The Market Inside the Market: Odd-lot Quotes[1]

Robert P. Bartlett
University of California, Berkeley School of Law
rbartlett@berkeley.edu

Justin McCrary
Columbia University Law School; National Bureau of Economic Research (NBER)
jrm54@columbia.edu

Maureen O'Hara
Johnson College of Business, Cornell University
mo19@cornell.edu

November 29, 2021

Abstract

We show how current market practices relating to odd lot quotes result in a large "inside" market where for many stocks better prices routinely exist relative to the National Best Bid or Offer (NBBO). We provide strong evidence that being able to see these odd lot quotes provides valuable information to traders with access to proprietary data feeds. We develop a XGBoost machine learning prediction algorithm that uses odd lot data to predict future prices, and demonstrate a simple and profitable trading strategy using odd lot data. We show that the SEC's new approach of changing the definition of a round lot reduces, but does not ameliorate, the high incidence of superior odd lot quotes within the NBBO.

JEL classification: G14, G21, G23, G24

Keywords: Odd-lot quotes, inside market, NBBO, round lots, Reg NMS II, machine learning.

[1]We thank May Street for providing the data used in this paper. We also thank Andrew Baker, David Easley, Phil Mackintosh, James Hicks, Frank Partnoy, Bart Yueshin, and seminar participants at the International Virtual Finance seminar for helpful comments.

The National Best Bid or Offer (NBBO) establishes the "best" price to buy (the offer) or sell (the bid) every U.S. listed equity security at each point in time in the trading day – except that it doesn't.[2] The NBBO only reports quotes based on "round lots" (100 share orders), so quotes arising from smaller orders (called odd lots) are not generally included. In 1975 when the system was implemented, this omission was largely irrelevant, but that is no longer the case. The rise of algorithmic trading which chops orders into smaller pieces, the advent of very high-priced stocks (a round lot of AMZN costs over $320,000; GOOG is approximately $275,000), and the renaissance of retail traders have all made odd lot quotes far more numerous. As a result, there is now a "market inside the market" where better quotes routinely exist compared to those disseminated as the NBBO.[3]

In this paper we investigate the odd-lot quotes that populate this inside market. For some stocks we show that odd-lot quotes play a minor role, but this is not the case for others – in our sample period, for example, 60% of the time a better odd lot bid was available for Amazon, 54% of the time for Google, 53% of the time for Tesla, and over 25% of the time for Facebook. How important are these odd-lot quotes in general and what do they imply for the efficiency and execution quality of equity trading? Equally important are the implications for information asymmetries between market participants. Odd-lot quotes are not visible to those watching the "tape" but they are visible to those purchasing the proprietary data feeds from each of the exchanges. How much of an advantage does knowing this information convey relative to knowing only the national "pretty good" bid or offer? Finally, we address a policy question: will changing the size of what constitutes a round lot as proposed by the SEC remove the issues posed by the inside market?

To address these questions, we use a new data set that contains virtually every message, quote and trade on the 16 U.S. stock exchanges during our January-March 2021 sample period. These data, provided to us by May Street, are the same data the Securities and Exchange

[2] The National Market System (NMS) was implemented in 1975 to enhance transparency, fair competition and best execution in U.S. equity markets. A cornerstone of this system, the National Best Bid or Offer (NBBO) is formed from the aggregation of the best quotes to buy (the bid) and sell (the ask) each security from each of the 16 stock exchanges at each point in time. This information is collected by two Securities Information Processors (the "SIPs") and then disseminated on the "tape", giving traders contemporaneous information on quotes and spreads in each security. More information on the National Market System and its evolution can be found at https://www.finra.org/rules-guidance/guidance/national-market-system-plans.

[3] There are other causes of better prices inside the spread such as hidden orders or mid-point orders but these are not observable to other market participants.

Commission (SEC) uses for its high frequency market surveillance platform called "MIDAS." We note at the outset that the scale of these data poses major challenges for doing empirical research. Several of our tables, for example, involve 32 billion data points, an astonishing number given the length of our sample period. But such is the nature of today's high frequency markets, and these data give us the rare ability to see the entire market, both inside and out. Given such 'big data', we use machine learning techniques to evaluate the information content of odd-lot information, allowing us to assess the value of knowing the rich montage of odd-lot information available (or not available) to some market participants.

Our research provides a number of new results on this inside market. We divide our sample stocks into 5 price-based buckets, with stocks in group 1 having prices $20.00 or below, up to group 5 stocks with prices $250.00 and over. We find that traders submit odd-lot quotes for all stock groups and that the rate of odd lot orders ranges from 5.6% of all submitted orders for less than 500 shares in group 1 to 46.9% of all such orders in group 5. The incidence of odd lot orders differs markedly across exchanges, with the taker-maker platform Nasdaq BX having far more odd-lot orders (on a percentage basis) than any other venue. The odd-lot spread is smaller (i.e. better) than the spread of the NBBO for all stock groups, and this advantage increases with stock price, reaching on average 26% better for group 5 stocks. Perhaps most intriguing are our results on the incidence of superior odd lot quotes relative to the NBBO. While for the lower price stocks this is only the case an average 5.1% of the time, this incidence reaches almost 30% for group 4 stocks and it averages 42% for our highest price group. We use regression analysis to predict the incidence of such superior pricing, with our results highlighting the important interactions between variables such as price, NBBO spreads, volatility, and volume.

But does any of this really matter? Odd-lot quotes are for small amounts so even in large numbers they could still have little impact on the market as a whole. We address this issue in three ways. First, we focus on depth and ask: can any meaningful trade size be executed inside the spread? We aggregate all odd-lot orders that are inside the NBBO to get a measure of odd-lot depth relative to depth at the NBBO at each point in time. Our results suggest that, at first glance, for many stocks the answer is no – for the three lowest priced groups the average inside depth is less than 4% of the depth displayed at the NBBO. But this result is driven by the fact that if spreads are at one tick then there can be no odd-lot quotes. Conditioning on superior odd lot quotes, the picture is very different – with inside depth across our stock groups averaging almost 15% of the NBBO depth. Moreover, our regressions show how this odd-lot depth is affected by other market variables, providing a road map for smart trading algorithms to exploit this inside depth. From a trade execution perspective, our results show that odd-lot orders can play an important role, particularly for higher priced stocks.

Second, we ask about the information content of these odd-lot orders – how well do odd lot metrics predict whether a stock's price will be higher in one minute? This issue is of particular importance given that only some traders purchase proprietary data feeds for all exchanges, while others rely on information from the two Securities Information Processors (the "SIPs") that are responsible for collecting and disseminating exchanges' best bids and offers throughout the trading day. Relative to subscribing to all exchanges' proprietary data feeds, acquiring data from the SIPs is simple and cheap, but the SIP data only include round lot quotations from the exchanges.

To explore the informational advantage of subscribing to proprietary data feeds, we focus on group 1 stocks (the lowest price group) and group 5 stocks (stock prices at or above $250.00). We postulate a simple "SIP model" based on data from the SIP feed and a rival "OL model" that adds odd lot quotation data from the proprietary feeds. We use regression analysis to investigate whether quotation data can explain the log difference between the volume weighted average price between minute *t* and minute *t+1* ($VWAP_{t+1}$) and the SIP midpoint observed at minute *t*. We find strong evidence that odd lot quotes have information distinct from that in the SIP data. While most SIP variables are not significant, OL spreads and OL imbalance show strong positive associations with VWAP returns. Moreover, the OL model's R-squared is five times that of the SIP model, although the overall fit of both models remains low.

Given the massive size of our data, we turn to supervised machine learning—in particular, XGBoost, a gradient boosting decision tree algorithm—to investigate further how well SIP and OL data predict whether the label VWAP at time *t+1* is greater than the SIP midpoint at time *t* given features at time *t*. We find strong predictability for both the SIP and OL models, but particularly so for the odd lot model – with accuracy, precision, and recall metrics of 55%, 56% and 58% respectively. We use the Shapley Additive exPlanations (SHAP) method to explore why the odd lot data provide such superior performance and to establish the differential predictive ability of particular data features.

Third, based on our machine learning results, we design a simple trading strategy to ask whether the predictions of the XGBoost algorithm can provide profitable trading strategies. The short answer is yes. When the model predicts with at least 50% probability a positive label (i.e., the stock price going up), we calculate the round-trip trading profit (net of the effective half spread calculated over all stock-days in our test data) for a purchase at the beginning of a stock minute and a sale at $VWAP_{t+1}$. Two points here are particularly salient. One is that trading costs overwhelm the profit potential of this simple strategy for both the SIP and OL models. But raising the probability threshold of a positive label to 65% shows a different story: the SIP model provides negligible returns while the OL model produces average daily returns of 1.42% on reasonable volume. Indeed, we find positive returns to the OL model at thresholds as high as 90%. We interpret these results as demonstrating that odd lot data is informative and valuable.

