Where is the Value in High Frequency Trading? *

Álvaro Cartea[†] and José Penalva[‡]

November 25, 2010

Abstract

We analyze the impact of high frequency trading in financial markets based on a model with three types of traders: liquidity traders, market makers, and high frequency traders. Our four main findings are: i) The price impact of the liquidity trades is higher in the presence of the high frequency trader and is increasing with the size of the trade. In particular, we show that the high frequency trader reduces (increases) the prices that liquidity traders receive when selling (buying) their equity holdings. ii) Although market makers also lose revenue to the high frequency trader in every trade, they are compensated for these losses by a higher liquidity premium. iii) High frequency trading increases the volatility of prices. iv) The volume of trades doubles as the high frequency trader intermediates all trades between the liquidity traders and market makers. This additional volume is a consequence of trades which are carefully tailored for surplus extraction and are neither driven by fundamentals nor is it noise trading.

Keywords: High frequency traders, high frequency trading, flash trading, liquidity traders, institutional investors, market microstructure

[*] We would like to thank Harrison Hong for his valuable comments and discussions. We are also grateful to Andrés Almazán, Gene Amromin, Pete Kyle and Eduardo Schwartz for their comments. The usual caveat applies. We welcome comments, including references we have inadvertently missed. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Banco de España.

[†] alvaro.cartea@uc3m.es, Universidad Carlos III de Madrid.

[‡] jpenalva@emp.uc3m.es, Universidad Carlos III de Madrid and Banco de España.

1 Introduction

Around 1970, Carver Mead coined the term "Moore's law" in reference to Moore's statement that transistor counts would double every year. There is some debate over whether this "law" is empirically valid but there is no discussion that the last forty years have seen an explosive growth in the power and performance of computers. Financial markets have not been immune to this technological advance, it may even be one of the places where the limits of computing power are tested every day. This computing power is harnessed to spot trends and exploit profit opportunities in and across financial markets. Its influence is so large that it has given rise to a new class of trading strategies called high frequency (HF) trading. The term HF trading is often used interchangeably with that of algorithmic trading (AT). We prefer to restrict the use of "HF trading" to refer to a subset of AT trading strategies that are characterized by their reliance on speed differences relative to other traders to make profits and also by the objective to hold essentially no asset inventories for more than a very short period of time.

The advent of AT has changed the trading landscape and the impact of their activities is at the core of regulatory and financial discussions. The explosion in volume of transactions we have witnessed in the last decade, and speed at which trades are taking place, is highly suggestive that AT is very much in use, and that these strategies are not being driven out of the market as a result of losses in their trading activities. Indeed, different sources estimate that annual profits from AT trading are between \$3 and \$21 billion (Brogaard [2010] and Kearns et al. [2010]). These strategies have supporters and detractors: on one side we find trading houses and hedge funds who vigorously defend their great social value, whilst being elusive about the profits they make from their use; and on the other hand there are trading houses that denounce high frequency traders (HFTs) as a threat to the financial system (and their bottom line). Although AT in general and HF trading in particular have been in the market supervisors' spotlight for quite some time, and efforts to understand the consequences of HF trading have stepped up since the 'Flash Crash' in May 6 2010 (SEC [2010], Commission et al. [2010], and Kirilenko et al. [2010]) there is very little academic work that addresses the role of these trading strategies. The objective of this paper is to provide a framework with which to analyze the issues surrounding HF trading, their widespread use, and their value to different market participants.

To analyze the impact of HF trading in financial markets we develop a model with three types of traders: liquidity traders (LTs), market makers (MMs), and HFTs. In this model LTs experience a liquidity shock and come to the market to unwind their positions which are temporarily held by the MMs in exchange of a liquidity premium. On the other hand, the HFT has the ability to process information and execute trades more quickly than anybody else in the market; an ability that allows her to act as an intermediary and extract part of the trading surplus from the transactions between LTs and MMs. In addition, if there are no HFTs in the market, our model reduces to that of Grossman and Miller [1988] which we use as benchmark to analyze some of the consequences of HF trading.

We highlight our main findings. First, the price impact of the liquidity trades is higher in the presence of the HFT and is increasing with the size of the trade. In particular, we show that the HFT reduces (increases) the prices that LTs receive when selling (buying) their equity holdings. Second, although the MMs also lose revenue to the HFT in every trade, the price they pay to purchase shares is lower than the price paid in the absence of the HFT, and the price at which MMs sell is also lower than that received if there were no HFTs extracting trading surplus. However, they are compensated by a higher liquidity premium. Third, HF trading increases the volatility of prices. Fourth, we show that the volume of trades doubles because the HFT act as an intermediary for all trades between the LTs and MMs. The additional volume is neither driven by fundamentals (only the original trades, without the HFT, are driven by fundamentals) nor is it noise trading. Far from it, the extra volume is a consequence of trades which are carefully tailored for surplus extraction.

Two contemporaneous empirical papers lend strong support to the stylized features that our theoretical model captures as well as the implications concerning the impact that HF trading has on financial markets. The recent work of Zhang [2010] firmly concludes that HF trading increases stock price volatility and that this positive correlation between volatility and HF trading "is stronger for stocks with high institutional holdings, a result consistent with the view that high-frequency traders often take advantage of large trades by institutional investors".

Kirilenko et al. [2010] study the impact of HF trading during the Flash Crash on May 6 2010. Their findings about the activities of HFTs also provide strong support for the theoretical description we use to include HFTs as pure surplus extractors in our theoretical model. They find that HFTs have among all types of traders the highest price impact and that "HFTs are able to buy right as the prices are about to increase. HFTs then turn around and begin selling 10 to 20 seconds after a price increase." Moreover, they find that "The Intermediaries sell when the immediate prices are rising, and buy if the prices 3-9 seconds before were rising. These regression results suggest that, possibly due to their slower speed or inability to anticipate possible changes in prices, Intermediaries buy when the prices are already falling and sell when the prices are already rising." These findings strongly support our assumption that HFTs (due to their speed advantage) can for the most part effectively anticipate and react to price changes as a key part in their strategies for surplus extraction.

Before delving into our analysis of HF trading, we provide a brief overview of HF trading and HFTs, what HFTs could be doing, and what is it about trading speed that is so profitable for some and damaging for others, we do this in Section (2). After this quick overview, in Section (3) we develop our framework and analysis. In Section (4) we use the model to discuss the main issues raised by the presence of HFTs and in Section (5) we conclude and discuss some key features about HFTs that require further research (and quality data).

2 Trading Algorithms, High Frequency Traders, and Financial Markets

2.1 Financial Market Developments

Over the last years all major exchanges have revamped their systems to give way to the new era of computerized trading. Speed of trading and volume figures speak for themselves. In the SEC's report on "Findings regarding the market events of may 6, 2010" (SEC [2010]) we read that NYSE's average speed of execution for small, immediately executable orders was 10.1 seconds in January 2005, compared to 0.7 seconds in October 2009. Also, consolidated average daily share volume in NYSE-listed stocks was 2.1 billion shares in 2005, compared to 5.9 billion shares in January through October 2009. Consolidated average daily trades in NYSE-listed stocks was 2.9 million trades in 2005, compared to 22.1 million trades in January through October 2009. Consolidated average trade size in NYSE-listed stocks was 724 shares in 2005, compared to 268 shares in January through October 2009.

Other important metrics that intend to capture market efficiency and information transmission have also undergone considerable changes as a result of modifications of market rules and the prominent role that computing has taken in financial markets. For example, Chordia et al. [2010] focus on comparisons of pre- and post-decimal trading in NYSE-listed stocks (subperiods from 1993-2000 and 2001-2008). Some of their findings are that average effective spreads decreased significantly (from \$0.1022 to \$0.0223 cents for small trades (<\$10,000) and from \$0.1069 to \$ 0.0267 for large trades (>\$10,000)), while average depth available at the inside bid and offer declined significantly (from 11,130 shares to 2,797 shares). From 1993-2000 the mean trade size is \$82,900 and from 2001-2008 \$36,400 while the mean number of transactions is 1,136 and 14,779 respectively.

2.2 What characterizes Algorithmic and HF Trading

We adopt Hendershott et al. [2010]'s definition of AT: "the use of computer algorithms to automatically make certain trading decisions, submit orders, and manage those orders after submission". We distinguish HF trading as a subset of AT. An HF trading strategy is an algorithmic trading strategy that is based on exploiting greater processing and execution speed to obtain trading profits while holding essentially no asset inventory over a very short time span—usually measured in seconds, mostly less than a few minutes, and certainly less than a day.[1] One sometimes finds these strategies described as *latency arbitrage*. HF traders (HFTs) are proprietary firms and proprietary trading desks in investment banks, hedge funds, etc, that based on these strategies have the ability to generate large amounts of trades over short periods of time, Cvitanić and Kirilenko [2010]. There are other AT strategies that are in use for other

[1] In their study of the 'Flash Crash', Kirilenko et al. [2010] find that, holding prices constant, HFTs reduce half their net holdings in 115 seconds.

purposes. For example, there are AT liquidity strategies which strategically post and cancel orders in the order book to exploit widening spreads, or AT strategies designed to execute large orders with the smallest price impact. Our analysis focuses exclusively on HF trading strategies, which we believe are the ones most critics of AT have in mind.

