Liquidity with High-Frequency Market Making

Jungsuk Han[*], Mariana Khapko[†] and Albert S. Kyle[‡]

March, 2014

Abstract

We study a simple model of market making in which high-frequency market makers can cancel limit orders quickly after receiving an adverse signal. The resulting winner's curse induces low-frequency market makers to widen bid-ask spreads. Liquidity in the market may deteriorate unless high-frequency market makers fully replace low-frequency market makers in liquidity provision. Our result suggests that some restrictions on high-frequency trading, such as minimum resting times, may improve market liquidity by leveling the playing field among market makers with different speeds.

JEL Classification: D82, G14, G18

Keywords: high-frequency trading, market making, order cancellation, bid-ask spread, winner's curse, informed trading

[*] Corresponding Author: Department of Finance, Stockholm School of Economics, Drottninggatan 98, Stockholm, Sweden, SE 111-60, E-mail: jungsuk.han@hhs.se, Telephone: +46 8 736 9158, Fax: +46 8 312327.

[†] Stockholm School of Economics, Stockholm, Sweden, E-mail: mariana.khapko@hhs.se.

[‡] Robert H. Smith School of Business, University of Maryland, U.S.A, E-mail: akyle@rhsmith.umd.edu

1 Introduction

The empirical and theoretical finance literature has been engaged in an ongoing debate concerning whether high-frequency trading improves or harms market liquidity. The results are still inconclusive. Some empirical papers claim that high-frequency trading has a positive effect on market liquidity (e.g., Brogaard (2012) and Jovanovic and Menkveld (2012)), while others claim either ambiguous (e.g., Baron et al. (2013)) or negative effects on liquidity (e.g., Breckenfelder (2013)). Similarly, some theoretical papers suggest that high-frequency trading improves informational efficiency in the market (e.g., Li (2012)), but others suggest that market liquidity may be reduced by the participation of high-frequency traders due to increased adverse selection (e.g., Biais et al. (2013) and Budish et al. (2013)) or front-running (e.g., Li (2013)).

We study a simple theoretical model that shows how relative differences in order cancellation speeds affect bid-ask spreads. High-frequency market makers can cancel or update their quotes instantaneously. Low-frequency traders submit their quotes once at the beginning of a trading round and cannot cancel or update them afterwards. Low latency gives high-frequency traders an advantage over low-frequency traders if there is a release of value-relevant public information during the period when low-frequency traders cannot change bids and offers. When new public information arrives, high-frequency traders cancel their stale quotes and submit new updated quotes. Low-frequency traders incur adverse selection costs because their stale quotes have a higher probability of being hit when public information makes a trade less profitable and a lower probability of being hit when public information makes a trade more profitable.

In our analysis, we focus on quote adjustment during the period between the arrival of consecutive executable (market) orders. Like Glosten and Milgrom (1985), market makers post bids and offers which are executed against incoming order flow. Market makers face an adverse selection problem due to the fact that some of the incoming order flow contains information. Unlike Glosten and Milgrom (1985) we consider two types of market makers: (1) low-frequency market makers and (2) high-frequency market makers. Low-frequency market makers are always present in the market, submitting bids and offers expected to make non-negative profits. High-frequency traders are not always present in the market. Instead, high-frequency traders are present in the market with some assumed known probability reflecting an un-modeled economic cost or regulatory friction associated with their ability to participate with low latency.

We focus on how high-frequency traders exploit their ability to cancel and update orders quickly, leaving low-frequency traders to suffer the risk of trading at stale prices which do not reflect new public information. Theoretical papers (e.g., Ait-Sahalia and Saglam (2013), Biais et al. (2013), Budish et al. (2013) and Li (2013)) tend to focus on how low- and high- frequency traders trade with one another. In contrast, in our model high-frequency and low-frequency traders do not trade with one another because we assume the new public information implies quote updates smaller than the width of the bid-ask spread. Both high-frequency and low-frequency market makers choose quotes to avoid losses from trading against incoming informed order flow. This is consistent with the empirical literature which shows that high-frequency traders use their ability to cancel orders quickly to avoid losses (e.g., Hasbrouck and Saar (2010), Hasbrouck and Saar (2013)). It explains, for example, why human marker makers, operating at speeds with which eyeballs capture incoming signals and fingertips send outgoing messages, lose market share when competing with electronic market makers whose algorithms process and respond to information in milliseconds, microseconds, or nanoseconds.