Overall, our results underscore that the "market inside the market" is important enough that analyses using information only on round lot quotes has the potential to lead to erroneous conclusions, especially for high-priced stocks. The issues surrounding odd-lots, as well as broader issues connected with what data in general should be visible to the market, have attracted extensive industry and regulator attention.[4] Many of these issues are beyond our focus here, but the questions of whether and how odd-lot quotes should be reported as "core", publicly-available data are certainly germane. As part of its new rules ("Reg NMS II"), the SEC recently adopted changes to the definition of what constitutes a round lot depending upon a stock's price.[5] For stocks with prices at or below $250.00, a round lot stays at 100 shares; for stock prices greater than $250.00 and up to $1000.00, a round lot would be 40 shares; for stocks above $1000.00 up to 10,000.00, 10 shares; and for stocks above that level, 1 share. Under this definition, and when the rules are actually implemented, the number of odd-lot quotes will be reduced, but how well will this fix the problems posed by odd-lot quotes?

We answer this question in a counter-factual manner by looking at how this proposed round lot definition would have changed odd-lot incidence and depth in our January-March 2021 sample period. Our analysis shows that for our sample stocks in the 40-share category, the

[4] See, for example, The Reg NMS II Debate Begins, Traders Magazine, May 20 2020; or Nasdaq Comment Letter on Proposed Rule on Market Data Infrastructure (Release No. 34-88216; File No. S7-03-20, RIN 3235-AM61), May 26, 2020 available at https://www.nasdaq.com/docs/2020/05/27/Reg-NMS-II-Comment-Letter.pdf

[5] For description of these proposed changes, see SEC Press Release 2020-211, Dec. 9, 2020, "SEC Adopts Rules to Modernize Key Market Infrastructure Responsible for Collecting, Consolidating, and Disseminating Market Data".

incidence of superior odd lot quotes (relative to the NBBO using the new round lot definitions) would fall by approximately 4.8%. But even with this new definition, there is a better odd lot bid 35.1% of the time and a better odd lot ask 33.2% of the time. For sample stocks falling into the 10-share category, the improvement is greater, with the incidence falling approximately 22% of the time. Here, again, though, the inside market remains large: the odd lot bid (ask) is better 38.8% (37%) of the time. We discuss some of the many interesting dimensions of this policy, and the odd lot problem in general, further in the conclusion.

Our paper joins a small literature looking at odd lot equity trading and a much larger literature considering equity market structure. To our knowledge, our paper is the first academic research to examine odd-lot quoting and the information content and importance of the inside market.[6] Other research examining odd lot trading are O'Hara, Yao, and Ye [2014] who establish the importance of odd lot trades and their exclusion from the consolidated tape and TAQ data, and Bartlett [2021] who examines whether odd lot trades receive best execution. Market structure research addresses a wide range of topics including tick size (see, for example, Buti et al [2019]; O'Hara, Saar, and Zhou [2020]; Bartlett and McCrary [2020]; Chao, Yao, and Ye [2019]); latency (Brogaard et al [2015]; Bartlett and McCrary [2019]; Khapko and Zoican [2021]), and routing decisions (Battalio, Corwin, and Jennings [2016]; Cimon [2021]). Our work is most closely related to the literature examining proprietary data feeds (Cespa and Foucault [2017]; Easley, O'Hara, and Yang [2018]). As we show here, traders with access to proprietary data on odd lots have a significant advantage over traders relying on the SIP data, raising important questions of information asymmetry connected to the inside market. These issues are at the heart of Reg NMS II, which sets out wide-ranging changes to the collection and dissemination of market data. Our analysis suggests that their proposed definition of odd lots may reduce, but not substantially improve the problems posed by the inside market.

Our research also contributes to the growing literature using big data for finance research.[7] Chinco, Clarke-Joseph and Ye (2018) apply LASSO techniques to make 1-minute ahead equity return forecasts; Isil, Stern, Tan and Weisbach (2021) consider how well machine learning algorithms assess director performance; Philip (2019) uses reinforcement learning to estimate the

[6] We note, however, that these issues have been of interest in the industry, with research by Mackintosh [2019] being among the first to raise concerns about odd lot quoting and the Consolidated Tape Association [2019] proposing possible changes to dissemination of odd quotes, see https://www.ctaplan.com/oddlots
[7] For an overview of big data issues in finance see Goldstein, Spatt, and Ye [2021].

permanent price impact of a trade; Rossi and Utkus (2020) use machine learning to investigate winners and losers in robo investing; Gu, Kelly and Xiu (2020) apply multiple machine learning regression algorithms in asset pricing; and Easley, Lopez de Prado, O'Hara and Zhang [2021] apply random forest to investigate the predictive power of microstructure variables. Our work contributes to this literature by showing how information can be extracted from large data sets using XGBoost, and how these results can be interpreted using SHAP (SHapley Additive exPlanations), an "explainable ML" method based on game theoretic Shapley values. Our analysis provides sharp results on the value of knowing odd lot data in equity markets.

This paper is organized as follows. The next section sets out the data, sample period, stock sample selection criteria, and our categorization of stocks into price-level based groups. In section 3 we analyze the inside market, examining the distribution of orders sizes, odd lot bid and ask prices, odd lot spreads, the incidence of better inside quotes relative to the NBBO, and depths. We also provide regression analyses of stock level predictors of superior odd lot prices and of depth inside the NBBO. Section 4 provides results from our machine learning analysis of the predictive information contained in odd lot information. We discuss the accuracy and precision of our predictions and we identify the specific contributions of each feature. We also provide results from a simple trading strategy based on our predictions. In Section 5 we provide results from a counter-factual analysis of the SEC's upcoming redefinition of odd lots. We conclude in Section 6 with a discussion of the problems posed by the inside market and how these might be addressed.

2. Data

Our analyses rely primarily on a set of proprietary data feeds. As noted, we obtain these data feeds from May Street, the data vendor that supplies to the SEC the microdata underlying the MIDAS system.[8] For recent years, these data capture all (non-hidden) trade and quote messages for all exchanges in the United States. Traditional market microstructure analyses rely instead on the Trade and Quote (TAQ) data. For our purposes, there are two major advantages of the May Street data, relative to the traditional TAQ data analyzed in much of the market

[8] Prior to 2019, the data vendor was Thesys Technologies, but the contract is currently with May Street. See https://maystreet.com/maystreet-to-provide-market-data-for-the-u-s-securities-and-exchange-commissions-midas-platform/, last accessed September 15, 2021.

microstructure literature. First, they have information on odd lot quotes. Second, they have complete order book information for each exchange. In contrast, the TAQ data have information on round lot quotes only and only have top-of-book quote information for each exchange. These twin advantages allow us to speak to both price and volume: the extent to which the best odd lot quotes are inside the national best bid and offer (NBBO), as well as how much total odd lot quotation volume rests inside the NBBO.

Our analysis sample encompasses trading days from January 4, 2021, to March 31, 2021. We focus on a recent time period because odd lots are increasingly prevalent, so studying them is easier in more recent data.[9] To substantiate this point and also to convey a sense of what is known about odd lots in recent years using traditional data, we turn briefly to TAQ trade data. As many previous papers have noted, TAQ does not (yet) have information on odd lot quotes, but it does reflect odd lot *trades*. Figure 1 (panel A) presents a time series plot of the fraction of trades that are odd, both in terms of transactions (black dots, left axis) and in terms of shares (gray dots, right axis).[10] In late 2016, the fraction of trades that are odd was around 25 percent, but since that time it has increased substantially; by 2021, over 50 percent of trades were odd. Since odd lots are by definition for fewer than 100 shares, the fraction of shares traded that are part of an odd lot trade is lower in terms of its level than the fraction of trades that are odd, but as panel A shows, the volume-weighted fraction odd exhibits similar trends, rising from just over 4 percent to over 8 percent.