Making the distinction between HF trading and AT is important because it highlights the substantial difficulty one encounters when measuring the impact that HFTs have on markets according to metrics such as volume, spreads, and liquidity. For example, estimates of HF trading volume in equity markets vary widely depending on the year or how they are calculated, but they are typically between 50% and 77% of total volume, see SEC [2010] and Brogaard [2010]—although how much is actual HF trading versus generic AT is unclear. Also, Hendershott et al. [2010] find that for large-cap stocks AT improves liquidity and narrows effective spreads. They also find that AT increases realized spreads which indicates that revenue to liquidity suppliers has increased with AT, but it is difficult to infer how much of these effects are due to AT that is not HF trading. Similar identification problems are present in another recent study, Brogaard [2010], which finds that HFTs contribute to price discovery and reduce volatility. Thus, this identification problem, as well as possible collateral effects on other AT strategies, have to be taken into account in any regulatory implications one may draw from our analysis, as we focus exclusively on HF trading.

As per our definition, paramount to the activities of HF traders is the speed at which they can: access and process market information; generate, route, cancel, and execute orders; and, position orders at the front of the queue in the trading book so as to avoid having stale quotes in the market. Their speed or low latency is mainly due to two key ingredients: capacity (software and hardware), and co-location. Co-location allows HFTs to place their servers in close physical proximity to the matching engines of the exchanges. Surprisingly, being near the exchanges can shave the speed of reaction by a sufficient number of fractions of a second to provide HFTs a valuable edge when trading in the market—to the extent that they are willing to pay millions of dollars for this service.

Perhaps the most revealing behavior of HFTs is how they make use of cancelations to poke the market and extract valuable information. For instance, the strategy known as 'pinging' is based on submitting immediate-or-cancel orders which are used by HFTs to search for and access all types of undisplayed liquidity SEC [2010].[2] Another strategy, known as ‘spoofing’, consists of sending out a large amount of orders over a short period of time before immediately canceling most of them and only a few are executed. This burst of activity is expected to trigger other algorithms to join the race and start buying or selling (and slow down information flows to other market participants).

2.3 How is it that HFTs could be making money?

Here we provide four stylized examples that show how HFTs could be exploiting their speed advantage by posting, executing, and canceling orders so as to position their orders at front of the queue and intermediate in market transactions for a negligible period of time. Although the first three examples outline different strategies used by HFTs, the underlying feature common to all three is the ability that HFTs have to extract trading surplus.

Example 1. One way in which HFTs can extract surplus is by exploiting their speed to alter market conditions in a way that encourages buyers to accept a slightly higher price and sellers a slightly lower one. The idea is relatively simple and works in a setting where liquidity traders split their trades in small packages and market makers do not have large outstanding offers in the books. Suppose a trader (LT) needs liquidity and wants to sell a block of shares. As the first shares come into the system (say at the best buy price of $5.50 per share), the HFT cancels her outstanding posted buy offers that have not been executed. She then posts additional sell offers, adding to the increased selling pressure, so as to help clear the remaining posted buy orders in the book. Once the book is clear, the HFT quickly reposts a significant number of offers at lower prices (say $5.48) so that she is first in the buying queue. This is only possible if she can move quickly enough that by the time the market maker (MM) reacts to the increased selling pressure, new posted offers by the MM sit behind those posted by the HFT at $5.48 per share. The liquidity trader finds that the market around $5.50 has dried up and can only sell at $5.48. These shares are bought by the HFT, who is at the front of the

[2] There are circumstances when orders are placed close to best buy or sell with no intention to trade, this is known as book layering, and up to 90% of these orders are immediately canceled, see SEC [2010].

queue. Also, having posted substantial orders at the front of the queue at \$5.48 allows the HFT to be the first to notice when the liquidity pressure eases. At that moment, the HFT cancels her posted buy offers and starts posting sell offers at slightly higher prices (say \$5.49). The MM sees the selling pressure shift and the price rebound. Although the MM is sorry to see the lower prices disappear, he is still willing to buy at \$5.49, which allows the HFT to sell her earlier purchases and end up with a zero net position. During this process, the HFT has been able to make profits on the shares bought at \$5.48 and sold at \$5.49, while taking a loss on the shares executed at the beginning, bought at \$5.50 and sold at \$5.49. An HFT who is fast enough will have bought many more shares at \$5.48 than at \$5.50 by selectively canceling and reposting her offers to make the strategy profitable. As for the other traders: The MM bought some shares below the initial best buy price so he is satisfied; the other liquidity trader is also satisfied from having been able to sell his shares even though for some of them he received a cent less than what the MM paid for them.

Example 2. Assume that the market opens after the release of good news about a company's performance. At opening, shares are selling at \$5.50 and slow traders (traders who are not HFT) are posting buy orders. Due to the high latency of the slow traders, HFTs see the buy orders arriving in the system and decide to purchase as many shares as possible at the current price. Immediately, HFTs issue low volume immediate-or-cancel sell orders to gauge how much are slow traders willing to pay for the shares. For example, an immediate-or-cancel sell order goes out at a price of \$6.00 per share and if it does not get filled it is immediately canceled and a new immediate-or-cancel sell order is sent at \$5.99 and so on until it is filled, say at \$5.70. At this point the HFTs unload all shares that were purchased making a profit and holding no inventories.

Example 3. Similarly, HFTs may take advantage of the so-called flash orders which give them an informational edge over other traders in the market. For example, a buy order for 1,000 shares at \$5.50 is 'flashed' to a reduced number of traders some of which are HFTs. Traders that are flashed the information not only know about such a potential trade before the vast majority of the market, but they are also capable of acting upon the information before it reaches the rest of the market. If the HFTs believe or are able to correctly anticipate that this buy order is part of a large lot that will trickle through the system in small batches of, say 1,000 shares at a time, they have time to react and purchase all shares in the market; use sell-or-cancel orders to find out the upper limit at which the counterparty that initiated the buy order is willing to pay to liquidate his entire lot of shares; and complete the round trip of buying and selling whilst making a profit and carrying no inventory, all of this in the intraday market, and possibly within a couple of minutes.

Example 4. Other trades that could be profitable for HFTs, but do not profit from extracting surplus, are those that are designed to collect rebates offered by the exchange. Market centers attract volume by offering a liquidity rebate to MMs that post orders. Rebate strategies that break even by selling (buying) and then buying (selling) shares at the same price are profitable because they earn the rebate for providing liquidity. If the rebate is around 0.25 cents per share then a round trip rebate accrues half a cent per share to the HFT that pursues these type of rebate trading.

We note that although the examples above have been contrived so that the HFTs make positive profits these strategies are not arbitrages. This is because there are states of the world where they can also deliver a loss. These strategies are not riskless and the HFTs face different types of risk, for instance volume and price risk. In all cases the HFT must take a view on what side of the market she will initially take and how many shares and at what prices she is willing to buy (sell) before turning around to sell (buy) her holdings with the objective of not carrying inventories for too long. Here we take the view (supported by the findings in Kirilenko et al. [2010]) that as a result of their speed and their ability to post and cancel orders, it is presumed that more often than not the HFT will earn a profit, or break even, and be flat after a short period of time. It is the value of these trades we want to analyze.

3 The Model

Our analysis sets up a framework in which a stock market has social value because it facilitates the financing of economic activity, adding value to equity holders by providing a way for them to convert their equity into cash (and viceversa) quickly and at a reasonable price. This idea is captured by the Grossman-Miller (GM) model, Grossman and Miller [1988]. In the GM model, equity investors (liquidity traders) quickly find counterparties for their trading needs.[3] These counterparties are MMs who are willing to take the other side of investor trades and hold those assets temporarily–until another investor enters the market to eliminate the temporary order imbalance. Holding these assets entails price risk, which MMs are willing to bear in exchange for a small payment. This payment is calculated as the difference between the price at which the transaction takes place and the expected value of the traded asset. We introduce HFTs in this model, where an HFT is a trader who, thanks to her rapid information processing and quick execution ability, can extract part of the trading surplus from the transaction between equity investors and MMs. However, even though our analysis centers on the effect of HFT's surplus extraction due to a temporary trading imbalance, it also applies to a broader set of circumstances. In particular it applies to a much larger number of market participants, which includes mutual fund managers, hedge funds, insurance companies and other large investors and for a broader range of transactions, not only immediate liquidity needs, but also other trades executed either to build up or unwind an asset position, for hedging, etc.

The setting for our model is a simple world with three dates, $t = 1, 2, 3$ in which a temporary order imbalance of size i affects conditions in a stock market with a cash asset (with a return normalized to zero) and a risky asset. There are two "outside investors", which we refer to as liquidity traders (LT1 and LT2). LT1 wants to sell i shares of the risky asset at date 1, while LT2 wants to buy i shares at date 2 (trade $-i$ shares).[4] If both traders met at the same time the asset price would not be affected by i but, as they do not, the M intermediaries, the MMs, accept LT1's order at date 1 and hold that position until LT2 enters the market at date 2. Date 3 serves as a reference point to determine the expected cash value of the asset, which is described by the random variable P_3. The asset has price risk as public information about P_3 enters the market at dates one and two. Thus, MMs face price risk between dates one and two, and in order to compensate for this risk when holding the assets, market prices will adjust.