When high-frequency traders are present in the market with some finite probability, their ability to cancel and update orders quickly increases the adverse selection problem for low-frequency market makers, who lose more good trades than bad ones. Low-frequency traders respond to this adverse selection problem by posting wider bid-ask spreads. The increased adverse selection results from order cancellation and updates by high-frequency traders in response to new public information, not from changes in the incoming order flow (which does not respond to new public information). If the probability of high-frequency market makers being present in the market equals zero, order cancellation does not increase adverse selection, and low-frequency traders post low break-even bid-ask spreads. If the probability of high-frequency traders' participating in the market equals one, the increased adverse selection from order cancellation also vanishes because the remaining market makers, all of whom have low latency, are all able to cancel and update quotes quickly in response to new public information; they therefore also post low bid-ask spreads on average. If the probability of high-frequency traders being present in the market is strictly greater than zero and strictly less than one, the bid-ask spread may be wider than when the probability is zero or one. We show that calculation of this wider bid-ask spread involves a non-trivial average of a relatively higher pre-information-arrival spread with a relatively lower post-information-arrival spread.

2 Model

Consider a simple trading model over time interval $[0, 1]$. The liquidation value $\tilde{V}$ of a single risky is high, V^G, with probability q or low, V^B, with probability of $1 - q$. The risk-free rate is normalized to zero. Public information about the risky asset in the form of a signal $\tilde{S} \in \{S^G, S^B\}$ may arrive at most once at a random time $s \in [0, 1]$, distributed as an exponential random variable with arrival rate parameter λ. The signal gives a correct indication of the asset value with probability $r > 1/2$.

A market order of unit size arrives at $t = 1$. It is submitted by either an informed trader, with probability ϕ, or a noise trader, with probability $1 - \phi$. An informed trader knows the true value of the asset, $\tilde{V}$, buys if this value is higher than the ask price, and sells if it is lower than the bid price. A noise trader submits an order to buy or sell one unit with equal probability, distributed independently from the fundamental value of the asset.

We assume a dealer market where the order flow is cleared by market makers (or dealers) who submit limit orders.[1] There are two types of market makers: low-frequency (or slow) and high-frequency (or fast). Numerous, competitive, risk-neutral low-frequency market makers are always present in the market, submit their bids and offers at $t = 0$, and are not able to alter them afterwards. When they submit their bids or offers, the slow market makers do not know whether fast market makers will subsequently be present in the market. With probability μ, numerous, competitive, risk-neutral high-frequency market makers arrive into the market and remain continuously present until the market order arrives at $t = 1$. When present in the market, the fast market makers are able to alter their limit orders instantaneously in response to a new public signal $\tilde{S}$. When the market order arrives at $t = 1$, it is matched to the limit order with the best bid or ask price. With probability $1 - \mu$, the fast market makers remain absent from the market—as a result of frictions such as technology costs—and the market order executes against the best bid or offer price posted by slow market makers.

At each point in time, each market maker posts at most one bid and one ask prices for one unit of the asset. The execution of an incoming market order is divided equally ("*pro rata*") among the market makers offering the best bid or offer price. Let p^L_{bid}

[1] That is, the market makers do not trade with one another by assumption. Alternatively, we can assume that r is not too big (so that quote updates are smaller than the width of the existing bid-ask spread).

and p^L_{ask} denote the bid and ask prices posted by the slow market makers, respectively. Similarly $p^H_{bid}(t)$ and $p^H_{ask}(t)$ denote quotes of the fast market makers at $t \in [0, 1]$. These bids and offers must be "feasible" in the sense that they are measurable with respect to the information set of market makers, which we denote $\mathcal{F}_t$. The joint distribution of the liquidation value $\tilde{V}$, the public signal $\tilde{S}$, and the market order is common knowledge. In addition, the information set $\mathcal{F}_t$ includes the public signal $\tilde{S}$ if this signal has arrived on or before date t.