These trends are not driven by illiquid or obscure stocks. Panel B shows the same trend line for S&P500 stocks and non-S&P500 stocks. In late 2016, about 30% of trades for S&P500 stocks were odd, but by 2021 that figure had risen to over 65%. The corresponding figures for non-S&P500 stocks are 20% and 50%, respectively. These fractions are of course lower if weighted by the number of shares involved, but panel C shows that there are stocks where odd lot trades account for a large fraction of overall volume; the examples shown are Chipotle, BlackRock, Netflix, Tesla, and Facebook, which are companies spanning a wide space of industries. For each of these example stocks, the volume-weighted fraction odd has risen over

[9] A side benefit of our focus on the most recent period is that May Street data may not have complete coverage for historical data. Prior to 2019, for example, Thesys Technologies was the data vendor supplying to the SEC the microdata undergirding the MIDAS system. During the Thesys era, packets were frequently lost in transmission—to such an extent that TAQ might have had better coverage than MIDAS.

[10] The figure starts on 10/24/2016, the first day one can exclude test stocks using information in the TAQ master file.

the period. Fully 25% of Facebook volume these days is odd, and for the remaining stocks that figure is 32%, 44%, 47%, and 56%, respectively. Notably, these are all expensive stocks, which raises the interesting point of why companies do not stock-split their way out of odd-lot dominance. An obvious indicator of the influence of high price on odd lots is the five-for-one stock-split for Tesla on August 31, 2020, as is clear from the figure. Tesla's volume-weighted fraction odd fell decisively on that day from 51% to 28%. Panel D emphasizes the heterogeneity of the shift towards odd lots by juxtaposing the 2017 and 2021 histograms of the stock-day distribution of the fraction of trades that are odd. Whereas in 2017 three-quarters of stock-days were below 35% odd, by 2021, three-quarters were above 35%.

The interface for the May Street data is query-driven, so the next step in our study design is to pick a set of stocks to query. A core consideration for us is liquidity. Part of our study is a set of prediction models seeking to understand the extent to which odd-lot quote activity is predictive of future changes to the VWAP. Consequently, we need observations on VWAP throughout the trading day, and this means our focus will be on stocks with sufficient liquidity for this to be the case.

To determine which stocks might be sufficiently liquid, we turn to the same TAQ trade data from Figure 1, but focus on December 2020, just before our analysis period.[11] For each stock-day in December, we compute the largest gap in time between two consecutive trades during the trading day.[12] For each stock, we take the maximum of those maxima, and restrict attention to the set of stocks for which that maximum gap is less than 5 minutes.

This leads to a set of n=1,751 stocks, which we refer to as our "study stocks." These are by design relatively more liquid than other stocks. They contain nearly all (501 of 505) of the stocks comprising the S&P500.

As noted, odd lots are particularly common among expensive stocks. We divide our set of 1,751 stocks into 5 groups based on the average VWAP observed across full trading days in December 2020. We use cut points of $20, $50, $100, and $250. Figure 2 gives the time series of the fraction of trades that are odd for our set of 1,751 stocks (black line), as well the time

[11] We exclude Christmas Eve, 12/24/2020. We also use fields in the TAQ master file to exclude test stocks and to keep only common stocks.
[12] Following Bartlett and McCrary (2019), we use the participant timestamp unless it is zero or prior to the SIP timestamp, in which case we replace it with the SIP timestamp. We allow for all timestamps from 9:30:00 to 16:00:00. We also exclude trades with a trade stop indicator of Y and require that the trade correction indicator be 00 or 01.

series for each of the 5 groups (gray lines). The figure shows two things. First and unsurprisingly given how we chose them, our stocks are generally similar to the S&P500 in terms of odd-lot prevalence. That is, the black line in Figure 2 is similar in level and trend to that in panel A of Figure 1. Second, the 5 groups of stocks are highly dissimilar in terms of odd-lot prevalence. Over time, the fraction of trades that are odd has risen for all 5 groups, but the levels are notably different. Speaking generally, price roughly orders the extent of odd lots. For example, for group 1, the fraction of trades that are odd rises from just over 15% to 40% and for group 5, the fraction rises from 40% to 83%. For convenience, Figure 2 displays a gray area ("recession shading") corresponding to the sample period.

The next step in our study design is to query the May Street system for information on these stocks during our sample period. We use two sets of queries from the May Street data.[13] The first of these relates to our analysis of the size of individual messages to add liquidity. For this analysis, we pull every message to add liquidity across the 1,751 study stocks during the sample period.

The second set of queries focuses on pulling data for the full order book each day for the study stocks at the beginning of each minute in our study period. In particular, we examine the entire order book at 390 snapshots in time—once at the beginning of each minute between 9:30am and 4pm—each day. Economizing on analysis in this way is efficient: although this discards much of the data, for our 1,751 study stocks during the 61 trading days of our sample period, examining 390 snapshots during the day still involves analysis of 41.6 million records. Variables included are the set of 20 best bids (price and size), 20 best offers (price and size), prevailing NBBO (price and size), and VWAP during the ensuing minute.

Figure 3 provides an example of how the top of the book looks to a consumer of the SIP/TAQ data relative to a consumer of the exchanges' proprietary data feeds using Amazon common stock at 13:25:00 on March 31, 2021. At the beginning of this minute, the TAQ data (Panel A) indicate that the national best bid (NBB) stood at $3,110.67 due to a round lot bid at

[13] In TAQ, the identifier for a security is the pair (sym_root,sym_suffix). In the May Street data, the identifier is a single field, symbol. We developed a crosswalk from TAQ to May Street as follows. For tape A securities, symbol is equal to the concatenation of the TAQ identifiers, with a period inserted (e.g., BRK.B). For other securities, symbol is equal to the concatenation of the TAQ identifiers with no period inserted (e.g., DISCA). The one exception to this of which we are aware is the stock TrueCar Inc., which in TAQ has a sym_root value of TRUE and blank sym_suffix, but which in the May Street data has a symbol value of True.

this price on both Nasdaq and MIAX, and the national best offer (NBO) stood at $3,113.91 due to a single round lot order to sell on Nasdaq. Thus, the quoted spread in Panel A is $3.24.

In contrast, as shown in Panel B, a consumer of the exchanges' proprietary data feeds would see very difference prices due to the presence of odd lot quotes. Specifically, Panel B shows that the best bid price at this time was $3,112.59 ($1.92 higher than the NBB) due to an order to buy 36 shares at this price on BATS Z and an order to buy 1 share at this price on Nasdaq. Panel B also shows that the best offer price at this time was $3,113.19 ($0.72 less than the NBO) due to an offer to sell 2 shares at this price on Arca. Overall, the quoted spread for these odd lot orders is just $0.60, or less than one-fifth of the quoted spread of the NBBO. More generally, Panel B also highlights the large number of odd lot quotes that rested at superior prices to the NBBO. For instance, while the SIP/TAQ data show total bid depth of 200 shares, or approximately $622,000, the May Street data show that the aggregate depth of superior priced bids was 405 shares, or approximately $1.3 million. On the offer, the SIP/TAQ data show total offer depth of 100 shares, or $311,391, while the May Street data show aggregate depth of superior priced offers amounting to 131 shares, or approximately $408,000. Lastly, Panel B also highlights how the exclusion of odd lot quotes in the SIP/TAQ data can overlook additional depth at the NBBO. In Panel B, this is illustrated by the fact that an order to buy 50 shares at the NBB price rested on Arca but does not contribute to the depth at the NBB in Panel A. The SIP/TAQ data in Panel A also exclude orders to sell shares at the NBO on BATS Y and BATS Z for 10 shares and 1 share, respectively.

3. Odd Lot Orders and the NBBO

3.1 Summary Statistics

Table 1 summarizes our study stocks during the sample period. As noted above, we organize our 1,751 study stocks into 5 groups based on the December 2020 VWAP. These groups are defined as follows: group 1 (VWAP<=$20), group 2 ($20<VWAP<=$50), group 3 ($50<VWAP<=$100), group 4 ($100<VWAP<=$250), and group 5 (VWAP>$250). These groups each have 100 or more stocks, but there are more stocks in group 1 (504) than there are in group 5 (109). Unsurprisingly, the VWAP is lowest for group 1 and highest for group 5. For groups 1-4, there is little difference between the mean and median VWAP, but for group 5, where the price range for VWAP is unbounded, there is notable skew to the right in the distribution: median VWAP is $353.32, but mean VWAP is a large $507.06.

Using obvious terminology, we call the best (highest) round lot bid the SIP Bid, and we call the best odd lot bid the OL Bid. Similarly, we call the best (lowest) round lot ask the SIP Offer and the best odd lot ask the OL Offer. The SIP Bid and SIP Offer are together the NBBO, and the OL Bid and OL Offer are together the OL NBBO. By regulation, only round lots are eligible for inclusion in the NBBO. The OL NBBO thus by definition has the potential to reflect better prices than the NBBO. The spread is defined as SIP Spread = SIP Offer – SIP Bid. The corresponding odd lot concept is something we refer to as the OL spread and is defined as OL Spread = OL Offer – OL Bid. By definition, OL Spread < SIP Spread.