[3] The other interpretation of the model is intended for the futures market. Everything we say for the stock market is also equally valid for trading in the futures market in exactly the same terms. We refer the interested reader to the original article for details of how to reinterpret the model.

[4] We make the assumption that LT1 wants to sell and LT2 wants to buy to streamline the presentation. The analysis is equally valid if LT1 wants to buy i shares and LT2 wants to sell them.

All these traders (LT1, LT2, and MMs) are price-taking and risk averse. They have expected utility from wealth at date 3 so that they choose their asset positions so as to maximize $\mathbb{E}\left[U\left(W_3\right)\right]$, where $U\left(W\right) = -\exp\left(-aW\right)$. We assume that market prices at each date, P_t, $t = 1, 2, 3$, are normally distributed, and we use the notation $\theta_t^x\left(P_t\right)$ to describe the demand for assets by trader x at date t for a given price P_t ($\theta_t^x > 0$ implies holding (long) a positive number of shares).

3.1 HF trading and the Extraction of Trading Surplus

For simplicity we will assume that there is a single, monopolistic HFT in the market. In fact there are several HFTs in the market which represent a very small fraction of active traders (out of the almost 12,000 active traders during the flash crahs, Kirilenko et al. [2010] identify 16 as HFTs). Also, entry is limited by high investment costs (co-location and hardware, but more importantly, access to algorithms and detailed data) and these act as effective barriers to entry. Thus, it is reasonable to conclude that HFTs can currently exercise a substantial amount of monopolistic power.

The HFT exploits her greater speed to extract trading surplus from transactions at each date. This is modeled in a highly stylized way by allowing the HFT to take over any trader's position in two blocks: one block is executed at the market price, P_t, and the other at the market price plus or minus a haircut, Δ. The profits the HFT can obtain from this ability depend on market conditions and is illustrated by the following simplified example which is followed below by the description of the general model.

Suppose that a liquidity trader (LT) wants to sell $1,000$ shares, the expected price of the asset is \$5.50 and there are nine MMs. Suppose further that all traders are risk averse and they have linear demands for assets given by

$$\theta^{MM}\left(P\right) = \theta^{LT}\left(P\right) = 5000\left(5.50 - P\right)$$

so that they will only buy shares if the price is below \$5.50.

In a trading equilibrium, the holdings of all traders (one LT and nine MMs) have to add up to the supply of LT (1000 shares) so that

$$\begin{aligned} 9\theta^{MM}(P) + \theta^{LT}(P) &= 1000 \\ \Rightarrow 9 \times 5000\,(5.50 - P) + 5000\,(5.50 - P) &= 1000 \\ \Longleftrightarrow P &= 5.48\,. \end{aligned}$$

At the price of \$5.48 per share, each trader (including the LT) will hold 100 shares. This implies that (the price sensitive) LT who wanted to sell 1,000 shares (at \$5.50) ends up selling 900 shares at a price of \$5.48.

The HFT can enter the market and make profits by introducing a haircut, say of 1 cent per share, inducing the LT to sell part of his shares at a price of \$5.47 and inducing MMs to buy part of the package at the slightly higher price of 5.49 (using one of the strategies we discussed in Section 2.3).

The HFT's profits are determined by traders' asset demands. At a price of \$5.47, LT will hold $5000\,(5.50 - 5.47) = 150$ shares, thus selling only 850 shares. At a price of \$5.49, MMs will hold $5000\,(5.50 - 5.49) = 50$ shares each, that is 450 shares. These do not cancel and implies the HFT would be left holding 400 shares, but the HFT wants to hold zero inventory at the end of the transaction, so the HFTs strategy is: After offering to buy 850 shares from LT at 5.47, she offers to buy another batch of 50 shares from the LT at \$5.48. The LT finds these trades acceptable and executes them. Having acquired all LT's shares, the HFT turns around and first offers to sell 50 shares to each of the MMs at \$5.49. Then, she offers to sell another 50 shares at \$5.48. Again, this combination of offers is acceptable to the MMs and they agree to execute them. The final result is that the HFT has made 2 cents on each of 450 shares sold at \$5.49 (and bought at \$5.47) and another cent on 400 shares bought at \$5.47 and sold at \$5.48 for a total of 13 dollars. Furthermore, she ends up holding no inventory. The market has seen a doubling of the trading volume, plus some transactions taking place at prices above and below the "market price" of the asset.

This example represents what happens in the extended general model from which we can derive equilibrium prices, haircuts, volumes and the effect of the HFT on the number of market makers in the economy. Although the example (and the model) is highly stylized—it exaggerates the profits that an HFT could extract from the transaction, and involves knowledge that is not directly available to any trader—it does capture the advantage gained by HFTs through their speed when it comes to extracting information from the order flow and using their execution speed to cancel and advantageously repost trades in the order book, as was discussed above.

3.2 The general model: Trading at $t = 2$

To solve the general model we construct each trader's problem by backward induction. The second liquidity trader, LT2, enters at $t = 2$ and wants to buy i shares. He knows that he will not be able to acquire all the shares he demands at the "market price" $(i + \theta_2^{LT2}(P_2))$ but will have to pay the haircut, Δ_2^{LT2} for some trades, where Δ_t^x denotes the haircut applied to trader x at trading date t–this notation allows us to maintain a generic analysis and postpone for later a discussion of how the HFT may (or may not) discriminate between different traders and trading dates. We assume traders accept this haircut as part of the cost of doing business.[5] Then, let $i + \tilde{\theta}_2^{LT2}\left(P_2 + \Delta_2^{LT2}\right)$ denote the number of shares he acquires at the (higher) price, $P_2 + \Delta_2^{LT2}$, so that at the market price he acquires $\theta_2^{LT2}(P_2) - \tilde{\theta}_2^{LT2}\left(P_2 + \Delta_2^{LT2}\right)$.

Also, given all current information, the date 3 value of the asset P_3 is distributed normally with mean μ and variance σ^2. Let $\mathbb{E}[X|\mathcal{F}_t]$ denote the expectation of X conditional on public information at date t. Then, given an initial wealth of W_2^{LT2}, LT2's final wealth will be

$$W_3^{LT2} \;=\; W_2^{LT2} + \theta_2^{LT2}(P_2)\,P_3 - \left(\theta_2^{LT2}(P_2) + i\right)P_2 - \left(\tilde{\theta}_2^{LT2}\left(P_2 + \Delta_2^{LT2}\right) + i\right)\Delta_2^{LT2},$$

[5] It is possible to introduce a strategic element whereby the trader alters his bidding behavior in anticipation of the HFT. This can greatly complicate the model but will only affect how much of the trading surplus is extracted by the HFT, without any qualitative changes in the results.

which, if we drop the superscripts and the functional dependence of asset holding on prices, is

$$W_3 = W_2 + \theta_2 P_3 - (\theta_2 + i) P_2 - \left(\tilde{\theta}_2 + i\right) \Delta_2.$$

Substituting for wealth in the utility function, we obtain:

$$\mathbb{E}\left[U(W_3)|\mathcal{F}_2\right] = -\exp\left(-a\left(\theta_2\left[\mu - P_2\right] - iP_2 - W_2 - \left(\tilde{\theta}_2 + i\right)\Delta_2\right) + \frac{1}{2}a^2\sigma^2(\theta_2)^2\right).$$

Trader LT2's final asset demand, θ_2^{LT2}, is such that it maximizes this expression. Hence,

$$\theta_2^{LT2}(P_2) = \frac{\mu - P_2}{a\sigma^2}. \tag{1}$$

Similarly, LT1's and MMs' asset demands at $t = 2$ can be computed and we obtain

$$\theta_2^{LT1}(P_2) = \theta_2^{MM}(P_2) = \frac{\mu - P_2}{a\sigma^2}. \tag{2}$$

Market clearing requires that changes in asset holdings equal to zero:

$$\left(i + \theta_2^{LT2}\right) + \left(\theta_2^{LT1} - \theta_1^{LT1}\right) + M\left(\theta_2^{MM} - \theta_1^{MM}\right) = 0.$$

Using the date 1 market clearing condition $\left(i - \theta_1^{LT1}\right) - M\theta_1^{MM} = 0$ implies that

$$\theta_2^{LT2} + \theta_2^{LT1} + M\theta_2^{MM} = 0,$$

so that the equilibrium price is $P_2 = \mu$, and the optimal asset positions at the end of period 2 are $\theta_2^{LT2} = \theta_2^{LT1} = \theta_2^{MM} = 0$. The HFT, on the other hand, will extract trading surplus from the assets transacted equal to[6]

$$\left|\tilde{\theta}_2^{LT2}\left(P_2 + \Delta_2^{LT2}\right) + i\right|\Delta_2^{LT2} + \left|\tilde{\theta}_2^{LT1}\left(P_2 - \Delta_2^{LT1}\right) - \theta_1^{LT1}\right|\Delta_2^{LT1} + M\left|\tilde{\theta}_2^{MM}\left(P_2 - \Delta_2^{MM}\right) - \theta_1^{MM}\right|\Delta_2^{MM}.$$

[6] We are assuming that the market makers and LT1 are net sellers of the asset. This is confirmed as in equilibrium MMs will be net buyers in the first period and have zero net final holdings (and $\theta_2^{LT1} - \theta_1^{LT1}$ has the same sign as MMs' changes in asset holdings at date two).

