Stealth trading in modern high-frequency markets

Alejandro Garcia, Thomas Kelly, Joan Segui

Abstract

This paper aims to investigate stealth trading in modern US equity markets. Stealth trading by informed participants has been documented as the primary mechanism by which private information is incorporated into market prices, thus driving price discovery. This paper extends the stealth trading literature by analyzing several different trade characteristics with a particular focus on the role of high frequency traders. First, we analyze how trade size relates to price formation, building on several previous important works with a specific focus on odd-lot trading. Next, we look at how different types of market participants impact price formation, specifically utilizing our dataset to examine more closely the role of high frequency traders in the price discovery process. We further extend this model to analyze briefly the role of trade size and trader type contingent on market capitalization. As a last analysis, we use high frequency return windows of less than 1 minute to analyze how cumulative price changes coming from odd-lot trades vary with the type of market participant. Overall, our results confirm previous analysis which indicates that small trades play an outsized role in the price discovery process at the daily level, and our results hint that high frquency traders engage in stealth trading and play a significant role in the price discovery process. The results revealed in our high frequency analysis indicate that the role of HFTs in driving price movements over very short time horizons are more subtle.

Background, literature review and objectives

Introduction

A key focus in market microstructure is to understand the determinants of price formation in financial markets. One influential postulate on why different types of trades may have different price contributions than what their share in the overall trading process would suggest is the Stealth Trading Hypothesis coined by Barclay and Warner in their 1993 article *Stealth trading and volatility: which trades move prices?*. This theory posits that heterogenous cumulative price impact from trades of different characteristics, such as size, can be traced back to how much information each of these groups of trades inputs into the market over a certain time window. In turn, the choice of certain types of trades might respond to strategic behavior on the part of traders with superior information[1] regarding the true value of the asset. Therefore, to the extent that features of trades associated to traders in possession of

[1]Note that this does not necessarily mean that the agent is engaging in *insider trading*. For instance, some traders may have better proprietary forecasts than others.

profitable private signals can be isolated, it is to be expected that such transactions drive a relatively larger share of cumulative price impact. Barclay and Warner also put forward a couple of competing hypotheses asserting that, if all trades conveyed the same amount of information regardless of their characteristics, the relative contribution to price formation of each category should mirror its share in some measure of aggregate trade over the period. Specifically, these two conjectures, known as the Public Information and Trading Volume hypotheses, predict that price contribution from trades with specific traits should mimic, respectively, their share in the overall number of transactions or total volume of trade in the stock over the interval considered.

Against this backdrop, using a dataset that has formed the basis of prior literature on price formation over the past few years, our thesis focuses on measuring the relative price contribution from trades of varying sizes in which high-frequency traders (HFTs) take either or both sides of the transaction.

Literature Review

The literature on stealth trading dates back to Barclay & Warner (1993), who first designed and tested the manner in which trade size impacts the cumulative price movement of a stock during a certain time interval. Building upon the work done by Kyle (1985), which concluded that informed traders with private information act strategically, the authors devised the weighted price contribution (WPC) procedure to test the cumulative price impact of trades coming from different trade size categories over a certain time interval. By applying the WPC methodology to a dataset of corporate acquisitions in the early 1980's, they find that over 92% of the cumulative price change in the days leading to the announcement of the merger came from medium sized trades. They also find that over 82% of daily cumulative price change in NYSE listed stocks came from transactions in this size category during the 1981 to 1984 period. Together, these results provided evidence that informed traders attempt to camouflage their trading activity by splitting large trades into several pieces (a strategy frequently called *slicing and dicing*) over one or several trading sessions. Barclay and Warner hypothesize that their result is due to the fact that traders in the large trades category are typically uninformed liquidity takers or suppliers that reveal their intentions (thus decreasing the price concession required by market makers due to smaller adverse selection costs), whereas anonymous large block orders would result in large price concessions and information spillovers. In addition, small traders were generally believed to be retail traders and therefore classified as uninformed. Using this logic, Barclay and Warner postulate that medium sized trades are the typical size category used by informed traders. The results obtained by the authors were then underpinned by Chakravarty (2001), who using the same methodology extended the understanding of the importance of the impact of informed traders using medium sized trades by explicitly analyzing the type of trader as well as the trade size. Chakravarty's dataset of NYSE stocks from November 1990 to January 1991 includes the full audit trail of the data, which identifies participants as institutions or individuals. Over the full sample without discriminating by trader type, Chakravarty found that medium sized orders contributed just over 79% of the cumulative price change of a stock, corroborating the results from Barclay and Warner. Moreover, when subsetting the data by trader type, the author found that 102% of the cumulative price change in the

medium sized trade category came from trades initiated by institutions, whereas -2% of the price change came from medium sized trades initiated by individuals. Overall, these results supported the initial results of Barclay and Warner, and provided additional understanding of the investor responsible for the cumulative price changes of stocks. Under the assumption that institutional investors have a higher tendency to acquire profitable private signals, Chakravarty demonstrated that institutions were the primary drivers of share price changes, and that they typically acquired stakes using medium sized trades in order to prevent leaking private information into the market.