For stocks in group 1, the SIP Spread is quite narrow for group 1, and larger for other groups, ranging from roughly one to two cents for group 1 to almost a dollar for group 5. The average SIP Spread is 2.5 cents for group 1, 6.6 cents for group 2, 13.4 cents for group 3, 24.8 cents for group 4, and 91.8 cents for group 5. For each of the five groups, the median SIP Spread is somewhat tighter. For example, for group 1 it is 1.1 cents and for group 5 it is 61.1 cents.

As noted, it is definitionally true that the OL Spread has to be tighter than the SIP Spread, and Table 1 shows that there is variation in how much tighter it is across groups. Specifically, the OL Spread is somewhat tighter than the SIP Spread for group 1, but much more so for higher priced stocks, especially group 5. For example, the average OL Spread for group 1 is 12% tighter than the SIP Spread, but for group 5 it is over 26% tighter than the SIP Spread.

The fact that the OL Spread is tighter than the SIP Spread, particularly for higher priced stocks, does not necessarily say much at all about whether there is meaningful liquidity among odd lots. A stock could have a single share on offer inside the NBBO—even far inside the NBBO—but that might not put a trader in a position to get filled at a meaningfully better price if they were seeking execution for a larger number of shares. To assess the extent to which there is significant odd lot depth inside the NBBO, we next compare the depth inside the NBBO to the depth at the NBBO.

To do so, we look at three metrics. To describe these, consider the bid side. We examine the dollar value at the SIP Bid ("Depth at Top of SIP Bid"); the dollar value at the OL Bid ("Depth at Top of OL Bid"); and finally the cumulative dollar value on offer at prices strictly better than the SIP Bid, all the way up to the OL Bid ("Total Depth at OL Bids"). If there are no odd lots better that the SIP Bid, then Depth at Top of OL Bid is set to missing, and similarly for Total Depth at OL Bids. Thus, the Table 1 statistics for these measures indicate the amount of depth inside the SIP Bid when the OL Bid is better than the SIP Bid. These metrics are constructed analogously on the offer side.

The idea behind these metrics is as follows. Suppose a trader wants to buy 100 shares of stock XYZ. Suppose there are 100 shares on offer at the SIP Offer, 1 share on offer at the OL Offer, and 99 shares on offer from just below the SIP Offer all the way down to just before the OL Offer. If our trader submitted a marketable buy order to this venue for 100 shares, the order would execute against the 100 shares on offer at prices superior to the SIP Offer because exchanges honor price-time priority. As a result, this trader would purchase 100 shares of XYZ at better prices than were available at the SIP Offer due to the odd lot quotes.

We begin by examining the dollar value available at the SIP Bid ("Depth at Top of SIP Bid"). This variable exhibits a long right-hand tail, so medians might be a more reliable measure of the distribution's spread. Medians for groups 1 through 3 are approximately \$20,000 to \$30,000. For groups 4 and 5, median Depth at the Top of the SIP Bid rises to \$50,000 and \$100,000, respectively. The dollar value available at the SIP Offer shows a similar cross-group pattern as the SIP Bid, with the SIP Offer rising from a little over \$20,000 for group 1 to just over \$100,000 for group 5.

As might be expected, the dollar value on offer at the OL Bid is notably smaller than the dollar value on offer at the SIP Bid, and this pattern is stronger for group 1 than for the more expensive groups. For example, for group 1, median Depth at Top of OL Bid is under 2% of the dollar value associated with Depth at Top of SIP Bid. This rises to 4% for group 2, 5% for group 3, 6% for group 4, and nearly 7% for group 5.

Taking into account all of the odds lots between the SIP Bid and the OL Bid has little effect for groups 1 and 2, but roughly doubles the above figures for group 5. For group 1, the median Total Depth at OL Bids is 2%, for group 2 it is 5%, for group 3 it is 7%, for group 4 it is 10%, and for group 5 it is almost 14%. For group 5, in particular, this is meaningful volume on offer inside of the NBBO.

While these figures were specific to the bid side, the numbers in Table 1 regarding the offer side are highly similar quantitatively, and certainly convey the same qualitative conclusions. For example, for group 5, the median Total Depth at OL Offers is 14% of the dollar value associated with the Depth at Top of SIP Offer.

To develop further intuition regarding these high level conclusions, we next examine the distribution of order sizes. We begin by considering two very different sample stocks during our sample period. The two stocks occupy opposite ends of the price spectrum. The first is GEVO (Gevo, Inc.) whose price during our sample period rose from around $4.50 at the beginning, to a peak around $15 during the middle, before eventually falling back to approximately $9 by the end. The second stock is AMZN (Amazon.com, Inc.), which throughout out sample period hovered around $3,200.

In general, GEVO is not nearly as liquid as AMZN—but at its price point, it is liquid *enough* that the penny tick is typically binding. On average, the SIP Spread for GEVO was a penny for 336 of the 390 observed stock minutes (87%) for each sample day, and it was two pennies in 37 of the 390 observed stock minutes (9%). Whether the penny tick binds is an important consideration, because when it does, no odd lots can exist inside the NBBO by virtue of the prohibition on sub-penny quotes. For GEVO, the NBB and NBO are usually so close to one another that there is either (a) literally no space for aggressively-priced odd lots—any quote that improved either side would lock the market—or (b) only one penny of space for odd lots better than the NBBO. Were the penny tick to be binding for most stocks, then odd lots could not be worth examining.

AMZN furnishes a different example, one where odd lots have the potential to be relevant. While AMZN is one of the most liquid stocks in the market today, it is so expensive that the penny tick almost never constrains the spread. Across the 23,790 stock minutes observed for AMZN in the sample, only 2 had a SIP Spread of two cents or less. In contrast, the SIP Spread was over 50 cents for 98.7% of all observed stock minutes for AMZN. In other words, for a stock like AMZN, there is ample space for odd lot quotes inside the NBBO.

To assess whether the potential importance of odd lots for AMZN is in fact a reality, we turn to an examination of the distribution of order sizes for these two stocks. To do so, we queried the May Street system for all messages adding liquidity for these two stocks during our sample period. Figure 4 summarizes our results in this regard.

The figure is the histogram of order size for all messages adding liquidity during our sample period for these two stocks, GEVO and AMZN. For visual ease, we restrict attention to orders of size less than 500. For GEVO, messages adding liquidity are overwhelmingly round. Over 60% of such messages are for 100 shares, roughly 18% of such messages are for 200 shares, with smaller fractions for 300 and 400 shares. Only 8% of such messages are for odd lots. Interestingly, orders that are less than 100 shares are most common at 1 share.

For AMZN, and at the risk of understatement, the picture is quite different. Almost 40% of messages adding liquidity are for a single share, and 8-12% of such messages are for two or three shares. For AMZN, the fraction of messages adding liquidity that are 100 shares is around one-third, with very little mass to the distribution beyond that.

Figure 5 shows that the order book patterns shown in Figure 4 are also reflective of actual transactions and shows the histogram of trade size for GEVO and AMZN during our sample period. For GEVO, round lot trades are clearly more common than odd lots, but odd lot trades are not negligible. Focusing on trades of size less than 500, fully 45% of GEVO trades are odd. The sizeable number of odd lots trades relative to odd lot quotes is due to the fact that the order book analysis reflects only orders on exchanges, whereas trades can and often do occur on off-exchange venues, reflect internalization, etc. (i.e., "D" trades). Notably, in unreported results, we find that nearly 60% of non-exchange trades for GEVO during the sample period are odd lots, which may reflect the internalization of retail order flow that is conventionally associated with odd lot trading.

For AMZN, the story is simple. Again, focusing on trades for fewer than 500 shares, fully 36% of trades during the sample period are for a single share; 9% are for 2 shares. Overall for AMZN, the fraction of trades for less than 500 shares that are odd is 95%.

In Figure 6, we generalize the order book analysis that we conducted for GEVO and AMZN, providing a histogram of order size (in bins) for our five groups of stock. The bins are for 1-49 shares, 50-99 shares, 100-199 shares, 200-299 shares, 300-399 shares, 400-499 shares, and 500 or more shares. The results show that grouping comes close to ordering the distribution. This analysis is quite demanding in terms of computation, as there are for our study stocks during our sample period fully 32.3 billion orders adding liquidity during trading hours.