3.3 The general model: Trading at $t = 1$

Traders at date $t = 1$ anticipate what will happen at $t = 2$. They also know that there will be public information revealed prior to date two trading so that P_2 is normally distributed with mean μ and variance σ^2 (we keep a constant variance solely to reduce notational clutter).

Traders' future wealth can be written as

$$\begin{aligned} W_3^{LT1} &= W_0^{LT1} + \theta_2^{LT1} P_3 + \left(\theta_1^{LT1} - \theta_2^{LT1}\right) P_2 - \left(\theta_1^{LT1} - \tilde{\theta}_2^{LT1}\right) \Delta_2^{LT1} - \theta_1^{LT1} P_1 + \left(i - \tilde{\theta}_1^{LT1}\right) \Delta_1^{LT1} \,, \\ W_3^{MM} &= W_0^{MM} + \theta_2^{MM} P_3 + \left(\theta_1^{MM} - \theta_2^{MM}\right) P_2 - \left(\theta_1^{MM} - \tilde{\theta}_2^{MM}\right) \Delta_2^{MM} - \theta_1^{MM} P_1 - \tilde{\theta}_1^{MM} \Delta_1^{MM} \,. \quad (3) \end{aligned}$$

Using $\mathbb{E}\left[P_2|\mathcal{F}_1\right] = \mathbb{E}\left[\mathbb{E}\left[P_3|\mathcal{F}_2\right]|\mathcal{F}_1\right] = \mu$, simplifies the expression for traders' wealth, and it is straightforward to derive optimal demands:

$$\begin{aligned} \theta_1^{LT1} &= \frac{1}{a\sigma^2}\left(\mu - P_1 - \Delta_2^{LT1}\right) \,, \quad (4) \\ \theta_1^{MM} &= \frac{1}{a\sigma^2}\left(\mu - P_1 - \Delta_2^{MM}\right) \,. \quad (5) \end{aligned}$$

Notice that traders anticipate that their current asset demand (and hence positions at the end of trading at $t = 1$) will affect future trading and hence the haircuts they will have to pay. The market clearing condition is now

$$-i + \theta_1^{LT1} + M\theta_1^{MM} = 0 \,.$$

From this we obtain the market clearing price:

$$P_1 = \mu - \frac{\Delta_2^{LT1} + M\Delta_2^{MM} + ia\sigma^2}{M+1} \,, \quad (6)$$

and traders' asset demand:

$$\begin{aligned} \theta_1^{LT1} &= \frac{i}{M+1} + \frac{1}{a\sigma^2}\frac{M}{M+1}\left(\Delta_2^{MM} - \Delta_2^{LT1}\right) \,, \quad (7) \\ \theta_1^{MM} &= \frac{i}{M+1} - \frac{1}{a\sigma^2}\frac{1}{M+1}\left(\Delta_2^{MM} - \Delta_2^{LT1}\right) \,. \quad (8) \end{aligned}$$

Again, the HFT is able to extract trading surplus and generate profits of

$$\left|i - \tilde{\theta}_1^{LT1}\left(P_1 - \Delta_1^{LT1}\right)\right| \Delta_1^{LT1} + M \left|\tilde{\theta}_2^{MM}\left(P_1 + \Delta_1^{MM}\right)\right| \Delta_1^{MM},$$

where

$$i - \tilde{\theta}_1^{LT1}\left(P_1 - \Delta_1^{LT1}\right) = \frac{M}{M+1} i - \frac{M}{M+1} \frac{\Delta_2^{MM} - \Delta_2^{LT1}}{a\sigma^2} + \frac{\Delta_1^{LT1}}{a\sigma^2}\,.$$

In summary, from the above analysis we extract the following conclusions:

Theorem 3.1. *For a given order imbalance of magnitude i:*

1. *Market clearing prices are*

$$P_1 = \mu - \frac{\Delta_2^{LT1} + M\Delta_2^{MM} + ia\sigma^2}{M+1}, \quad P_2 = \mu\,.$$

2. *Asset trading is:*

Price	*LT1*	*LT2*	*MM (total)*
$P_1 - \Delta_1^{LT1}$	$-\frac{M}{M+1}\left(i - \frac{\Delta_2^{MM} - \Delta_2^{LT1}}{a\sigma^2}\right) + \frac{\Delta_1^{LT1}}{a\sigma^2}$		
P_1	$-\frac{\Delta_1^{LT1}}{a\sigma^2}$		$M\frac{\Delta_1^{MM}}{a\sigma^2}$
$P_1 + \Delta_1^{MM}$			$\frac{M}{M+1}\left(i - \frac{\Delta_2^{MM} - \Delta_2^{LT1}}{a\sigma^2}\right) - M\frac{\Delta_1^{MM}}{a\sigma^2}$
$P_2 - \Delta$		$i - \frac{\Delta_2^{LT2}}{a\sigma^2}$	
P_2	-$\frac{\Delta_2^{LT1}}{a\sigma^2}$	$\frac{\Delta_2^{LT2}}{a\sigma^2}$	$-M\frac{\Delta_2^{MM}}{a\sigma^2}$
$P_2 + \Delta_2^{LT1}$	$-\frac{1}{M+1}\left(i + M\frac{\Delta_2^{MM} - \Delta_2^{LT1}}{a\sigma^2}\right) + \frac{\Delta_2^{LT1}}{a\sigma^2}$		
$P_2 + \Delta_2^{MM}$			$-\frac{M}{M+1}\left(i - \frac{\Delta_2^{MM} - \Delta_2^{LT1}}{a\sigma^2}\right) + M\frac{\Delta_2^{MM}}{a\sigma^2}$

The HFT acts as counterparty to all these trades.

Thus, the HFT affects the market clearing price. Traders, anticipating the haircuts imposed by the HFT, amplify the selling pressure of the liquidity trader, driving the market clearing price in the first period down further than it would have been without HF trading.

Also, the presence of the HFT introduces additional prices at which transaction takes place. These price movements around the market clearing price can be interpreted as additional microstructure volatility. Furthermore, it can be seen from the table that most of the volume takes place around the market clearing price, and very little *at* the market clearing price, which may lead to erroneous inference about the "true" market price of the asset at each date. In addition, we find that

Corollary 3.2. *The number of assets that change hands is twice what it would have been without the HFT.*

And,

Corollary 3.3. *If there is a rebate of c cents per share, the HFT get half of all rebates, and the M market makers split a quarter of all rebates between them.*

3.4 Optimal Haircuts

The HFT, as a monopolist, realizes that there is a trade-off between a larger haircut, and a smaller number of assets traded using that haircut. Thus, she will adjust her behavior by setting haircuts that maximize the profits of her trading activity.

So far we have used notation that allows the HFT to distinguish between different traders and different periods of time. Of course, this is a highly stylized interpretation of what might be going on in markets, and the amount of information the HFT can extract from the order flow. Nevertheless, we do not believe that the order flow allows HFTs to condition on such fine distinctions between traders and trading times. We do, however, believe that HFTs can extract sufficient information to distinguish a large trade coming in (the initial trades by LT1 and LT2) from regular (and smaller) trades from other market participants (MMs and, at $t = 2$, LT1). Thus, we will focus our analysis on the case where the HFT discriminates according to relative size of orders and sets one haircut for large trades Δ_L, and one for small trades Δ_S. The haircut for the large trades is

$$\Delta_L = \Delta_1^{LT1} = \Delta_2^{LT2} \, ,$$

and for the small trades:

$$\Delta_S = \Delta_1^{MM} = \Delta_2^{MM} = \Delta_2^{LT1} \, .$$

In the Appendix we discuss another case (where the HFTs make no distinctions and apply the same haircut to all trades).

Given these assumptions on the HFTs informational advantages, the HFTs problem is to maximize profits, which from Theorem 3.1 are:

$$\Delta_L \left(\frac{M}{M+1} i - \frac{\Delta_L}{a\sigma^2} \right) + \Delta_S \left(\frac{M}{M+1} i - M \frac{\Delta_S}{a\sigma^2} \right) + \Delta_L \left(i - \frac{\Delta_L}{a\sigma^2} \right) + \Delta_S \left(i - \frac{(M+1)\,\Delta_S}{a\sigma^2} \right) .$$

Lemma 3.4. *The optimal haircuts are*

$$\begin{aligned} \Delta_S &= \frac{1}{2} \frac{1}{(M+1)} i a \sigma^2 , \\ \Delta_L &= \frac{1}{4} \frac{2M+1}{M+1} i a \sigma^2 = \Delta_S \left(M + \frac{1}{2} \right) . \end{aligned}$$

This result backs the intuition that the haircut imposed on large trades is bigger than that applied to small trades, and the increase in haircut for large trades is proportional to the (average) number of other, non-liquidity seeking, market participants.

Moreover, we see that with more MMs present in the market, the HFT increases the haircut imposed on the LT1 in date 1 and on LT2 in date 2 ($\Delta_L \to 1/2$ and $\Delta_S \to 0$ as $M \to \infty$). Figure 1 below shows that the haircut for the large (small) trades is increasing (decreasing) in M.