However, since the time of Chakravarty's study, numerous developments in equity markets have decreased the costs of breaking up orders into smaller trades. Notably, since the early 1990s there has been a large decline in fees and increasing electronification of markets, both of which serve to decrease the costs of splitting orders. Thus, as the costs of executing smaller orders decreases, one would expect that the stealth trading tactics of informed traders would increasingly gravitate toward smaller trade sizes as they intend to minimize the dissemination of private signals while profiting from them. Market microstructure developments indeed point in this direction, with several more recent studies pointing to a dramatic decline in the size of trades which lead contribution to price discovery in stock markets. Notably, these events have coincided with a marked rise in the participation of HFTs over the last two decades.

O'Hara, Yao & Ye (2014) use a unique dataset to meaningfully update the stealth trading literature for the modern market environment through their analysis of odd-lots (i.e. trades sized 100 shares or less) and their focus on continuous trading daily intervals. First, a history of odd-lot trading on the NYSE is provided, showing that these trades were generally a declining share of total volume in the period from 1950 to 2000, amounting to less than 2% of total volume by 2000. From there, the authors use more recent data from the period between 2008 and 2012 to show steady increases in both the number of trades and the volume coming from odd-lots. As measured at the end of 2011, odd-lots had increased to 6% of the total volume, the largest share seen in US equities since the early 1970's. They also demonstrate that trades initiated by HFTs have a higher probability of being odd-lots than trades initiated by non-HFT market participants. Next, O'Hara, Yao and Ye follow Barclay and Warner's weighted price contribution approach to analyze the price impact of different trade sizes in a dataset which spans 120 representative stocks during the period from 2008 and 2009[2], showing that odd-lot trades contribute a larger share to price formation than what their participation in overall activity would suggest. In all, these results point to price discovery shifting toward small trades over the past twenty years, which the authors argue is tightly linked to increased participation from algorithmic traders.

Brogaard, Hendershott & Riordan (2014) more fully analyze the buying and selling activity of HFT firms. Using a state-space model to decompose price movements of stocks into intraday permanent and temporary components, the authors analyze if HFT trading activity is predictive of future price movements. The authors provide evidence that HFTs tend to trade in the direction of permanent price movements, while trading in the opposite direction of transitory noise. They also show that HFT trading activity predicts price movements at a horizon of 3 to 4 seconds. Their results suggest that HFT firms acquire private information through the trading process itself, and then exploit these signals over very short time

[2]We use this dataset in our empirical analysis and discuss it in further detail in latter sections of the work.

windows.

Objectives

Our project aims to build on the results of the surveyed literature in several respects. First, although there is evidence that HFTs channel a relatively large share of their trading activity through odd-lots (O'Hara, Yao & Ye, 2014), we test how this claim interacts with the relative information content from odd-lots at the daily level using an extension to the conventional WPC procedure. Furthermore, theoretical models suggest that informed traders would only sacrifice price certainty for execution certainty if the value of their private signal were sufficiently close to expiry, usually characterized as intervals shorter than a minute (Kaniel & Liu, 2006). The WPC methodology and our data are well-suited to study a first refinement of this proposition, specifically, if limit orders drive a larger share of price discovery over the trading session according to their size and the type of traders at either side of the transaction. Nevertheless, the WPC approach is not appropiate for assessing these dynamics at higher frequencies. For reasons that will be touched upon in latter sections, frictions related to the trading process which tend to fade for data aggregated at the daily level may be present transitorily and lead to confounding effects in the relevant econometric tests. Therefore, to study how trade price contribution varies by size and type of traders involved at shorter-than-a-minute intervals, we use a fixed effects specification that seeks to capture the relative price contribution of odd-lot trades at higher frequencies. We then use this model to analyze how the relative information content of odd-lots varies when the type of parties involved in the transaction are accounted for. In all, our results should contribute to the understanding of how the information content of trades differs according to their size and the type of parties involved over varying time horizons in the session. Along the way, we take a brief detour from the key focus of our work to shed some light on the debate of whether a systematic relationship exists between odd-lot trades, frequently linked to informed investors, and the market capitalization category of the stock being traded (Roseman, Van Ness & Van Ness, 2018).

Data and empirical strategy

The price contribution literature has focused on building an understanding around which trade sizes convey a disproportionate amount of information and potential explanations for this phenomenom. Throughout this section, the WPC methodology is developed and applied to our data. Next, extensions that allow us to explore how the information content varies depending on both order sizes and type of traders involved, as well as whether any systematic relationship between odd-lot trades and market capitalization appears plausible are estimated. Before reaching that stage, however, a description of the dataset at hand is warranted.

The NASDAQ HFT dataset

The dataset used in this project is usually referred to as the "NASDAQ HFT" database in a number of research articles in the price contribution literature (Hasbrouck & Saar, 2013). The file is built around a random sample of 120 stocks listed on the NYSE and NASDAQ exchanges, and contains transaction-by-transaction data timestamped to the millisecond for all trading dates in 2008 and 2009. Specifically, the sample comprises information on the following eight variables:

- ID: A unique identifier for each of the 120 stocks.
- Date
- Time: Time at which the trade occurred, in milliseconds from midnight.
- Shares: Size of the trade in number of shares.
- Price: Price at which the transaction was executed.
- Direction: Indicates if trade was a purchase (1) or a sale (-1).
- Type: A key feature of the NASDAQ HFT dataset, identifies each transaction by parties involved; HH and NN for a trade between two HFTs or two non-HFTs, respectively, HN for transactions between HFT liquidity takers and non-HFT providers, and NH to denote non-HFT takers trading with HFT makers. The labelling is performed on the basis of NASDAQ's proprietary knowledge on the identity of its customers.
- Capgroup: The sample is grouped into three market capitalization size categories according to 2009 year-end data; large, medium and small, each containing 20 stocks from NASDAQ and 20 from NYSE. The top size batch is built from the largest 40 market capitalizations in the sample, while the mid and small categories include 40 stocks each with sizes around the 1000^{th} and 2000^{th} largest firm values, respectively, in the Russell 3000 index.