3 Equilibrium

We define an equilibrium as a set of feasible quotes p^L_{bid}, p^L_{ask}, $p^H_{bid}(t)$ and $p^H_{ask}(t)$, such that all market makers expect non-negative profits from trading at equilibrium quotes and no market maker would expect strictly positive profits by posting different feasible quotes when all other market makers post equilibrium quotes. This definition focuses on an equilibrium in which the actions of the market makers are symmetric within the same class. We will show that all equilibria have the same execution prices, and quotes which have a positive probability of execution are unique.

The following proposition shows how the fast market makers always set their bid and ask prices equal to the expected value of the asset conditional on their information set. This implies that when new public information arrives, the fast market makers immediately cancel their existing stale limit orders and replace them with up-to-date orders reflecting new public information.[2] The width of the bid-ask spread is positive because the fast market makers also condition on the information content of the buy or sell order against which they execute (see, for example, Treynor (1995) for a detailed discussion on this).

Proposition 1. *The fast market makers always choose the following zero-profit bid and ask prices for any $t \in [0, 1]$:*

$$p^H_{bid}(t) = E[\tilde{V}|\mathcal{F}_t, sell];$$
$$p^H_{ask}(t) = E[\tilde{V}|\mathcal{F}_t, buy].$$

[2] Notice that the fast market makers do not need to change their quotes immediately; they can wait until any time before $t = 1$. However, we rule out those equilibria with such behaviors because it is not sustained even with small perturbations of the arrival time of the market order (e.g., assuming tiny randomness in the arrival time).

Proof. We prove the claim for the bid price. The same proof can be applied to the ask price due to symmetry. If $p^H_{bid}(t) > E[\tilde{V}|\mathcal{F}_t, \text{sell}]$, the expected profit is negative (i.e., $E[\tilde{V} - p^H_{bid}(t)|\mathcal{F}_t, \text{sell}] < 0$) whenever the order is executed at the bid price. This cannot be an equilibrium price with a positive probability of execution. If $p^H_{bid}(t) < E[\tilde{V}|\mathcal{F}_t, \text{sell}]$, then the expected profit is positive conditional on order execution. If there is a positive probability of executing against an incoming market order at this price, then a fast market maker could make a larger expected profit by changing the price to a slightly higher price (due to not having to divide the order *pro rata* with other fast market makers). Therefore, this also cannot be an equilibrium price. This finishes the proof. □

Let us now turn to the equilibrium quoting strategy of the slow market makers.

If the fast market makers are never present ($\mu = 0$), then it is an equilibrium for the slow market makers to set zero-profit quotes equal to the expected liquidation value conditional on whether the market order is a buy or sell, $p^L_{bid} = E[\tilde{V}|\text{sell}]$ and $p^L_{ask} = E[\tilde{V}|\text{buy}]$. The equilibrium becomes equivalent to a special case of Glosten and Milgrom (1985) with a single trading date in the sense that information is contained in the order flow and not in public information.

If the fast market makers are always present, then the slow market makers cannot possibly make a profit. The equilibrium quotes of the slow market makers are based on the worst possible public information associated with execution of their order, bad public information when buying and good public information when selling. Otherwise, the slow market makers suffer from a winner's curse because they would have an opportunity to trade only when it incurs a loss. We thus have $p^L_{bid} = E[\tilde{V}|S_B, \text{sell}]$ and $p^L_{ask} = E[S_G, \tilde{V}|\text{buy}]$. The slow market makers break even trading at these wide quotes (because they only trade when unfavorable public information prevents the trade from being profitable), a deviating slow market maker would lose money posting bids or offers tighter than this (because he would lose the profitable trades to fast market makers and execute only the unprofitable trades), and the slow market makers never trade posting wider quotes (because the fast market makers' quotes are always tighter).