Interestingly, the initial holdings of LT1 are the same as without a HFT:

$$\theta_1^{LT1} = \frac{i}{M+1} .$$

Figure 2 shows the optimal asset holdings, θ_t and $\tilde{\theta}_t$, for the liquidity traders and MMs.

Figure 1: Optimal Haircuts for large and small trades

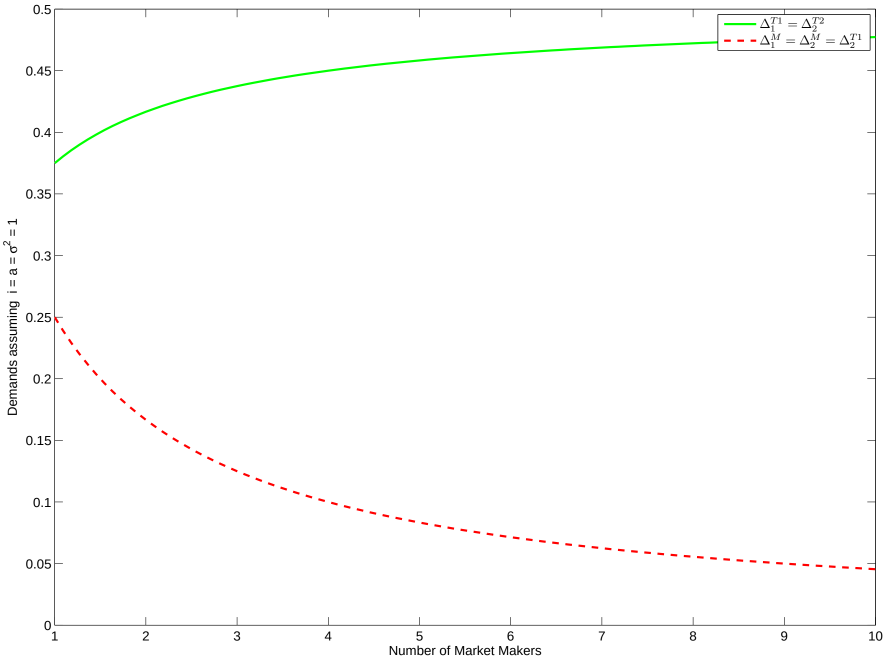
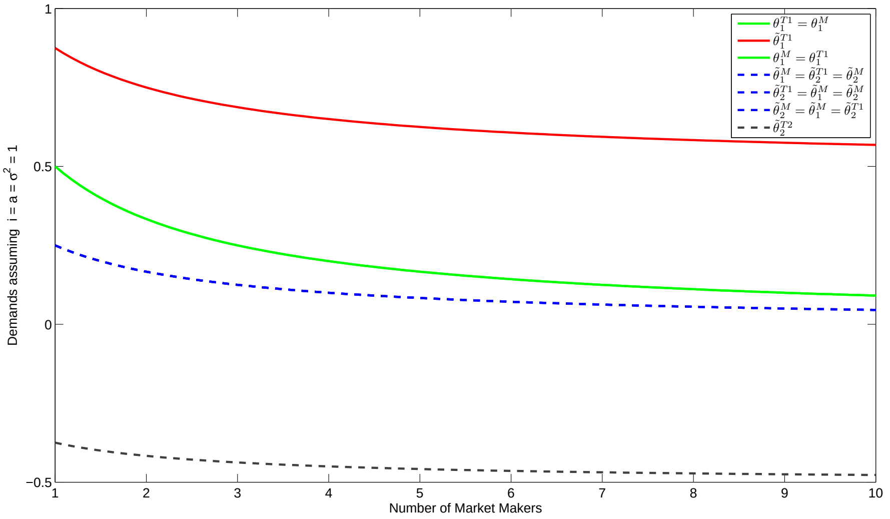


Figure 2: Optimal asset holdings when the HFT sets a haircut for large trades and another haircut for small trades.



4 Measuring the impact of HFTs on financial markets

To the best of our knowledge there are only two academic articles that have measured the impact of AT in the financial markets.[7] As discussed in the introduction Hendershott et al. [2010] find that AT improves liquidity and narrows effective spreads and Brogaard [2010] finds that HFTs contribute to price discovery and reduce volatility. Although the study of Brogaard [2010] has data of what he labels as HFTs, his data set cannot differentiate how much of the activity of the 26 firms in his study is AT and how much is HF trading; the same applies to the results of Hendershott et al. [2010].

4.1 Basic findings

In the absence of ultra-high-frequency data that provides details of HFTs trades, postings, cancelations, flash-trades, etc, the second best is to use our model to show how the presence of HFTs affects different financial metrics such as: volume, liquidity, price impact, and price volatility. We summarize four basic findings for the particular case analyzed above: the HFT sets two types of haircuts depending on whether trades are large or small (we discuss the implications of the case where the HFT applies a unique haircut to all trades she intermediates in the Appendix).

First, we show that the volume of trades doubles because the HFT intermediates all trades between the LTs and MMs. The additional volume is neither driven by fundamentals (only the original trades, without the HFT, are driven by fundamentals) nor is it noise trading. Far from it, the extra volume is a consequence of trades which are carefully tailored for surplus extraction. Second, the price impact of liquidity trades is higher in the presence of the HFT and is increasing with the size of the trade. In particular, we have seen that the HFT depresses the average prices the LT1 receives in dates 1 and 2, and that the increase in the liquidity premium and the haircut is proportional to the size of the trade. Interestingly, the equilibrium quantities that the LT1 sells in date 1 is not affected by the presence of the HFT because the

[7] In Section 2 we discussed the work of Kirilenko et al. [2010] which focuses on the Flash Crash and employs three days of data.

HFT intermediates every trade in order to ensure it does not end up with a non-zero stock of assets. Third, although the MMs also pay a haircut in each transaction, the average price they pay to purchase shares at date 1 is lower than the price paid in the absence of the HFT, and the average price at which MMs sell at date 2 is also lower than that received if there were no HFTs extracting trading surplus. The overall effect of lower buy prices and lower sell prices for MMs is that expected returns from holding inventories from date 1 to 2 is higher when a HFT extracts surplus on both rounds of trades. However, the higher return does not imply greater expected profits for MMs. Fourth, HF trading increases the volatility of prices.

4.2 Number of Market Makers

Our model allows us to derive the impact of HFT's surplus extraction activities on the number of MMs, and hence on the "true" supply of liquidity in the market. We proceed as in GM and assume that the cost of entry for the MM is c. This cost is sunk before knowing the liquidity shock i which we assume to be normally distributed and independent of the other shocks affecting prices. At time $t = 0$ the expected utility of an individual MM is $\mathbb{E}\left[U\left(W_3^{MM} - c\right)\middle|\mathcal{F}_0\right]$, where W_3^{MM} is given by (3). Free entry of MM will occur until

$$\mathbb{E}\left[U\left(W_3^{MM} - c\right)\middle|\mathcal{F}_0\right] = \mathbb{E}\left[U\left(W_0^{MM}\right)\middle|\mathcal{F}_0\right] , \tag{9}$$

and recall that $\theta_2^{MM} = 0$. The expected value and variance of wealth in period three are

$$\mathbb{E}\left[W_3\middle|\mathcal{F}_1\right] = W_0^{MM} - c + a\sigma^2\left(\theta_1^{MM}\right)^2 - \theta_1^{M}\Delta_1^{MM} + \frac{1}{a\sigma^2}\left(\left(\Delta_2^{MM}\right)^2 + \left(\Delta_1^{MM}\right)^2\right)$$

and

$$\mathrm{Var}\left[W_3^{M}\middle|\mathcal{F}_1\right] = \sigma^2\left(\theta_1^{MM}\right)^2 . \tag{10}$$

Hence the entry condition (9) becomes

$$e^{ac}\mathbb{E}\left[e^{-\frac{1}{2}a^2\sigma^2\left(\theta_1^{MM}\right)^2 + a\theta_1^{M}\Delta_1^{MM} - \frac{1}{\sigma^2}\left(\left(\Delta_2^{MM}\right)^2 + \left(\Delta_1^{MM}\right)^2\right)} \middle| \mathcal{F}_1 \right] \quad = \quad 1\,. \tag{11}$$

In Equation (11) we can see the effect of the HFT. First, we can observe that the risk assumed by the MM depends on his final holdings at date 1 (θ_1^{MM}). The HFT's desire to hold zero assets at the end of each period implies that it will eventually unload all assets and (in our model) leave the MM with the same asset holdings as if the HFT had not been there. Then, overall, the risk assumed by the MM will be the same. On the other hand, the HFT distorts prices. In Equation (11) we can see that the MM makes some extra profit as the liquidity premium, the difference between expected value and market clearing price, increases (by $a\theta_1^{M}\Delta_1^{MM}$). But the MM also loses revenue, twice: one from each of the haircuts at dates $t = 1$ and $t = 2$ ($-\frac{1}{\sigma^2}\left(\left(\Delta_2^{MM}\right)^2 + \left(\Delta_1^{MM}\right)^2\right)$).