Turning to the results, consider first group 1—stocks more like GEVO. For group 1, there is very little mass below 100 shares. Panel A of Table 2 quantifies this result: the first two bins are approximately 5% of the distribution for group 1. In contrast, approximately 60% of the distribution is 100-199 shares. Roughly 20% of the distribution is for 200-499 shares, and 14%

are for 500 or more shares. These are quite large orders and are notable in part because for the other four groups, orders of that size are unusual.

Turning to group 5—the group that is more like AMZN, or stocks such as Chipotle, BlackRock, Netflix, Tesla, and Facebook that we examined in panel C of Figure 1—a large 42% of orders are for 1-49 shares. Orders for 50-99 shares are also prevalent, around 5% of the total. A smaller portion—about half—of orders are for 100-199 shares, and there are few orders for 200 or more shares.

Turning to the other groups, and considering the first bin for 1-49 shares, the fraction of the distribution in the first bin is lowest for group 1, second lowest for group 2, and so on, culminating in the highest fraction for group 5. Considering the largest bin for more than 500 shares, the pattern is the opposite: the largest fraction is for group 1, the second largest is for group 2, and so on, culminating in the smallest fraction for group 5. The same pattern repeats for bins 200-299, 300-399, and 400-499. For the intermediate bins—that is, for bins 50-99 and 100-199, these two conflicting patterns are battling, and the results are no longer cleanly ordered. For example, for 100-199 shares, the group with the greatest prevalence is neither group 1 nor group 5, but group 2. For group 2, 70% of all orders are for 100-199 shares. For group 2, the tendency towards larger orders is not as strong as that of group 1, but stronger than that for the other groups.

The final row of panel A of Table 2 summarizes across all groups. This shows that for all 1,751 of our study stocks, over 15% of orders adding liquidity are odd. Some 64% of such orders are for 100-199 shares; the fraction falls to 7% for 200-299, 3% for 300-399, 2% for 400-499, before rising to nearly 7% for over 500.

Panel B of Table 2 shows that there is substantial heterogeneity across venues in these patterns. Inverted exchanges are attractive for odd lot liquidity providers, but they are not the only exchanges with high fractions of odd lot quotes. Nasdaq's BX is clearly the exchange with the greatest extent of odd lot quotes: odd lots represent fully 37% of orders adding liquidity on BX. The other two inverted exchanges, EDGA and BATSY (CBOE BYX), have 18% and 16% odd lot quotes, respectively. However, the fraction odd lot quotes for AMEX, NYSE, and NASDAQ is 8%, 13%, and 22%, respectively, and EDGX has a higher fraction odd lot quotes (23%) than its sister exchange EDGA.

3.2 Do Odd Lot Quotes Provide Superior Pricing?

So far, we have established that: (a) odd lots quotes are a common form of liquidity provision in the market today, (b) that they are unevenly distributed across stocks, but particularly prominent among more expensive stocks, (c) concentrated on some exchanges, and (d) that there is non-negligible volume available at odd lots with superior pricing to the NBBO, particularly for those same expensive stocks that attract odd lot liquidity in the first place.

A. The incidence of Odd-Lot Pricing

We now turn to an examination of four related questions: (1) How often is it right that odd lot quotes offer better prices than those at the NBBO? (2) On a minute-by-minute basis, what characteristics are associated with odd lot prices being better than the NBBO? (3) How much odd lot depth is there relative to depth at the NBBO? (4) On a minute-by-minute basis, what characteristics are associated with high levels of better-priced odd lot depth relative to depth at the NBBO?

We approach answering the first question in a simple but direct way. For each stock, we first calculate the number of minutes on a given day where the OL Bid is better (higher) than the SIP Bid. For each stock, we then examine the distribution of those minutes across the 61 trading days in our sample. Figure 7 provides an illustration using again the comparison of GEVO and AMZN. For GEVO, it is extremely common for the SIP Bid is be the best available bid price displayed across all exchanges. More precisely, across the 390 minutes observed in each trading day, the OL Bid offered superior pricing to the SIP Bid for six or fewer minutes in over 70% of the 61 trading days. The results for AMZN could not be more different. For AMZN, of the 390 minutes during the trading day, there is an odd lot better than the SIP Bid for 200-300 of the minutes across 100% of the 61 trading days.

Table 3 summarizes the results of this analysis for all stocks in the study. As in Figure 7, for each stock-day, we compute the fraction of the 390 end-of-minute observations where the OL NBBO is better than the SIP NBBO ("Superior Odd Lot Pricing"). We do so separately for bids (panel A) and offers (panel B). To convey a sense of the distribution of Superior Odd Lot Pricing, we present the mean, standard deviation, and 25th, 50th, and 75th percentiles. The average stock-day for stocks in group 1, has Superior Odd Lot Pricing about 5% of the time. The distribution exhibits strong right-hand skew: the median is far below the mean at 1%, the 25th percentile is well below 1%, but the 75th percentile is a high 7%. Unsurprisingly given earlier results, the distribution of Superior Odd Lot Pricing shifts strongly to the right as you move from group 1 to group 5. For example, the mean Superior Odd Lot Pricing moves from 5% for group 1, to 14% for group 2, to 22% for group 3, to 30% for group 4, to fully 42% for group 5. For at least 27 stocks (those in the 75th percentile or above for the 109 group 5 stocks) just over half the trading day (50.5%) goes by where the OL Bid is better than the SIP Bid.[14]

Figure 8 highlights the shift to the right described above. It displays the full distribution of Superior Odd Lot Pricing for groups 1 (blue) and 5 (red). For group 1, there is a decisive mass point at zero (for over 40% of stock-days, at the end of none of the 390 minutes during the trading day is it true that the OL Bid is better than the SIP Bid) and a rapid decline thereafter. However, for group 5, the distribution is centered at 0.4 and the distribution is approximately symmetric. That means that for about half of the stock-days, 40% of the 390 minutes during the trading day had Superior Odd Lot Pricing.

Returning to panel B of Table 3 covering Superior Odd Lot Pricing for offers instead of bids, we see largely the same picture. The correlation between the 25 entries in panel A and those in panel B is 0.999. For example, the mean Superior Odd Lot Pricing for group 1 in panel A (bids) is 5%, whereas that in panel B (offers) is 4%; the 75th percentile of Superior Odd Lot Pricing for group 5 in panel A is 51%, whereas that in panel B is 49%. The similarity of the distribution on the bid and offer sides is reinforced by Figure 9, which gives a histogram (overall across groups) of Superior Odd Lot Pricing, separately on the bid and offer side. The figure shows that these are two highly similar distributions.

In Table 4, we turn to descriptive regressions for the bid and offer side separately, where the outcome is an indicator for whether in the given minute the OL Bid (odd columns) or OL Offer (even columns) is better than the SIP Bid and SIP Offer, respectively. Because we intend this simply as a descriptive regression, the covariates are measured as of the same minute as the outcome. The specific covariates we consider include stock fixed effects, date fixed effects, the log of the SIP Spread, the log of the VWAP, the log of volume (total number of shares transacted within the minute), the volatility (standard deviation of the return), and return (the change in log price from the one minute to the next).

Overall, the regressions reported in Table 4 are modestly, but not overwhelmingly predictive: across the four columns, the R^2 ranges from 0.13 to 0.15. Focusing attention on column (1) and considering first the effect of the SIP Spread, the coefficient on log spread is

[14] We say "at least" because there may also be such stocks in groups 1-4.

about 0.10. The statistical significance is high (with a t-ratio of over 55).[15] However, statistical significance does not necessarily speak to economic significance. To interpret the coefficient, consider the following: suppose the spread increased by 10% and suppose the coefficient is 0.10 exactly. Then the predicted increase in the probability that the OL Bid would be better than the SIP Bid is 0.01.[16]

The coefficient on log price is similar in magnitude but of the opposite sign, and marginally significant despite a sample size of some 40 million observations. For a 10% increase in price, the predicted change in the probability that the OL Bid would be better than the SIP Bid is -0.001, i.e., a small effect. The effects of volume are even smaller—a 10% increase in volume is predicted to lead to a change in probability of -0.00002.

On the other hand, volatility and return seem to have larger effects. A ten percent increase in volatility is associated with a 1.3 percentage point decrease in the chances that the OL Bid is better than the SIP Bid. Similarly, suppose that the return to the stock over the last minute increases by 10 percentage points. Then the chances that the OL Bid is better than the SIP Bid decrease by 1.3 percentage points.