Since the focus of our work is on price discovery, we remove observations outside the continuous trading hours between 9:30 and 16:00[3].

Given the granularity of the dataset at hand, traditional summary statistics provide very limited information on its characteristics. Therefore, following Brogaard, Hendershott & Riordan (2014), we calculate equally-weighted daily averages of volume traded by type of parties involved and market capitalization of the underlying stock which may be more informative of features relevant to the models we will estimate in later sections. Results from this procedure are reported in Table 1.

Weighted Price contribution

The WPC literature proposes assessing the relative contribution of trades of certain sizes to price discovery in a given timeframe[4] by focusing on how the ratio of cumulative price

[3] As suggested in Brogaard, Hendershott & Riordan (2014).
[4] The methodology is usually applied to daily or multiday intervals.

change within that subgroup to overall price change in all stocks in the sample compares to the proportion of transactions or volume attributable to that size category over the period.

Consider N trades for stock s in period t, each of which falls in one of J size categories. Then, the price contribution of trades belonging to category j for stock s in window t is given by

$$PC_j^{s,t} = \frac{\sum_{n=1}^{N} \delta_{n,j} r_n^{s,t}}{\sum_{n=1}^{N} r_n^{s,t}} \qquad (1)$$

where $\delta_{n,j}$ is a binary variable that equals 1 if trade n in stock s falls within size group j and 0 otherwise. In turn, $r_n^{s,t}$ is the difference in price between trade n and its immediately preceding transaction for stock s in period t. We remove observations with daily cumulative changes larger than 300% in absolute value to avoid pushing $PC_j^{s,t}$ towards zero due to outliers[5]. Furthermore, to mitigate heteroskedasticity from firms with small daily cumulative price changes, the methodology suggests weighting each stock's price contribution derived from (1) using the factor

$$w^{s,t} = \frac{|\sum_{n=1}^{N} r_n^{s,t}|}{\sum_{s=1}^{S} |\sum_{n=1}^{N} r_n^{s,t}|} \qquad (2)$$

Then, the weighted price contribution of trades in size bin j in timeframe t is given by

$$WPC_j^t = \sum_{s=1}^{S} w^{s,t} PC_j^{s,t} \qquad (3)$$

Finally, assuming the sample is composed of T subperiods of length t, the aggregate weighted price contribution of trades in category j over the full sample is defined as

$$WPC_j = \sum_{t=1}^{T} \frac{WPC_j^t}{T} \qquad (4)$$

Table 2 reports the results from applying the procedure displayed in (3) defining J as six order size categories at daily frequency t and taking averages across dates. Our results from this exercise are consistent with prior literature, pointing to trades sized 100 shares or smaller carrying a disproportionate load of information, while the opposite holds for larger transactions. While displaying WPC results against their share of transactions or volume within a given period may hint on the direction of which trade sizes have relatively larger price impact, there is a series of tests conventionally applied in the price contribution literature which we need to carry out before making statistical statements on the matter. We thus turn to addressing this point by detailing a testing procedure for the WPC methodology and extending the analysis relative to what has been done in previous research in the next section.

Econometric tests on price contribution

As has been established in prior sections, the Stealth Trading Hypothesis provides a logical framework to understand why some trade sizes may be more informative than others.

[5]We follow Barclay & Warner (1993) in this respect. Observations removed constitute less than 4.5% of the sample.

However, this construct stands against a couple of competing theoretical backdrops, both implying that trade characteristics (and in particular, transaction sizes) themselves should be irrelevant for price discovery. As such, the Public Information and Trading Volume hypotheses both posit the testable prediction that the price contribution of order size categories to the overall change of a stock price in a given period should be identical to their share of total trade, where the relevant proportion is calculated with respect to number of transactions in the former conjecture and volume traded in the latter.

The empirical significance of different trade sizes' price contribution can be thus assessed through the following linear specification

$$PC_j^{s,t} = d_{<100}\,\delta_{j<100}^{s,t} + d_{\geq 100}\,\delta_{j\geq 100}^{s,t} + \beta\, prop_j^{s,t} + \varepsilon_j^{s,t} \;\ldots\; (A)$$

Where d is the coefficient on δ, a dummy variable that equals 1 for price contribution from trades within a certain size category and 0 otherwise, β is the coefficient on the proportion of total transactions/volume that a given size j has relative to overall transaction/volume for stock s during period t, and the dependent variable represents the price contribution of category j for stock s on date t as expressed in (1). The way the trade size binary variables are defined in (A) makes for a very intuitive interpretation from the estimation. β, which is estimated from the full trade/volume proportion series tells us how much of daily cumulative price change is explained by the relative share of trades in size bin j to overall trade or volume. In turn, the model intercept, which is split into two by the trade size dummy, tells us whether and in which direction transaction size contributes to cumulative price in the period. Under the null matching either the Public Information or Trading Volume hypotheses, no price contribution difference should arise from trade size category considered and the overall price change should be fully accounted for by the aggregate share in trade of both size categories, that is, $H_0 : \delta_{<100} = \delta_{\geq 100} = 0$ and $\beta = 1$. Therefore, the joint test on the null hypothesis provides a basis for statistical statements on whether trade size is relevant or not for cumulative price changes over the daily session, while the individual coefficient estimates on the binary variables tell us specifically if trade size j contributes disproportionately more (if positive) or less (if negative) than what its share in trade would suggest.