Using (5) we can express the optimal haircut for small trades as $\Delta_S = \theta_1^{MM}\frac{1}{2}a\sigma^2$. Substituting $\Delta_1^{MM} = \Delta_2^{MM} = \Delta_S$ in (11) the entry condition becomes

$$e^{ac}\mathbb{E}\left[e^{-\frac{1}{2}a^2\sigma^2\left(\theta_1^{MM}\right)^2} \middle| \mathcal{F}_1 \right] = 1\,,$$

which indicates that the gain from the lower price at date 1 exactly offsets the haircuts paid to the HFT in the round trade (buy from LT1 and sell to LT2). And, the number of MMs will be the same with or without HFT. The exact canceling of the two effects occurs because of the particular information assumptions we have made which allows the HFT to distinguish only large from small trades. In the Appendix we discuss the case where the HFT does not distinguish trades and charges the same haircut to all participants. In that case, the number of MMs increases, as the liquidity premium effect is greater than the haircuts.

4.3 Liquidity

Effective liquidity for a liquidity trader is measured by the market impact cost of his trades. This is usually hard to determine as the identity of traders is not available for empirical analysis, so that most studies focus on measuring liquidity in two ways. One measure of liquidity is how much of the quantity i the price sensitive liquidity trader (LT1) is willing to sell in period 1 at the market clearing price, in reaction to the price impact of his trade. According to this 'quantity metric' liquid markets would be those where LT1 carries the smallest holdings over to the next period (that is θ_1^{T1} is smallest —recall that LT1 sells $\left(i-\theta_1^{T1}\right)$ in period 1). In the particular case analyzed here, where the HFT sets a haircut for small trades and a haircut for large trades, the LT1 carries over an amount $\theta_1^{T1}=i/(M+1)$ which is the same as what LT1 carries over if there were no HFT, so that the presence of the HFT would not alter liquidity at all.

The other measure of liquidity is to focus on price impact. As we have seen above, our analysis indicates that price impact might be a better metric. This is because although LT1 sells the same amount in period 1 with and without an HFT, the amount of cash he obtains (and hence the liquidity he attains from the sale) is much lower due to the presence of the HFT—both because of the direct loss from the haircuts, but also the additional losses from lower market clearing prices. Thus, it is liquidity traders, and not MMs (whose expected utility and numbers stay the same), the ones who take the brunt of the presence of HFTs.

Using our model, we can calculate the effect of HFTs on the prices and returns faced by all traders, and compare them to those obtained in the absence of a HFT.

When there is no HFT the market clearing price (for buyers and sellers) are:

$$P_1^{\Delta=0}=\mu-\frac{ia\sigma^2}{M+1},\quad P_2^{\Delta=0}=\mu\,.$$

This implies that MMs receive a liquidity premium at date $t=1$ of $\frac{ia\sigma^2}{M+1}$, and obtain an expected return of

$$\mathbb{E}\left[r^{\Delta=0}\middle|\,\mathcal{F}_1\right]=\frac{ia\sigma^2}{M+1}\frac{1}{P_1^{\Delta=0}}\,. \tag{12}$$

On the other hand, when there is an HFT, buyers and sellers face different prices. The MM pays a haircut to the HFT in date 1 (when he buys $\tilde{\theta}_1^M$at $P_1 + \Delta_1^{MM}$ per share, and then $\theta_1^{MM} - \tilde{\theta}_1^{MM}$ at P_1 per share) so that the average price is

$$\begin{aligned} \bar{P}_1^M &= P_1 + \frac{\tilde{\theta}_1^M}{\theta_1^M}\Delta_1^M = \mu - \frac{ia\sigma^2}{M+1} - \frac{1}{4}\frac{ia\sigma^2}{M+1} \\ &= P_1^{\Delta=0} - \frac{1}{4}\frac{ia\sigma^2}{M+1}\,. \end{aligned} \tag{13}$$

From which we see that he obtains a 25% higher liquidity premium (after including the haircut paid to the HFT). Similarly, at date 2

$$\bar{P}_2^M = \mu - \Delta_2^M + \frac{\tilde{\theta}_2^M}{\theta_1^M}\Delta_2^M = P_2^{\Delta=0} - \frac{1}{4}\frac{ia\sigma^2}{M+1}\,, \tag{14}$$

so that he ends up selling back at a lower price (relative to the case without HFT), and thereby losing the initial extra liquidity premium garnered from the first transaction. Note that if we compute the expected excess return obtained by the MM: $r = \left(\bar{P}_2^M / \bar{P}_1^M\right) - 1$ we have

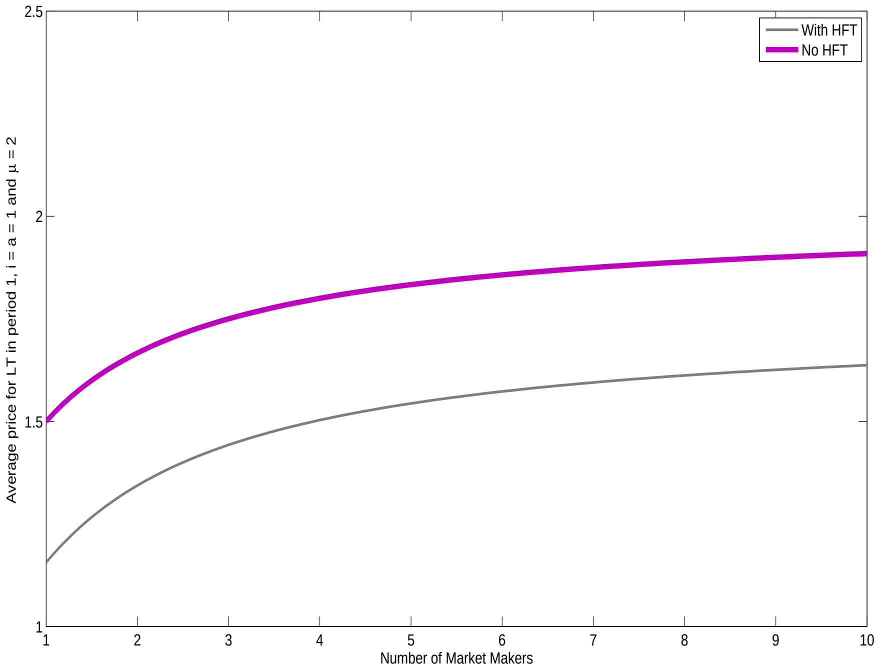
$$\mathbb{E}_1[r] = \frac{ia\sigma^2}{M+1}\frac{1}{\bar{P}_1^M} > \mathbb{E}_1\left[r^{\Delta=0}\right]\,.$$

That is, the expected returns in the presence of HFT are higher than in the absence of HFT. As mentioned above, this higher return is purely an artifact of the lowering of both buy and sell prices—a similar (though reversed) effect would appear if the initial liquidity imbalance had a different sign (that is, if $i < 0$).

As for liquidity traders, the HFT extracts surplus from LT1 in date 1 (by first buying $i - \tilde{\theta}_1^{LT1}$ at $P_1 - \Delta_1^{LT1}$ and then $\tilde{\theta}_1^{LT1} - \theta_1^{LT1}$ at P_1). The the average price received by T1 at date 1 (Figure 3) is

$$\bar{P}_1^{T1} = P_1 - \frac{i - \tilde{\theta}_1^{T1}}{i - \theta_1^{T1}}\Delta_1^{T1}\,. \tag{15}$$

Figure 3: Average price for LT1 in period 1



At date 2, after the haircut the average price he sells his remaining assets is:

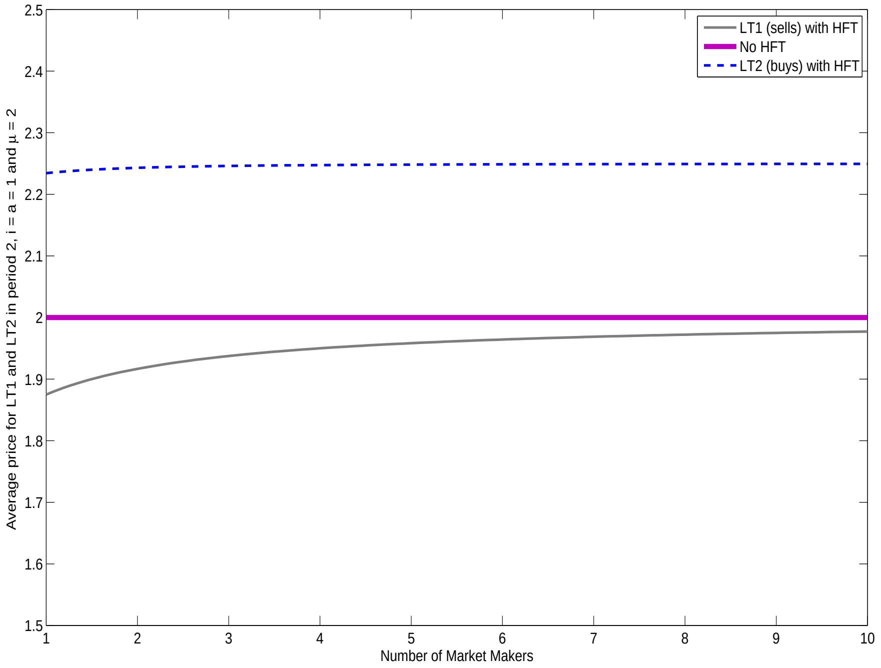
$$\bar{P}_2^{T1} \quad = \quad \mu - \Delta_2^{T1} + \frac{\tilde{\theta}_2^{T1}}{\theta_1^{T1}} \Delta_2^{T1} = P_2^{\Delta=0} - \frac{1}{4} \frac{ia\sigma^2}{M+1} \,. \tag{16}$$