Turning to the results in column (2), we reach broadly similar conclusions, but with occasional sign changes. Higher spreads are associated with a better chance that the OL Offer is inside the SIP Offer. Higher prices are associated with lower chances that the OL Offer is inside the SIP Offer. Volume has more of a statistical association with the offer side than the bid side, but the magnitude of the effect is still small. For volatility and return, the sign of the effect is the opposite what it was for the bid side. For return, this is an easy story. For the bid side, an increase in return means that more quotes become aggressive, diminishing the likelihood that the most aggressive bid will be less than a round lot. Likewise, for the offer side, an decrease in return means that more quotes become aggressive on the sell-side, again diminishing the likelihood that the most aggressive offer will be less than a round lot. For volatility, this is consistent with the idea that when a stock is volatile, the (OL) Spread widens.

Columns (3) and (4) extend this analysis to contemplate interactions of each covariate with indicators for group. For log spread and log price, these interactions do not appear to be very

[15] The inference approach we adopt here is clustering on the stock. This allows for arbitrary correlations
[16] That is, the predict change in the outcome associated with a 1 percent increase in the spread is the log spread coefficient relative to 100.

important. For log volume, it appears that there is substantial heterogeneity: log volume seems quite important for group 5 and is predictive of increased chances of the OL NBBO bettering the SIP NBBO for both the bid and offer side, and by similar magnitudes. For log volatility, the patterns across group are hard to interpret. For log return on the other hand, the patterns are quite clear. For group 5, the effect sizes are 5-6 times as large as they are overall. For group 1, the effect sizes are a fifth to a third as large as the effects overall.

B. Odd-Lot Depth

We now turn to addressing the third question outlined above—namely the extent of better-priced odd lot volume relative to the volume at the NBBO. Following the pattern of Table 3, Table 5 starts by calculating for each stock-day the cumulative odd lot volume on offer that is better than the round lot volume. There are two approaches to analyzing this question. The first is an unconditional analysis, whereby when there are no odd lots better than the SIP NBBO, the odd lot volume is considered to be zero. The second approach is a conditional analysis, where we ask: "When there are odd lots, how much odd lot volume is there relative to round lot volume?" Roughly speaking, the unconditional approach is "include the zeros" and the conditional approach is "truncate the zeros." Panel A of Table 5 adopts the unconditional approach and panel B of Table 5 adopts the conditional approach.

Turning to the results in panel A of Table 5, we see first that the table mimics the structure of Table 3. For each of our five groups, we put forward the mean, the standard deviation, and the 25^{th}, 50^{th}, and 75^{th} percentiles of the distribution across stock-days of the stock-day specific estimate of the average across 390 minutes of the ratio of Total OL Bid (Offer) Depth to SIP Bid (Offer) Depth. For group 1, the typical stock-day has an average Total OL Bid (Offer) Depth relative to SIP Bid (Offer) Depth of 0.005 and 0.005, respectively. For group 5, this rises to 0.09 and 0.08, respectively. That is, for group 5, the cumulative odd lot depth is about 9 percent of the inside depth. We take that to be a large and meaningful fraction of volume.

The structure of panel B is similar, but adopts the conditional approach. By construction, the numbers here are larger. What is surprising is the extent to which they are larger, quantitatively. For example, for group 1, for stock-day-minutes when there is better odd lot pricing than SIP pricing, the cumulative fraction of odd lot volume relative to inside depth is 16% for both the bid and the offer side. For group 5, this fraction rises to 20% for both sides.

That is, for group 5, when odd lot pricing is better than the SIP NBBO, a trader could find about a fifth as much volume in the odd lots as the trader sees at the SIP NBBO.

In Table 6, we turn to descriptive regressions designed to see what covaries with there being a great deal of odd lot depth. Turning to column (1) pertaining to OL Bid Depth relative to the SIP Bid Depth, we see that log spread is statistically significant with a t-ratio of over 44. However, this does not necessarily say much about the economic significance of the coefficient. The coefficient on log spread of 0.0246 means that a 10% increase in spread is associated with an increase in the fraction of depth that is odd of 0.00246—which we would characterize as something of a small effect.

The effects for log price are similarly muted: a 10 percent increase in price is predicted to lead to a decrease in the fraction of depth that is odd of 0.0007. The effects of volume are yet smaller. Effects for volatility and return are similarly small. For each variable, a 10% increase is predicted to lead to a decrease of about 0.004 in the fraction of depth that is odd. Perhaps predictably given these small effects, the R^2 values from the regressions are notably smaller than from the regressions in Table 4, and hover around 7%.

We turn next to predictive models, where we try to see whether information pertaining to odd lots is predictive of changes to market price in the subsequent minute.

4. Do Odd Lot Quotes Predict Future Trade Prices?

Given that odd lot quotes are presently excluded from the SIP data, does access to odd lot quotes provide traders with more valuable information than SIP data alone? We approach this question in two steps. First, using both linear regression and machine learning techniques, we ask whether odd lot quotes are predictive of future trade prices. Second, we ask whether odd lot quotes provide traders with exploitable trading opportunities that are not available to traders using only the SIP data.

4.1 Set up

In designing our predictive models, we focus on the capacity of quotation data to predict prices over the 1-minute interval that follows the observation of an odd lot quotation. In particular, because our data represent a snapshot of the quoting environment at the beginning of each minute throughout the trading day, we examine whether the quote data observed at the beginning of minute *t* are predictive of the volume-weighted average price observed through the commencement of minute *t+1*, which we refer to as $VWAP_{t+1}$. In our regression analysis, we focus on the extent to which quotation data can explain the log difference between $VWAP_{t+1}$ and the SIP midpoint observed at *t*. In our machine learning analysis, we focus on whether the quote data can predict whether $VWAP_{t+1}$ is above the SIP midpoint at *t*.

For our baseline "SIP model," the quotation data include only the following data that are available from the SIPs (all observed at the commencement of *t*):

SIP Spread:	The quoted spread based on the best bid and offer shown by the SIPs
SIP Midpoint:	The midpoint of the best bid and offer shown by the SIPs
SIP Imbalance:	The difference between the displayed depth at the SIP bid and SIP offer
Prior SIP MP Return:	The log difference between the *SIP Midpoint* from the start of the minute *t-1* to the start of the minute *t*

The rival "OL model" adds the following odd lot quotation data from the proprietary data feeds (all observed at the commencement of *t*):

OL Spread:	The quoted spread based on the best bid and offer across all exchange data feeds
OL Midpoint:	The midpoint of the best bid and offer across all exchange data feeds
Difference:	The log difference between the *OL Midpoint* and the *SIP Midpoint*
OL Imbalance:	The difference between the displayed depth at the best bid and offer based on exchange data
Prior OL MP Return	The log difference between the *OL Midpoint* from the start of minute *t-1* to the start of the minute *t*
OL Bid Depth	The cumulative depth at the bid at odd lot prices that are inside the best bid shown by the SIPs ("SIP Bid"), divided by the depth at the SIP Bid

OL Offer Depth	The cumulative depth at the offer at odd lot prices that are inside the best offer shown by the SIPs ("SIP Offer"), divided by the depth at the SIP Offer
Better Bid	An indicator for whether odd lot quotes exist that are superior to the SIP Bid
Better Offer	An indicator for whether odd lot quotes exist that are superior to the SIP Offer
Bid Level	Count of the number of price points where superior odd lot pricing exists relative to the SIP Bid
Offer Level	Count of the number of price points where superior odd lot pricing exists relative to the SIP Offer

Lastly, to account for the possibility that prices between t and $t+1$ are influenced by prior trades, we also include as parameters in both models the natural log of the volume of trades observed between $t-1$ and t, the natural log of the cumulative volume of trades since the beginning of the trading day through t, and the log difference between the return on the volume weighted average price during the minute $t-1$ and the minute $t-2$.

Note that we make no claim that these parameters or the model specifications are the optimal set of parameters or specifications a trader would use. On the contrary, we purposefully keep the models parsimonious under the principle that, to the extent the odd lot quotation data has meaningful informational content, this informational advantage should be apparent even in a simple predictive model. At the same time, we acknowledge that this approach places a considerable bias against a finding that quote data has informational content. In particular, our sample stocks represent very liquid exchange-traded equity securities, and our outcome of interest is observed over a full sixty seconds. Under these conditions, market efficiency suggests it should be difficult (if not impossible) to find that quote data observed at the beginning of a minute has the capacity to predict future prices.