The range of size categories presented in Table 2 was reduced to trades smaller than 100 shares and trades greater than or equal to this magnitude. Conveniently, replicating a specification that has been applied before in the existing literature allows for a "sanity check" on the pre-processing done on the data. In particular, if our first set of results resemble reasonably well those of O'Hara, Yao & Ye (2014), the most recent article applying the WPC methodology to assess price contribution based on this dataset, we can be confident that extensions to the model are being built from a common empirical grounding.

Table 3 displays the results of estimating equation (A) from the NASDAQ HFT database using a Weighted Least Squares procedure with regression weights equal to the series obtained from (2)[6]. The first column shows results for the regression with proportion calculated by number of transactions and the right panel shows the case where this share is computed by volume. As can be concluded from the F-statistic under both specifications, the joint hypothesis that trade sizes are irrelevant for explaining cumulative price changes and

[6]The literature suggests proceeding in this fashion to account for heteroskedasticity, which might be particularly problematic for stocks with small cumulative price changes

that, in turn, these can be exactly traced back to the relative proportion of trades in that group can be rejected at under the 5% significance level. The estimate on the proportion of trades/volume already hinted in this direction, being individually different from 1 at under 5% significance. Notably, the coefficient on the odd-lot binary regressor is significant and positive in both specifications, providing statistical evidence to the claim that these trades contribute disproportionately to cumulative price changes at the daily level. Both of these results are consistent with the Stealth Trading Hypothesis. In contrast, the price contribution from trades sized 100 shares or larger is on average smaller than their daily proportion of transactions/volume, although this is only significant at the 5% level in the volume share specification. Our results resemble those in O'hara, Yao & Ye (2014) in magnitude, sign and significance [7].

We perform a first refinement to the conventinal WPC approach that seeks to compute relative price informativeness by discriminating trades not only by their size (i.e. odd-lots vs non odd-lots) but also by the type of initiating trader. If we take a step back and look into the way the relative price contribution in model (A) was calculated in equation (1), it appears that j need not only accomodate for trade size categories, but potentially could account for more detailed characteristics to allow us to group the price contribution from a more refined set of trades and then test whether the stealth trading hypothesis holds. With this in mind, we ran the procedure used to calculate price contribution as in (1) and weights as in (2) by defining the price contribution from four distinct groups of trades: high-frequency initiated/odd-lot, non high-frequency initiated/odd-lot, high-frequency initiated/non odd-lot, and non high-frequency initiated/non odd-lot. The model we need to regress to test whether each of these categories contributes disprportionately to price discovery over the daily session is thus given by

$$PC_j^{s,t} = \sum_{h=1}^{2}\sum_{i=1}^{2} d_{h,i}\delta_{h,i} + \beta\, prop_j^{s,t} + \varepsilon_j^{s,t} \;\ldots\; (B)$$

Where the subscript h on the δ dummy variable represents the trade size considered (odd-lot/non odd-lot) and i characterizes the type of initiating trader (HFT/non HFT). Table 4, which maintains the overall structure of Table 3, presents our results from estimating model (B). Of note, the null hypothesis consistent with the trading volume or public information hypothesis is now given by $H_0 : \delta_{<100,HF} = \delta_{<100,nHF} = \delta_{\geq 100,HF} = \delta_{\geq 100,nHF} = 0$ and $\beta = 1$. In addition to an F-test on H_0, we now include joint tests for $\delta_{<100,HF}+\delta_{<100,nHF} = \delta_{\geq 100,HF}+\delta_{\geq 100,nHF}$ and $\delta_{<100,HF} + \delta_{\geq 100,HF} = \delta_{<100,nHF} + \delta_{\geq 100,nHF}$ to assess whether significant price contributions differences arise between odd and non odd-lot sized trades once the type of initiating party has been considered, and between HFT and non HFT initiated transactions once trade sizes have been accounted for.

We see in these results that accounting for the initiating trader provides us with a richer interpretation of the price discovery process. As in part (A), we see that the F-test for trade size is significant both when viewing the data by number of transactions as well as by volume. In addition, we reject the null hypothesis that the coefficients of all category groups are equal to zero, thus rejecting the public information and trading volume hypotheses.

[7] Small differences may arise, for instance, if the non-continuous trading hours or outsized returns excluded from the analysis in that piece differed slightly from our own.