If the fast market makers are present in the market with probability $\mu \in (0, 1)$, then the slow market makers face an adverse selection problem which forces them to widen their quotes beyond the case $\mu = 0$. The following proposition states that as the probability of participation by fast market makers μ increases from zero to one, the width of the quotes increases monotonically from the narrowest equilibrium quotes when

$\mu = 0$ to the widest equilibrium quotes when $\mu = 1$. The proposition also covers the polar cases $\mu = 0$ and $\mu = 1$.

Proposition 2. *Given $\mu \in [0,1]$, the slow market makers choose bid and ask prices as follows:*

$$p^L_{bid} = \omega_{bid}(\mu)E[\tilde{V}|sell] + (1-\omega_{bid}(\mu))E[\tilde{V}|S^B, sell];$$
$$p^L_{ask} = \omega_{ask}(\mu)E[\tilde{V}|buy] + (1-\omega_{ask}(\mu))E[\tilde{V}|S^G, buy],$$

where $\omega_{bid}(\cdot)$ and $\omega_{ask}(\cdot)$ are strictly decreasing functions in μ such that

$$\omega_{bid}(\mu) = \frac{(1-\mu)\left[\frac{1}{2}(1-\phi)+\phi(1-q)\right]}{(1-\mu)\left[\frac{1}{2}(1-\phi)+\phi(1-q)\right]+\mu(1-e^{-\lambda})\left[\frac{1}{2}(r(1+\phi)+q(1-\phi))-qr\right]};$$
$$\omega_{ask}(\mu) = \frac{(1-\mu)\left[\frac{1}{2}(1-\phi)+\phi q\right]}{(1-\mu)\left[\frac{1}{2}(1-\phi)+\phi q\right]+\mu(1-e^{-\lambda})\left[\frac{1}{2}(1-\phi)(1-r-q)+qr\right]}.$$

When the fast market makers partially provide liquidity (i.e., $\mu \in (0,1)$), the slow market makers will enjoy a strictly positive profit if it turns out that the fast market makers are not present, but will suffer losses otherwise. On average, they should have a zero expected profit because they are competitive. Therefore, the following should be true:

$$(1-\mu)P(\text{sell})\Big[E[\tilde{V}|\text{sell}] - p^L_{bid}\Big] + \mu P(s \le 1)P(S^B, sell)\Big[E[\tilde{V}|S^B, \text{sell}] - p^L_{bid}\Big] = 0.$$

From the zero profit condition, we obtain an explicit expression for the bid price of the slow market makers

$$p^L_{bid} = \omega_{bid}E[\tilde{V}|\text{sell}] + (1-\omega_{bid})E[\tilde{V}|S^B, \text{sell}],$$

where
$$\omega_{bid} = \frac{(1-\mu)P(\text{sell})}{(1-\mu)P(\text{sell})+\mu(1-e^{-\lambda})P(S^B,\text{sell})},$$
with $P(\text{sell}) = \frac{1}{2}(1-\phi)q + (\phi + \frac{1}{2}(1-\phi))(1-q)$ and $P(S^B, \text{sell}) = \frac{1}{2}(1-\phi)(1-r)q + (\phi + \frac{1}{2}(1-\phi))r(1-q)$. We can similarly prove the result for the ask price, and this finishes the proof of Proposition 2.

4 Liquidity

In this section we study the comparative statics properties of liquidity measures with respect to the probability of high-frequency trading μ. Let $\Delta_0(\mu)$ denote the bid-ask spread before the arrival of new information,

$$\Delta_0(\mu) = (1-\mu)\big(p^L_{ask} - p^L_{bid}\big) + \mu\big(p^H_{ask}(0) - p^H_{bid}(0)\big).$$

It can be shown that $\Delta_0(\mu)$ is hump-shaped with respect to μ, and the bid-ask spread becomes the narrowest when μ is either zero or one.[3] If information arrives before the market order, the expected bid-ask spread is equal to

$$\Delta_s(\mu) = (1-\mu)\big(p^L_{ask} - p^L_{bid}\big) + \mu\big[P(S^G)\big(p^L_{ask} - p^H_{bid}(s)\big) + P(S^B)\big(p^H_{ask}(s) - p^L_{bid}\big)\big].$$

The quotes of the fast market makers are updated on the arrival of information whereas those of the slow market makers stay stale. We can show that $\Delta_s(\mu)$ is lower than $\Delta_s(0)$ for any $\mu > 0$. That is, the presence of the fast market makers always improves post-announcement liquidity.