4.2 Regression Results

We first evaluate the capacity of the two models to explain the cross-sectional variation of our outcome of interest, the log difference between $VWAP_{t+1}$ and the SIP midpoint observed at t. In light of the very different levels of odd lot quotes across our five groups of stocks, we run the analysis separately for Group 1 stocks (low levels of odd lot quotes) and Group 5 stocks (high levels of odd lot quotes) to examine whether the informational content of odd lot quotes varies by stock price.

Table 7 presents the results. Columns 1 and 2 present the results for the baseline SIP Model and OL Model for Group 1 stocks, while Columns 3 and 4 present the same models with stock fixed effects. Across all four models, many of the variables suggest a strong correlation with VWAP returns over the ensuing minute. The positive, significant coefficient on the prior minute's VWAP return is consistent with the existence of an intra-day momentum effect. The prior minute's volume of trading likewise has a significant, positive association with VWAP returns; however, returns are negatively associated with the cumulative amount of volume through that point of the trading day.

Turning to the quotation variables, the SIP Model reveals a negative association between spreads and returns, as one might expect, though the result becomes insignificant with stock fixed effects. The coefficient for *SIP Midpoint* suggests a positive association with price and returns in column 1, but the association is reversed once we add stock fixed effects. In contrast, the estimates for *SIP Imbalance* and for *Prior SIP MP Return* are both strongly significant and constant across both Columns 1 and 2, indicating that strong relative bid depth and a negative return to the SIP midpoint over the past minute are associated with positive future VWAP returns.

Columns 2 and 4 similarly reveal a number of statistically significant associations between the odd lot quote data and the ensuing minute's VWAP returns. Most notably, while the coefficient estimate for *SIP Spread* remains negative, the estimate for *OL Spread* is positive, indicating that odd lot quotes (regardless of size) may have information that is distinct from that contained in the SIP data. A similar conclusion is suggested by the tightly estimated elasticity of 22% for the variable *Difference*. By this estimate, a 10% increase in the OL midpoint relative to the SIP midpoint predicts a 2.2% increase in the VWAP return over the next minute.

Despite these associations, however, the overall model fit is poor for both the SIP Model and the OL Model. Even with stock fixed effects, overall R-squared is just 0.004 for both models, suggesting that these data are quickly incorporated into stock prices. That the result holds across both models no doubt reflects the fact that, as shown previously, OL quotes rarely differ from SIP quotes for Group 1 securities.

Not surprisingly, the results differ for the Group 5 stocks, as shown in Columns 5 through 8. For these stocks, Table 3 showed that odd lot quotes regularly differ from those of the SIP, providing an opportunity to examine whether the associations detected in Columns 2 and 4 may nevertheless have stronger predictive capacity for higher priced stocks. Overall, the results suggest that this is the case. For instance, while VWAP returns are negatively related to *SIP Spread* in columns 5 and 7, the relationship becomes insignificant in columns 6 and 8 where the coefficient for *OL Spread* is statistically negative. Likewise, *OL Imbalance* shows a strong, positive association with VWAP returns in columns 6 and 8, which reduces the positive relationship between VWAP returns and *SIP Imbalance* shown in columns 5 and 8. Among the SIP Model's quote variables, only *Prior MP SIP Return* has a consistent coefficient across columns 5 – 8, which indicates a negative association with VWAP returns.

In addition, the overall model fit differs considerably between the SIP Model and the OL Model for Group 5 stocks. Consistent with Columns 1 through 4, R-squared for the SIP Model in columns 5 and 7 is a modest 0.005. In contrast, R-squared increases more than five-fold to 0.026 for the OL Model in columns 6 and 8. While an R-squared of 0.026 leaves roughly 97% of the variation of VWAP returns unexplained, these results nevertheless highlight a differential in the extent to which prices incorporate the odd lot quote data relative to the SIP data. In the next section, we turn to machine learning to examine whether this differential is sufficient to provide exploitable trading opportunities.

4.3 Machine Learning Analysis

A. Model Specification

In this section, we use a machine learning (ML) model to examine whether odd lot quotes predict future stock prices. Relative to structural regression models, ML algorithms have strong predictive properties for two primary reasons. First, these algorithms can fit complex and flexible functional forms to data in a way that can uncover relationships between a set of predictors and outcomes that are not specified in advance. Second, they rely on regularization techniques that avoid over-fitting to maximize out-of-sample predictive performance. In practice, the predictive performance of machine learning is also enhanced by the adoption of a "firewall principle" (Mullainathan and Spiess [2017]): None of the data that are used in fitting the model are used to evaluate the model's predictive performance. More specifically, prior to fitting the model, the data are split into a "training" dataset and a "test" (or hold out) dataset. The training dataset is used to fit the model and determine all regularization parameters, while the test dataset is then used to evaluate the prediction function that is produced.

For our purpose, we employ a supervised machine learning algorithm in which we fit a function to predict whether $VWAP_{t+1}$ is higher than the SIP midpoint at *t*. The particular function we fit uses *XGBoost*, a gradient boosting algorithm that is closely related to the more common random forest algorithm. *XGBoost*, however, contains several features that lead to modest improvements in a number of predictive tasks (Chen and Guestrin 2016). As in Table 7, we are interested in the predictive capacity of the odd lot quote data, relative to the SIP quote data, as well as in any differences in their predictive capacity by stock price. Accordingly, we fit a "SIP Model" and an "OL Model" using the same parameters for each model utilized in Table 7, and we do this separately for the Group 1 data and the Group 5 data. We thus fit two "SIP Models" and two "OL Models." We summarize the outcome of interest (or "labels") and the predictors (or "features") for the SIP Model and the OL Model in Panel A of Table 8.[17]

Consistent with the firewall principle, we split each dataset such that two-thirds of the observations are assigned to a training dataset, meaning that approximately 7 million stock-minute observations are assigned to the Group 1 training data and approximately 1.9 million stock-minute observations are assigned to the Group 5 training data.[18] Using these training datasets, we fit a SIP Model and an OL Model for each training dataset after "tuning" a set of hyperparameters to optimize predictive performance.[19] We evaluate the models' out-of-sample predictions on the held out portions of the datasets, which contain approximately 3.7 million stock-minutes observations for the Group 1 data and approximately 0.9 million stock-minute observations for the Group 5 data.

B. Model Predictions

1. Results

[17] The predictive performance of the OL Model is modestly improved in the training data if we use the raw difference between the OL Midpoint and the SIP Midpoint rather than its log difference. The results in Table 8 therefore show results for the OL Model that use the raw difference for the feature *Difference.*

[18] In both cases, we sort the data by symbol→date→hour→minute and split the data such that the first two-thirds of symbols constitute our training dataset. We adopt this protocol to ensure that the training data and test data comprise the same trading days.

[19] *XGBoost* contains a set of hyperparameters Θ that jointly determine model representation and optimization. In particular, a number of these hyperparameters affect the degree of regularization imposed during model fitting. To select the best set of hyperparameters Θ⋆ , we perform grid search across a large range of possible combinations of hyperparameters using 5-fold cross-validation. We select the hyperparameter set that maximizes the receiver operating characteristic during this process.

Panel B of Table 8 presents a number of evaluation metrics for the models that were applied to the held-out test data. Evaluating classification models typically requires multiple evaluation metrics in light of a model's potential to have differentials in the rate of true-positives, true-negatives, false-positives, and false-negatives. This can be illustrated by way of the following confusion matrix for the SIP Model for Group 1 data:

		Predicted Values	
		Not Up	Up
Actual Values	Not Up	1,084,088	794,196
	Up	775,270	1,089,731

In the hold-out data, the incidence of true "up" minutes was 49.8%. Given the roughly 50/50 split of outcomes, *accuracy* (i.e., (true positives + true negatives)/all observations) provides a straightforward evaluation metric of the true-positive and true-negative rate. Here, 1,089,731 stock-minutes were correctly predicted to be followed by a $VWAP_{t+1}$ that is higher than the SIP midpoint at *t* (true positives). Likewise, 1,084,088 stock-minutes were correctly predicted not to be followed by a $VWAP_{t+1}$ that is higher than the SIP midpoint at *t* (true negatives). Overall, the model's accuracy was 58%. This is a remarkably high figure given the poor fit for the SIP Model in Table 7, highlighting the classification power of machine learning algorithms. A similar conclusion arises if we focus on the ratio of correctly predicted "up" observations to the total predicted "up" observations (*precision*) or the ratio of correctly predicted "up" observations to all observations that were actually "up" (*recall*). Each of these alternative measures was also 58%. The model's F1 score, which seeks to capture the overall false positive rate and false negative rate, is also 58%, which is unsurprising given that the score is calculated as a weighted average of precision and recall.[20]

We also assess the receiver operator characteristic (ROC) area under the curve (AUC) in Figure 10. ROC is a common evaluation metric for classification models. By default, most classification algorithms determine a class prediction based on whether the predicted probability is greater than 50%. The ROC plot evaluates the true positive rate and the false positive rate of a model using different probability thresholds for a positive classification. A model that perfectly predicted positive cases under any threshold would have an ROC curve that was a vertical line at (0,0) and a horizontal line at (0,1), and the AUC would be 100%. A model that had no predictive

[20] The F1 score is calculated as 2 * (Recall * Precision)/ (Recall + Precision).

capacity would be as a good as a random guess at any threshold, and its ROC curve would be a 45-degree line with an intercept at the origin. The AUC of a model thus provides a measure of the power of the model to correctly classify across all classification thresholds. In Figure 10, Panel A, the AUC for the Group 1 SIP Model is 61%, confirming again the predictive capacity of even the SIP Model within the Group 1 data.