When zooming in to look at individual coefficients, we find a more nuanced picture. Using the transaction specification, we see that every category other than odd-lot/HFT-initiated trades have negative coefficients, indicating that these categories contribute less to price discovery than their proportion of the number of transactions. However, our coefficient for odd-lot/HFT-initiated trades is not significant. When looking at the data by share of total volume, all four of the coefficients are statistically significant, and the results are in line with our expectations in that odd-lot trades initiated by HFT traders contribute 13 percent more than their share of the volume to price discovery, the most of any of our four groups. This seems to provide support for the idea that HFTs engage in stealth trading. However, we do note that in both specifications we are unable to reject the null hypothesis that the coefficients on HFT initiated trades are statistically different from those of non-HFT initiated trades. This is likely due to the fact that as shown in part (A), odd-lot trades have a very strong impact on price discovery, and those results continue to hold here as even non-HFT initated odd-lot trades contribute more than their share to price discovery. Given that the usage of odd-lots is also likely by informed traders who are not HFTs (for example, institutions who have the ability to trade algorithmically), we do not see this as a significant challenge to the thesis that HFTs engage in stealth trading, but instead find it more plausible that other informed institutional investors make up a signficant portion of the non-HFT initated odd-lot trades. Thus, an avenue for future research could include using data with the ability to separate trader type further into HFT, non-HFT institutions, and non-institutional clients in order to better identify informed traders.

We then study results drawn from equation (B) differ when we restrict our sample by market capitalization category. Although thoroughly touching upon the mechanisms that may lie behind distinct price contribution from trades in stocks with different market capitalization is beyond the scope of our work, we can get an idea as to whether average differences in the relative price contribution from different trade size - initiating trader combinations may arise when the market capitalization of the traded stock is considered. Results from this exercise are reported in Tables 5, 6 and 7.

In general, our results when running the regressions segmented by market cap are similar to our findings in part (B). In all specifications, we reject the null hypothesis that our dummy variable coefficients are equal to zero, and we also reject the null hypthesis that the coefficients for odd-lot trades are not statistically different from non-odd lot trades. In addition, just as in part (B), we see that our coefficients for the specification using share of volume are all statistically significant, whereas the specification using share of transactions can at times display coefficients that are not significant depending on the model specification. More interesting, however, we see a general trend as market cap decreases, the contribution to price discovery coming from odd-lot trades shows an increasing trend that is especially evident in our model specification using the share of volume. We postulate that stealth trading may be more visible in small-cap stocks because their lower liquidity creates larger potential price impact from inventory risk considerations in these stocks, leading informed investors in both the HFT and non-HFT category to use odd-lot trades more frequently. Lastly, we also note that small caps are the only category where non-HFTs contribute a larger share of price discovery than HFTs in the odd-lot category, which seems to indicate that HFTs trade less actively in than non-HFT informed investors due to liquidity concerns and adverse selection risk in these categories.

Now we turn to studying whether price contribution from different trade sizes varies depending on the type of traders involved in the transaction at intervals lasting only a few seconds. Importantly, however, the WPC approach is not properly suited for this purpose with the dataset at hand, which contains transaction-by-transaction prices. Intuitively, the probability that at any shorter-than-a-minute interval the best bid or ask price of the limit order book is wiped out thus shifting the mid-quote from its previous level is naturally smaller than the probability that the price moves higher or lower due to aggregate trading dynamics over a full trading session, particularly when focusing in odd-lot trades. The consequence is that transaction price data in very short intervals often exhibit a *bounce* between the best prices at either side of the book as buyers and sellers trade without shifting the overall spread up or down (Lerner, 2010). This dynamic would therefore introduce excessive noise into model (A), decreasing its statistical power and potentially leading to confounding effects if applied to our data. In fact, some extensions to the WPC methodology have focused on developing alternatives that allow for identification of relative price contribution under procedures that are robust to these features. For instance, using a dataset from the Spanish stock exchange that includes high-frequency quote level data in addition to transaction data, Abad and Pascual (2014) find that using the midpoint between the best bid and best offer in place of using transaction prices results in a large upward adjustment to the contribution of odd-lot trades to price discovery. Against this backdrop, they propose modifying the WPC when using high-frequency data in such a way that a frictionless estimate that strips out the bounce effect is obtained by using the midpoint price. Unfortunately, given that our dataset contains only transaction data and not quote level data, we are unable to validate the approach in the context of our project. Thus, we pursue a different approach when using high frequency data detailed below.

To assess differences in price contribution from different trade size-trader type combinations at higher frequencies we first compute forward looking log-returns for 1, 5, 10, and 30 second windows by stock from our dataset, multiplying each of these returns by the direction of the trade they followed (as stated in prior sections, 1 for purchase and -1 for sale orders). This implies that, if prices move in the direction of the accumulated trade in the window considered (i.e. prices increase after buy and decrease after sale the lagged order) the resulting computation will be positive, otherwise, unless cumulative return over the interval equals exaclty zero, it will be negative. In other words, positive returns mean transactions contribute to price discovery over the interval. We then compute the average log return for each of the windows considered at the stock-date level, which will form the basis of the econometric tests discussed in latter paragraphs.