Given μ, we can represent market illiquidity (or the expected bid-ask spread) at any $t \in [0,1]$ as $\Delta_t(\mu) = \Delta_0(\mu)P(t < s) + \Delta_s(\mu)P(t \geq s)$. Then, we can show that the average market illiquidity over the time horizon, $\bar{\Delta}(\mu) \equiv \int_0^1 \Delta_t(\mu)dt$, is equal to[4]

$$\bar{\Delta}(\mu) = \Delta_0(\mu) - P(s \leq 1)(1 - E\left[s \mid s \leq 1\right])\left(\Delta_0(\mu) - \Delta_s(\mu)\right).$$

Notice that $\bar{\Delta}(\mu)$ becomes smaller if (i) information is more likely to arrive (i.e., $P(s \leq 1)$ is higher), (ii) information is likely to arrive earlier (i.e., $E\left[s \mid s \leq 1\right]$ is smaller), and (iii) market liquidity improves more significantly by the arrival of information (i.e., $\Delta_0(\mu) - \Delta_s(\mu)$ is larger).

To illustrate our results, we choose the following parameter values for numerical examples: $V^B = 0, V^G = 1, q = 0.5, \phi = 0.5$ and $r = 0.65$.

[3] Notice that $\frac{\partial \Delta_0}{\partial \mu}\big|_{\mu=0} > 0$, $\frac{\partial \Delta_0}{\partial \mu}\big|_{\mu=1} < 0$ and $\frac{\partial^2 \Delta_0}{\partial \mu^2} < 0$.

[4] This is so because $\bar{\Delta}(\mu) = \Delta_0(\mu) - (\Delta_0(\mu) - \Delta_s(\mu))\int_0^1 P(s \leq 1)P(t \geq s|s \leq 1)dt$.

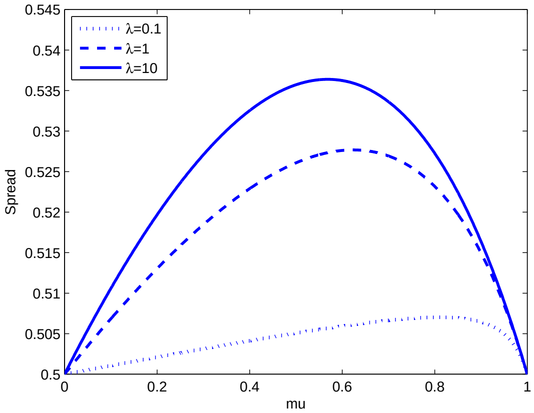

Figure 1: Bid-ask spreads before signal

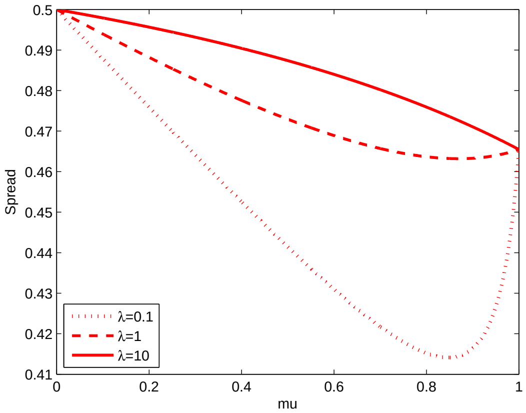

Figure 2: Bid-ask spreads after signal

Figure 1 illustrates that pre-announcement illiquidity has a hump shape with respect to μ, and Figure 2 illustrates that post-announcement illiquidity is lower for any value of μ if compared to zero probability of high-frequency trading. Figure 3 suggests a hump-shaped relationship between the average illiquidity and μ, and it is more pronounced for the intermediate values of λ (i.e., $\lambda = 1$). Notice that high-frequency trading is always beneficial when information arrives frequently (i.e., $\lambda = 10$), buy is mostly irrelevant to market liquidity if information arrives less frequently (i.e., $\lambda = 0.1$).