From Equation (16) we can see that the liquidity premium increases in proportion to the size of the trade. This implies that LT1 receives less money for liquidating his position at both dates (relative to the case without HFTs) as can be seen in Figure 3. Similarly, the average price paid by LT2 in period 2 is

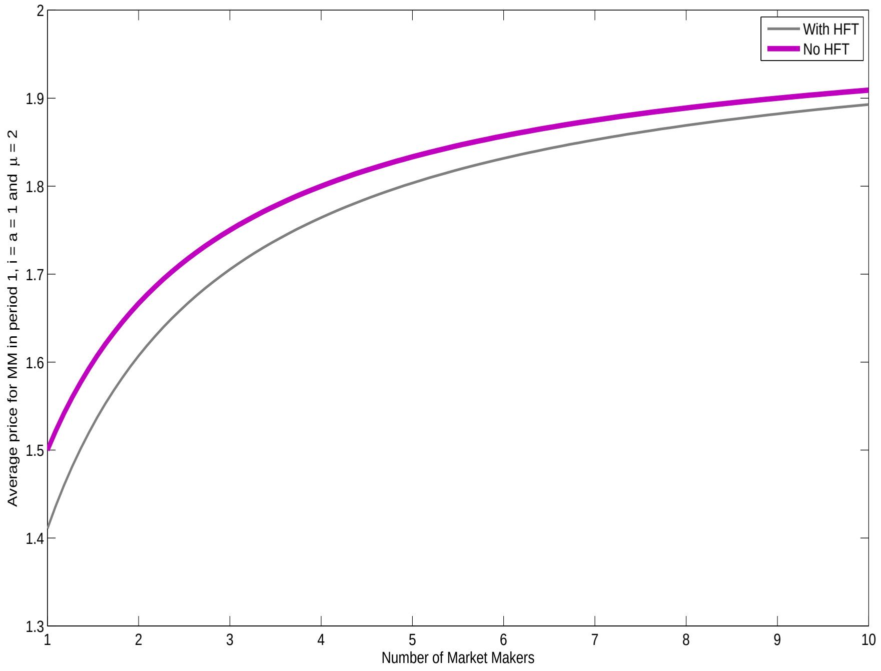
$$\bar{P}_2^{T2} = \mu - \Delta_2^{T2} + \frac{\tilde{\theta}_2^{T2}}{i} \Delta_2^{T2} \,. \tag{17}$$

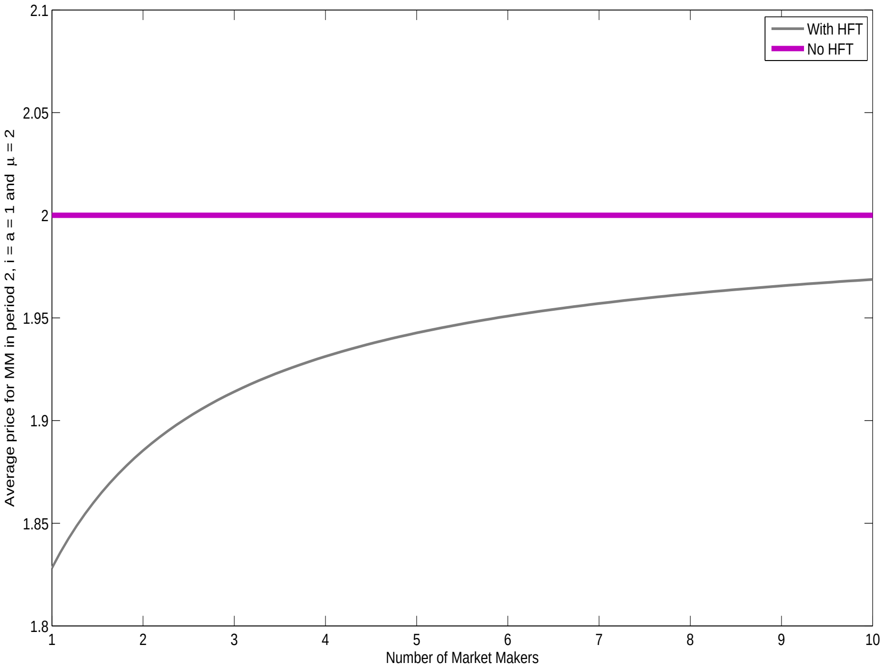
So that LT2 also suffers as she pays more to acquire her position than when there is no HFT. Note that LT1's asset demand at date 1 is distorted by the presence of the HFT, as he anticipates that he will pay a (lower) haircut when selling at date 2.

Figure 4: Average price for LT1 (sells) and LT2 (buys) in period 2



We see that the liquidity traders, LT1 and LT2, bear the brunt of the haircut imposed by the monopolist HFT. Although MMs suffer the loss of consumer (date 1) and supplier (date 2) surpluses, the liquidity premium received by the MMs for shares at date 1 is greater (the equilibrium price is lower) than in the absence of the HFT, and the net effect on his expected wealth is zero.


Average price for MM in period 1, i = a = 1 and μ = 2
Number of Market Makers
With HFT
No HFT
2
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1
2
3
4
5
6
7
8
9
10


Average price for MM in period 2, i = a = 1 and μ = 2
Number of Market Makers
With HFT
No HFT
2.1
2.05
2
1.95
1.9
1.85
1.8
1
2
3
4
5
6
7
8
9
10

4.4 Price volatility

When there is no HFT intermediating transactions between LTs and MMs there is one transaction price at date 1, $P_1^{\Delta=0}$, and another transaction price at date 2, μ. However, when a monopolistic HFT intermediates all transactions the tape will record four prices in date 1, $\left\{P_1 - \Delta_1^{T1},\, P_1,\, P_1 + \Delta_1^M,\, P_1\right\}$, and four (five or six if $\Delta_2^{MM} \neq \Delta_2^{LT1}$) in date 2, $\left\{\mu - \Delta_2^{MM},\, \mu,\, \mu + \Delta_2^{T2},\, \mu\right\}$. Therefore, it is inevitable to observe price volatility within dates 1 and 2 which is solely caused by the HFT's presence. Figure 5 shows the volatility of prices in dates 1 and 2 that result from the HFT intermediating all trades.[8] Furthermore, Figure 6 depict the standard deviation of prices at both dates. We see that price volatility is much higher when an HFT operates in the markets.

Note however that in our model this increased volatility does not generate additional risk for traders. Traders, in their evaluation of price risk evaluation in their objective functions, recognize that the microstructure noise generated by the HFT translates into a deterministic effect on their execution costs while leaving unaffected their final holdings.[9]

5 Conclusions

We have studied the impact of a monopolistic HFT that extracts trading surplus in transactions between liquidity traders and market makers. Clearly, the HFT comes out making profits. As a monopolist, she would enter the market only if these profits exceed the necessary entry costs. Introducing competition between HFTs would reduce these profits until HFT profits just cover investment costs for marginal HFTs. Thus the social value of HFT profits revolves around the

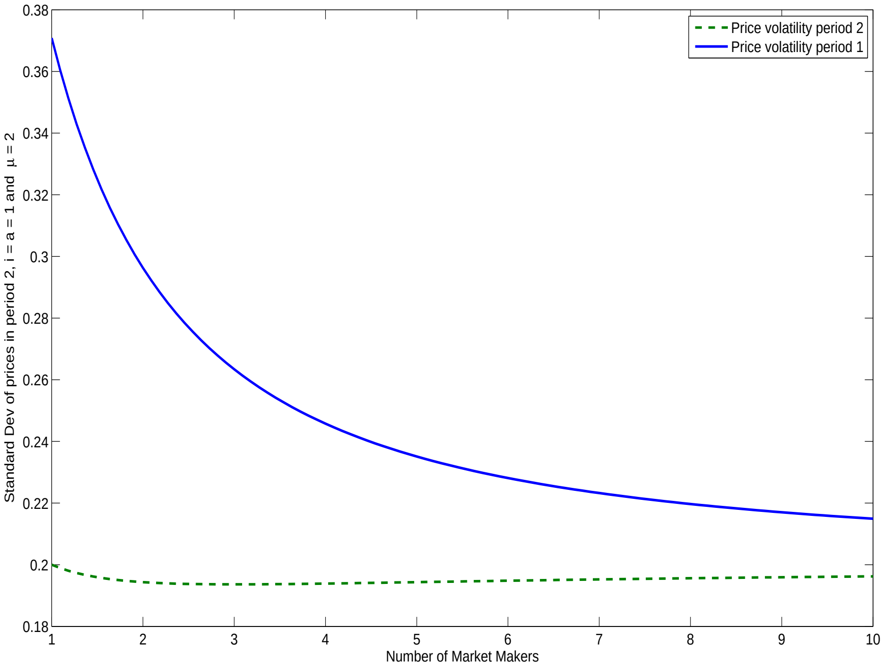
[8] To calculate the mean and variance of prices we also use the price at time $t = 0$ and assume it is the fundamental value μ, thus the mean and variance of the price in date 1 are

$$\bar{P}_1 = \frac{\mu + 4P_1 + \left(\Delta_1^M - \Delta_1^{T1}\right)}{5} \quad \text{and} \quad \mathbb{V}\left[P_1\right] = \frac{1}{5}\left\{\left(\mu - \bar{P}_1\right)^2 + \left(P_1 - \Delta_1^{T1} - \bar{P}_1\right)^2 + 2\left(P_1 - \bar{P}_1\right)^2 + \left(P_1 + \Delta_1^M - \bar{P}_1\right)^2\right\}.$$

Similarly, to calculate the mean and variance of prices with date 2 we assume that the first observation is μ.