To perform formal tests on if and how price contribution varies between trade sizes at high frequencies, our specification of interest is given by the following Fixed Effects model

$$r_{st} = k + d_1\delta_{<100,HFT} + d_2\delta_{\geq 100,HFT} + d_3\delta_{\geq 100,nHFT} + \sum_{s=1}^{120}\alpha_s S_s + \sum_{w=1}^{52}\lambda_w W_w + \upsilon_{it} \ ... \ (C)$$

where r_{st} is the daily average 1-, 5-, 10-, or 30-second forward looking log return for stock s on date t, $S_{s=1}^{120}$ and $W_{w=1}^{52}$ are binary variables to allow for stock and week fixed effects and d_h for h in {1, 2, 3} are the coefficients on dummy variables which identify average returns over the interval for odd-lot trades initiated by HFTs, non odd-lot trades initiated by HFTs

and non odd-lot trades initiated by non HFTs. Note that although model (C) resembles the structure of the specification used for daily tests, to avoid multicollinearity issues we omit the explicit inclusion of the odd-lot initiated by non HFT category, which is in turn represented by intercept k, thus our coefficients will be interpreted relative to that baseline. Table 8 reports our results from running model (C). As can be seen, none of our coefficients our statistically significant for the data when grouped as a whole. We provide a potential explanation in the next paragraph.

To mirror the procedure followed at the daily level, we run equation (C) restricting our sample by market capitalization category. Results from these estimations are reported in Tables 9, 10 and 11. Interestingly, we do achieve statistically significant results for both the large and mid-cap specifications at high frquencies. We also note that HFT initiated odd-lot trades play a relatively larger role in high frequency price discovery for large cap stocks than for mid-cap and small-caps, as indicated by the larger relative coefficients for the HFT initiated odd-lot trades in the large-cap specification. This is consistent with our understanding that HFTs preferentially trade in highly liquid securities, and thus they play a more dominant role in the price discovery process for large-cap stocks. For the small-cap specification, levels of significance of the coefficients vary across the different groups and time horizon. We suspect that these results may be impacted by the lack of continous trading at the small cap level. We also note that for small-cap stocks, the sign of the coefficient for HFT initiated odd-lot trades changes sign and is statistically significant. We believe this potentially points to HFTs being subject to higher adverse selection costs from informed non-HFTs using limit orders to accumulate positions in small-cap stocks. In turn, this change of sign potentially confounds our results when aggregated across market caps, leading to a lack of significance. We conclude that the price discovery process for small-cap stocks is fundamentally different than for larger market capitalizations, as it does not appear to rely on HFTs and is likely driven by informed investors who acquire fundamental information and operate with longer trading horizons.

Conclusions

Replicating the approach pioneered by Barclay and Warner, we first used the WPC approach to investigate if that small trades, and in particular odd-lots, contribute an outsized proportion to the process of price discovery in US equity markets. As shown in Table 2, using our 6 discrete trade size buckets, we calculate that odd-lots contribute to 37% of the total price discovery in our sample, which amounts to over 4x more than their share of the overall volume. In addition, round lots of exactly 100 shares contribute another 47.9% to price discovery, pointing to the vast majority of price discovery comes from trades less than or equal to 100 shares. Next, we collapse our trade size buckets into two groups denoted as less than 100 shares or greater than or equal to 100 shares, and we use dummy variables to analyze the statistical significance of our results price contribution differences. Using a joint test under the null hypothesis that the coefficients on our trade size dummies are equal to zero, we are able to reject the null hypothesis and determine that whether or not a trade is an odd-lot or not is a statistically significant determinant of price contribution. Finally, we briefly analyze the contribution of odd-lot trades segmented by market capitalization but find no additional significant information.

Next, we analyze whether trader type adds any significant information to our trade size model by discriminating by both trade size and whether the initiating trader is HFT or non-HFT. We find that, in line with our priors, our share of volume specifications show evidence of HFTs engaging in stealth trading, as HFTs initiating odd-lot trades contribute more to price discovery than their share of volume would suggest. We also do not that non-HFTs using odd-lot trades also contribute more to price discovery than their share of volume would suggest, which we believe is likely to be driven by non-HFT investors who are informed also gravitating toward small trades. Moreover, by running this same analysis conditional on market capitalization, we show that these results are consistent across all market caps, with the impact of odd-lots becoming stronger and the impact of HFT becoming slightly weaker as cap size decreases. Overall, our results lend credence to the hypthothesis that HFTs engage in stealth trading, and that their activity drives a greater proportion of price discovery than their share of volume alone would suggest.

Finally, due to challenges in uncovering HFT impact on price discovery with high frequency data, we shift to using a fixed effects model to analyze the role of trader type in post-trade price movements using short horizon windows of one second, five seconds, 10 seconds, and 30 seconds. We again segment the data by our binary specification of odd-lot vs larger than odd-lot, and use our four possible trader type interactions. Our results indicate profoundly different dynamics for short horizon price discovery dependent on market capitalization. While we find a statistically significant evidence of HFT impact on price discovery over short horizons in large and mid-cap stocks, our results for small-cap stocks diverge from the other categories and confound the signficance of our results when the data is aggregated. Overall, while our results continue to indicate that HFTs play a signficant role in short term price discovery among large-cap and mid-cap stocks, we see confounding effects coming from our analysis of the short-horizon price dynamics of small-cap stocks. The results point to HFTs playing a more important role in highly liquid securities, in line with our priors that HFTs preferentially trade in liquid securities, but this remains an area for further research.