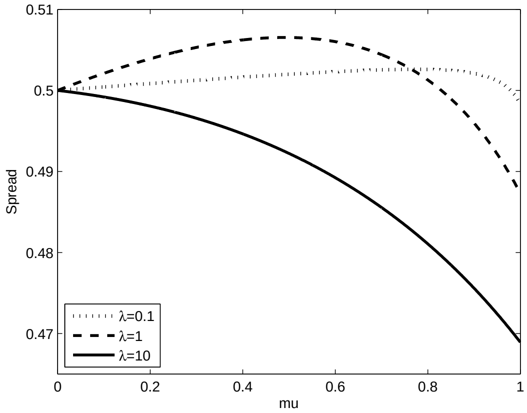

Figure 3: Average bid-ask spreads

High-frequency trading brings public information into prices faster, but it also drives out low-frequency trading by creating a winner's curse. Consequently, leveling the playing field among traders may create a trade-off between informational efficiency and liquidity provision by the low-frequency traders. Our results imply that this trade-off depends on the frequency of information arrival to financial markets.

5 Policy Implications

Our paper has several practical implications that contribute to the recent debate on regulating high-frequency trading. By introducing regulations that asymmetrically affect low- and high- frequency trading, regulators can either promote or limit high-frequency trading relative to traditional low-frequency trading. Our results imply that the optimum would be for high-frequency traders to be always present in the market if such presence were not costly. If high-frequency trading cannot fully replace low-frequency trading in terms of liquidity provision, imposing minimum resting times or cancellation fees may achieve improved market liquidity. Such restrictions can make the high-frequency traders less aggressive in posting their limit orders by limiting the opportunities to cancel orders. Thus, market liquidity may improve on average because reducing exposure to winner's curse allows the low-frequency traders to offer more competitive bid-ask spreads.

References

Ait-Sahalia, Yacine and Mehmet Saglam, "High Frequency Traders: Taking Advantage of Speed," 2013. Working Paper.

Baron, Matthew, Jonathan Brogaard, and Andrei Kirilenko, "The Trading Profits of High Frequency Traders," 2013. Working Paper.

Biais, Bruno, Thierry Foucault, and Sophie Moinas, "Equilibrium Fast Trading," 2013. Working Paper.

Breckenfelder, Johannes, "Competition between High-Frequency Traders, and Market Quality," 2013. Working Paper.

Brogaard, Jonathan, "High Frequency Trading and its Impact on Market Quality," 2012. Working Paper.

Budish, Eric, Peter Cramton, and John Shim, "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," 2013. Working Paper.

Glosten, Lawrence and Paul Milgrom, "Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders," *Journal of Financial Economics*, 1985, *14* (1), 71–100.

Hasbrouck, Joel and Gideon Saar, "Technology and Liquidity Provision: The Blurring of Traditional Definitions," *Journal of Financial Markets*, 2010, *12* (2), 143–172.

_ **and** _ , "Low-Latency Trading," *Journal of Financial Markets*, November 2013, *16* (4), 646–679.

Jovanovic, Boyan and Albert J. Menkveld, "Middlemen in Limit-Order Markets," 2012. Working Paper.

Li, Su, "Speculative Dynamics I: Imperfect Competition, and the Implications for High Frequency Trading," 2012. Working Paper.

Li, Wei, "High Frequency Trading with Speed Hierarchies," 2013. Working Paper.

Treynor, Jack, "The Only Game in Town," *Financial Analysts Journal*, 1995, *51* (1), 81–83.