *Review of Economic Studies 54*(4), 599–617.

Weill, P.-O. (2007). Leaning against the wind. *The Review of Economic Studies 74*(4), 1329–1354.