There are a number of potential implications for follow on research. In particular, some observers have noted that the predilection for odd-lot trades among modern informed traders may be driven by a regulatory rule that did not require odd-lot trades to be reported in the consolidated tape. This rule was changed in December 2013, so that odd-lots are now reported in the consolidated tape, and so replicating this analysis on a dataset after December 2013 would help triangulate whether this regulatory rule was a key driver of the use of odd-lots among informed traders. In addition, the difficulty in obtaining statistical significance for different trader types may be an artefact of our data, which includes only transaction-by-transaction prices. As a result, the statistical power of our analysis is decreased, most especially when looking at small trade sizes at high frequencies because small trades are probabilistically more likely to exhibit bid-ask bounce noise over short horizons. Obtaining quote level data would go a long way toward stripping out this noise, and would provide a better specification to analyze how HFTs, who tend to operate in small trade sizes over short horizons, contribute to price discovery. After all, if HFTs are skilled at stealth trading, we should expect their signature to be well hidden among the noise of financial markets.

Overall, our results corroborate previous results that indicate that price discovery has become increasingly driven by small trades in recent years, a phenomenon that has coincided with a larger presence of algorithmic traders. In addition, we show that HFTs using small

trades play a significant role in the process of price formation, especially when considering their presence among more liquid and larger capitalization securities. This is in line with the theoretical expectations of the stealth trading hypothesis, which states that informed investors attempt to hide their trades so as to prevent leaking information into the market while accumulating positions.

References

1. A.S. Kyle *Continous auctions and insider trading*. Econometrica, 1985.

2. M.J. Barclay and J.B. Warner. *Stealth trading and volatility: Which trades move prices?*. Journal of Financial Economics, 1993.

3. S. Chakravarty. *Stealth-trading: Which traders' trades move stock prices*. Journal of Financial Economics, 2001.

4. R. Kaniel and H. Liu. *So What Orders Do Informed Traders Use?*. The Journal of Business, 2006.

5. P. Lerner. *Theoretical analysis of the bid-ask bounce and Related Phenomena*. Aestimatio, 2010.

6. J. Hasbrouck and G. Saar. *Low-Latency Trading*. SSRN, 2013.

7. D. Abad and R. Pascual. *The Friction-Free Weighted Price Contribution*. SSRN, 2014.

8. M. O'hara, C. Yao and M. Ye. *What's Not There: Odd Lots and Market Data*. The Journal of Finance, 2014.

9. J. Brogaard, T. Hendershott and R. Riordan. *High-Frequency Trading and Price Discovery*. The Review of Financial Studies, 2014.

10. B.S. Roseman, B.F. Van Ness and R.A. Van Ness. *Odd-lot trading in US equities*. The Quarterly Review of Economics and Finance, 2018.

APPENDIX

Table 1: Average daily trading volume (million USD)

Type		nHFT^D	HFT^D	nHFT^S	HFT^S
Cap size					
	Small	0.74	0.25	0.88	0.11
	Mid	4.11	2.36	5.22	1.25
	Large	104.50	77.68	105.10	77.08

Superscripts on Type denote demand or supply side of transactions.

Table 2: Weighted price contribution by trade size

Size Category	WPC	Share of Trades	Share of Volume
$[1, 100)$	0.370	0.229	0.082
100	0.479	0.602	0.523
$(100, 200]$	0.071	0.095	0.139
$(200, 500]$	0.053	0.056	0.135
$(500, 5000]$	0.026	0.022	0.130
> 5000	0.000	0.000	0.036

Table 3: Estimation of model (A)

		Transactions	Volume
Trade size			
	< 100	0.120***	0.193***
		(7.56)	(14.15)
	≥ 100	-0.065*	-0.954***
		(-1.82)	(-11.29)
Proportion		0.945***	1.761***
		(20.31)	(18.87)
Adjusted R^2		.052	.051
F-test on H_0		74.9	510.7

(*), (**) and (***) denote significance at the 10, 5 and 1% levels. t-stats reported in parentheses.

Table 4: Estimation of model (B)

	Transactions	Volume
$\delta_{j<100,HFT}$	0.000	0.134***
	(0.03)	(21.12)
$\delta_{j<100,nHFT}$	-0.024***	0.080***
	(-4.06)	(14.85)
$\delta_{j\geq100,HFT}$	-0.227***	-1.157***
	(-17.90)	(-40.01)
$\delta_{j\geq100,nHFT}$	-0.202***	-1.100***
	(-15.13)	(-37.29)
Proportion	1.226***	2.020***
	(75.30)	(63.27)
Adjusted R^2	.206	.200
F-test on $\delta_{j<100} = \delta_{j\geq100}$	482.7	2243.7
F-test on $\delta_{HFT} = \delta_{nHFT}$	0.0	0.0
F-test on H_0	117.6	1632.4

(***) denotes significance at the 5% levels. t-stats reported in parentheses.

Table 5: Estimation of model (B) for large caps

	Transactions	Volume
$\delta_{j<100,HFT}$	-0.039***	0.086***
	(-2.85)	(7.28)
$\delta_{j<100,nHFT}$	-0.065***	0.029***
	(-5.70)	(2.803)
$\delta_{j\geq100,HFT}$	-0.342***	-1.632***
	(-11.89)	(-22.79)
$\delta_{j\geq100,nHFT}$	-0.315***	-1.575***
	(-10.31)	(-21.58)
Proportion	1.380***	2.546***
	(36.97)	(32.37)
Adjusted R^2	.198	.282
F-test on $\delta_{j<100} = \delta_{j\geq100}$	169.7	643.7
F-test on $\delta_{HFT} = \delta_{nHFT}$	0.0	0.0
F-test on H_0	42.1	479.5

(***) denotes significance at the 5% level. t-stats reported in parentheses.