[9] This changes if we allow for the HFT to (randomly) miss the opportunity to intermediate some trades between LT and MMs. While adding such a feature will make the model more realistic, it complicates the analysis substantially without much additional insight.

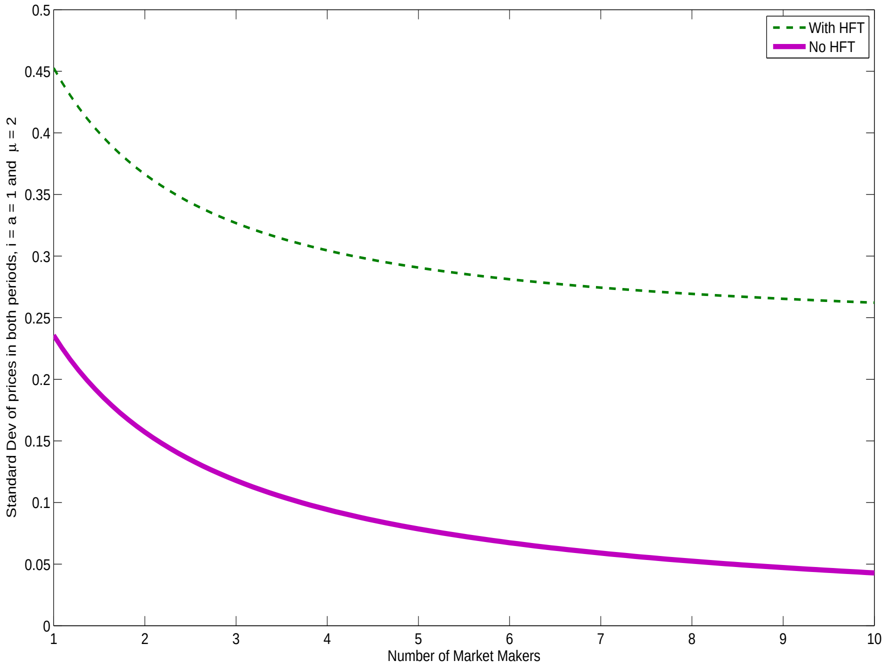
Figure 5: Volatility of prices in dates 1 and 2 induced by HF trading



issue of whether the existence of HFTs, and hence their investments, is valuable for the market as a whole.

As we have shown, the presence of HFTs doubles trading volume, increases the price impact of liquidity trades (in proportion to the size of the trade), has no effect on the number of traders (although this effect depends on the way HFTs vary their haircuts across trades), and increases price volatility. Thus, the presence of HFTs distorts market conditions, not through the amount of shares traded, but through prices. By exacerbating price impact, the HFT induces additional market impact costs on participants, especially the liquidity trader. This cost is proportional to the size of the trade which implies that large liquidity traders, such as institutional investors trading to change the composition of a portfolio, are the most affected by the presence of HFTs–an effect that is consistent with Zhang [2010]'s findings. In the particular case we have analyzed, it is the liquidity trader that bears all the costs from the presence of the HFT. Market makers find themselves unaffected because the revenue they lose to the HFT is compensated by a higher liquidity premium. Overall, the value of the stock market as a forum for providing a way for

Figure 6: Standard deviation of prices in dates 1 and 2 with and without HF trading



investors to convert their equity into cash (and viceversa) quickly and at a reasonable price falls because of the adverse effect the HFT has on prices. An aspect we have left unmodeled is the time to execution faced by liquidity traders. Clearly, if the HFT is to intermediate between liquidity trader and market makers, it must act quicker and hence execution time for liquidity traders must be lower. Whether the additional speed compensates for the additional trading costs is something that requires better data, but also it is something that traders are facing little choice on.

An issue that arises from our analysis is the question of how to measure (socially valuable) liquidity, when the analyst has no access to the identity of traders and hence cannot determine directly the market impact costs of trades. Clearly, just adding up the number of trades in the presence of rent-seeking hyperfast algorithms whose asset positions are essentially zero most of the time seems like a poor measure of the ability of the market to provide prompt and fair value to investors. Also, HFTs who generate additional microstructure noise at smaller intervals and accelerate market transactions raises several questions regarding how to measure price impact, whether we should adjust current measures to account for the increase in the speed of execution, and whether these measures adequately capture an investor's cost of executing a trade.

Finally, although in our knife-edge case HFTs had no effect on market makers. In the Appendix we consider an alternative parameterization where HFTs charge the same haircut per trade, and the effect is to increase the number of traders. This suggests that the overall effect on the "true liquidity providers" in not at all clear, and we need good empirical work to (a) determine the speed-cost effect on outside investors and liquidity traders, (b) determine if HFTs are raising the cost of business for market makers, and (c) if they do, whether the liquidity they provide is a good substitute for the one that is being driven out. HFTs clearly generate costs, but they also generate benefits, and the net effect requires more empirical analysis.

References

Jonathan Brogaard. High frequency trading and its impact on market quality. *SSRN Working Paper*, 2010.

Tarun Chordia, Richard Roll, and Avanidhar Subrahmanyam. Recent trends in trading activity. *SSRN Working Paper*, 2010.

U.S. Commodity Futures Trading Commission, the U.S. Securities, and Exchange Commission. Findings regarding the market events of may 6, 2010. Report, SEC, September 2010.

Jaksa Cvitanić and Andrei A. Kirilenko. High Frequency Traders and Asset Prices. *SSRN eLibrary*, 2010.

Sanford J. Grossman and Merton H. Miller. Liquidity and market structure. *Journal of Finance*, 43(3):617–37, July 1988.

Terrence Hendershott, Charles M. Jones, and Albert J. Menkveld. Does algorithmic trading improve liquidity? *Journal of Finance*, Forthcoming, 2010.

Michael Kearns, Alex Kulesza, and Yuriy Nevmyvaka. Empirical limitations on high frequency trading profitability. *SSRN worjing papers*, 2010.

Andrei A. Kirilenko, Albert (Pete) S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. The Flash Crash: The Impact of High Frequency Trading on an Electronic Market. *SSRN eLibrary*, 2010.

SEC. Concept release on equity market structure. Concept Release No. 34-61358; File No. S7-02-10, SEC, January 2010. 17 CFR PART 242.

Frank Zhang. The Effect of High-Frequency Trading on Stock Volatility and Price Discovery. *SSRN eLibrary*, 2010.

A Appendix

Above we argued that the HFT can discriminate across order size when trades come to the market which enables her to apply haircuts for large and small trades. If the HFT is not able to discriminate by size or type of trader she can still exercise her monopoly power by applying one haircut to all trades she intermediates. Hence we can repeat the analysis above while setting the following haircuts:

$$\Delta = \Delta_1^{T1} = \Delta_2^{T1} = \Delta_1^{MM} = \Delta_2^{MM} = \Delta_2^{T2}\,.$$

Then, the optimal delta set my the HFT is:

$$\Delta = ia\sigma^2 \frac{(2M+1)}{(M+1)\,(3+2M)}\,. \tag{18}$$

As in the case with two haircuts discussed in the body of the paper, LT1's optimal holding is $\theta_1^{T1} = i/(M+1)$, but as we show below, the equilibrium number of MMs in this case is *higher* than without HFT.

Other results are similar to those discussed above: First, the volatility of realized transaction prices increases. Second, the price impact of the liquidity trades in both periods is substantial: equilibrium sell (buy) prices are lower (higher) than the competitive price in the absence of the HFT. Third, the expected returns that the MM face from buying in period 1 and selling in period 2 increase (at the expense of traders), and the equilibrium number of MMs present in the market increases when compared to the number of MMs that are present in a market without a HFT. Fourth, the total volume of trades doubles relative to the number of trades observed in the absence of the HFT.

In the interest of space we only discuss the equilibrium number of MMs. We proceed as above where the entry condition (9) becomes

$$e^{ac}\mathbb{E}\left[e^{-\frac{1}{2}a^2\sigma^2\left(\frac{i}{M+1}\right)^2\beta(M)}\right] \quad = \quad 1 \qquad \text{where} \qquad \beta(M) = \frac{12M^2+12M+7}{(3+2M)^2}\,. \tag{19}$$

By inspecting (19) we know that the number of MMs with and without an HFT will be the same only when $\beta(M) = 1$ and this occurs for $M = 1/2$. When $M > 1/2$ we have more MMs in the presence of HFT. If we denote the number of MMs by M and the number of MMs in the absence of HFT by $M_{\Delta=0}$ we can show that

$$M_{\Delta} = \sqrt{\beta\left(M\right)}\left(M_{\Delta=0} + 1\right) - 1\,,$$

hence we have that $M > M_{\Delta=0}$. (The function $\beta(M)$ is increasing in M and we are interested in values for $M > 1$).