Table 6: Estimation of model (B) for mid caps

	Transactions	Volume
$\delta_{j<100,HFT}$	0.009	0.114***
	(0.80)	(10.65)
$\delta_{j<100,nHFT}$	-0.021***	0.044***
	(-2.40)	(5.26)
$\delta_{j\geq100,HFT}$	-0.282***	-1.600***
	(-15.56)	(-33.45)
$\delta_{j\geq100,nHFT}$	-0.252***	-1.530***
	(-12.62)	(-30.74)
Proportion	1.273***	2.486***
	(53.98)	(46.40)
Adjusted R^2	.241	.233
F-test on $\delta_{j<100} = \delta_{j\geq100}$	369.2	1430.4
F-test on $\delta_{HFT} = \delta_{nHFT}$	0.0	0.0
F-test on H_0	89.2	846.3

(***) denotes significance at the 5% level. t-stats reported in parentheses.

Table 7: Estimation of model (B) for small caps

	Transactions	Volume
$\delta_{j<100,HFT}$	0.023**	0.158***
	(1.79)	(13.31)
$\delta_{j<100,nHFT}$	0.054***	0.191***
	(5.40)	(21.28)
$\delta_{j\geq100,HFT}$	-0.025	-0.542***
	(-1.43)	(-17.09)
$\delta_{j\geq100,nHFT}$	-0.056***	-0.569***
	(-3.41)	(-17.90)
Proportion	1.002***	1.378***
	(47.05)	(38.97)
Adjusted R^2	.190	.182
F-test on $\delta_{j<100} = \delta_{j\geq100}$	47.5	740.1
F-test on $\delta_{HFT} = \delta_{nHFT}$	0.0	0.1
F-test on H_0	22.7	584.2

(***) and (**) denote significance at the 5 and 10% levels. t-stats reported in parentheses.

Table 8: Estimation of model (C)

	Window			
	1 sec	5 sec	10 sec	30 sec
Intercept	0.010	0.020	0.021	0.023
	(0.68)	(0.43)	(0.46)	(0.50)
$\delta_{<100,HF}$	-0.003	-0.008	-0.005	0.002
	(-1.15)	(-0.86)	(-0.49)	(0.26)
$\delta_{\geq 100,HF}$	-0.002	-0.004	-0.001	0.010
	(0.57)	(-0.43)	(-0.10)	(1.09)
$\delta_{\geq 100,nHF}$	-0.001	-0.004	-0.001	0.005
	(-0.36)	(-0.40)	(-0.10)	(0.53)

(*) Denotes significance at the 5% level. t-stats in parentheses.

Table 9: Estimation of model (C) for large caps

	Window			
	1 sec	5 sec	10 sec	30 sec
Intercept	0.005***	0.010***	0.013***	0.017***
	(22.44)	(30.63)	(32.10)	(31.63)
$\delta_{<100,HF}$	0.005***	0.106***	0.013***	0.015***
	(97.43)	(122.10)	(123.94)	(103.56)
$\delta_{\geq 100,HF}$	0.007***	0.014***	0.018***	0.021***
	(137.34)	(166.90)	(169.40)	(144.40)
$\delta_{\geq 100,nHF}$	0.003***	0.005***	0.006***	0.008***
	(54.90)	(61.16)	(61.05)	(53.17)

(***) Denotes significance at the 5% level. t-stats in parentheses.

Table 10: Estimation of model (C) for mid caps

	Window			
	1 sec	5 sec	10 sec	30 sec
Intercept	0.011***	0.018***	0.019***	0.017***
	(12.60)	(13.64)	(10.50)	(7.17)
$\delta_{<100,HF}$	0.002***	0.008***	0.012***	0.022***
	(8.56)	(21.71)	(25.76)	(36.63)
$\delta_{\geq 100,HF}$	0.004***	0.013***	0.019***	0.034***
	(16.87)	(35.95)	(40.04)	(55.70)
$\delta_{\geq 100,nHF}$	0.002***	0.005***	0.008***	0.014***
	(8.27)	(15.67)	(16.35)	(22.65)

(*) Denotes significance at the 5% level. t-stats in parentheses.

Table 11: Estimation of model (C) for small caps

	Window			
	1 sec	5 sec	10 sec	30 sec
Intercept	0.021	0.039	0.033	0.043
	(0.61)	(0.37)	(0.30)	(0.40)
$\delta_{<100,HF}$	-0.018***	-0.043	-0.040	-0.030
	(-1.97)	(-1.52)	(-1.40)	(-1.07)
$\delta_{\geq 100,HF}$	-0.017**	-0.039	-0.040	-0.025
	(-1.85)	(-1.41)	(-1.43)	(-0.87)
$\delta_{\geq 100,nHF}$	-0.008	-0.022	-0.017	-0.007
	(-0.91)	(-0.79)	(-0.61)	(-0.24)

(***) and (**) denote significance at the 5 and 10% levels. t-stats in parentheses.