Having established our primary evaluation criteria, we now examine whether access to the odd lot quote data provides any improvement to the model's predictive accuracy. In Panel B of Table 8, Column 2 summarizes the evaluation metrics for the OL Model applied to the Group 1 data. Overall, the evaluation metrics for the OL Model match those of the SIP Model, with the exception of *recall* which is modestly better for the OL Model than for the SIP Model. Likewise, Panel B of Figure 10 provides the ROC curve and AUC metric for the Group 1 OL Model, which is also the same as for the Group 1 SIP Model.

As in Table 7, the results are markedly different when we examine the Group 5 results in Columns 3 and 4 of Panel B of Table 8. Overall, the predictive capacity of both the SIP Model and OL Model is diminished with these data, which likely reflects the diminished volume of trading and quotation activity for higher priced securities. Nevertheless, the SIP Model has an *accuracy*, *precision*, and *recall* of 54%, 54%, and 56%, respectively, in test data that has a true positive rate of 50.6%. More importantly, in contrast to the Group 1 data, these metrics are all higher in the OL Model, where they are 55%, 56%, and 58%, respectively. Likewise, the F1 score for the OL Model is 2% higher, while the AUC score is 3% higher (Panels C and D of Figure 10).

2. Feature Importance

We use the SHapley Additive exPlanations (SHAP) method (Lundberg and Lee, 2017) to explore why the odd lot data are driving these performance enhancements within the Group 5 data. Originating in game theory, the SHAP method estimates the marginal contribution of each feature in explaining an individual predicted outcome, relative to the expected probability within the sample data. In this regard, the SHAP method is a local explanation method insofar that it explains the individual contributions of a coalition of features (the "players") to the outcome of an individual prediction (a "game"). For a classification model, a feature's SHAP value is calculated for each individual observation as the change in the predicted probability of a positive classification due to the addition of that feature, averaged across all possible combinations of the feature values for that observation.

As an illustration, consider the stock-minute for Chipotle depicted in Figure 11, which captures the OL Model at the beginning of 09:34:00 on January 8, 2020. According to the OL Model, the probability that $VWAP_{t+1}$ would be greater than the SIP midpoint at 09:34:00 was 25%. In Figure 11, this is the "model output value". Across all stock-minutes in the training data, the expected probability (or the "base vale") of this occurring was 50.74%; therefore, the OL Model was predicting that the observed features caused the predicted probability to be 26.2% *lower* than the expected probability. The SHAP values explain why. In the figure, the red and blue bars represent features, with the observed value printed below each bar. The blue bars represent a decrease in the predicted probability and the red bars represent an increase in the predicted probability. The width of each bar represents the feature's SHAP value for the stock-minute. Thus, the very wide blue bar for "difference" (which stretches from approximately 25% to approximately 68%) reflects the fact that its SHAP value was -42.9%, and the slightly narrower red bar for "Imbalance SIP" (which stretches from approximately 16% to 25%) reflects a SHAP value of 8.8%. In combination, the aggregate width of all red bars is 20.7%, which means that (absent any blue bars) all of these features would push the predicted probability from 50.74% to 71.47%. Likewise, the aggregate width of all blue bars is -46.9%, which means that (absent any red bars) all of these features would push the predicted probability down from 50.74% to 3.79%. Overall, the aggregate sum of the SHAP values for red and blue bars features is exactly -26.2%, which explains the predicted probability of 25%.

We repeat this analysis for each stock-minute in the training dataset and calculate SHAP values for all features across the 1,916,005 observations. To provide a measure of global feature importance, we take the absolute value of these SHAP values and calculate the mean for each feature. Figure 12 presents the results. Consistent with Table 7, the feature *Difference* has the greatest overall marginal contribution of roughly (4.1%), followed by *SIP Imbalance* (2.8%), *OL Spread* (1.5%), *OL Offer Depth* (0.08%), *Prior SIP MP Return* (0.07%), and *OL Bid Depth* (0.06%). Across the most influential six features, four therefore represented odd lot data.

In Figure 13, we plot all SHAP values for the features used in the OL Model. Each point on the summary plot is a SHAP value for a feature and an instance (i.e., a single stock-day-minute) in the dataset. Positive SHAP values indicate a positive ("up") classification, and the features on the y-axis are ordered according to their importance, as in Figure 12. The color of each point represents the high (red) to low (blue) values of the feature, thereby providing insight into how the marginal contribution of a feature changes according to its value. The influence of *Difference* is clearly evident, and the wide range of SHAP values for this feature is consistent with the example provided in Figure 11. Recall that in Figure 11, the two most influential features were *SIP Imbalance* and *Difference*. In that instance, the SHAP value calculated for the *SIP Imbalance* of 300 shares was a positive 8.8%. However, this positive effect on the predicted probability was overwhelmed by the SHAP value of -42.9% calculated for the -$3.65 value of *Difference* at the beginning of the minute.

Remarkably, the SHAP value for *Difference* indicates that it is the most influential feature in the OL Model despite the fact that it provides no information regarding the amount of depth at odd lot prices.[21]

4.3 Returns

We design a simple trading strategy to examine whether the predictions of the XGBoost algorithm provide potential trading opportunities. Using the held-out test data, we estimate the profits of a round trip trade that begins with a purchase at the commencement of a stock-minute and ends with a sale at $VWAP_{t+1}$. Where a trade does not occur in the next minute, we assume the trader exits the trade by means of selling at the SIP bid prevailing at *t+1.* Critically, however, we limit trading to those stock-minutes where the model predicts a positive label. To account for the trading costs, we calculate the mean effective half-spread for all stock-days in the test data by exchange. For each purchase order, we assume a trader must pay the mean effective half-spread for that stock on the exchange having the greatest trading volume for that stock on the day of the trade.

Applying this trading strategy to the Group 1 test data would result in approximately 80 trades per minute, or 30,000 trades per day between January 4, 2021 and March 31, 2021 for both the SIP Model and the OL Model. For the Group 5 test data, the strategy would result in approximately 20 trades per minute, or 8,000 trades per day.

---

[21] Notably, the overall incidence of $VWAP_{t+1}$ being greater than the SIP Midpoint at time *t* is 75% when *Difference* is greater than $1.00, and the overall incidence of $VWAP_{t+1}$ being less than the SIP Midpoint at time is 75% when *Difference* is less than -$1.00. This associations help explain the large magnitude of SHAP Values (positive and negative) for *Difference* in Figure F.

Visual observation of the data reveals that the imposition of trading costs results in negative returns across all models; therefore, purchases were additionally restricted to those stock-minutes that were especially likely to result in a higher $VWAP_{t+1}$ by increasing the probability threshold for a positive label. That is, rather than classify as a positive prediction any stock-minute with a predicted probability of 50% or higher, we incrementally increased the probability threshold for a positive classification to examine whether higher thresholds could yield positive returns to any of the models.

Panel A of Figure 14 shows the results for Group 1 stocks. As shown in the figure, increasing the probability threshold results in a sharp decline in the number of trades per day. For instance, at a threshold of 80%, many of the trading days have zero trades for both the SIP Model and the OL Model, and no trades occur at all at thresholds above 80%. Moreover, at almost all thresholds, average daily returns are negative. The only exception is that both the SIP Model and OL Model have a modest 0.002% average daily return at a threshold of 80%, but at this threshold, the strategy would permit an average of less than three trades per day.

Panel B of Figure 14 presents the results for Group 5 stocks. As in Panel A, total daily trades declines rapidly as we increase the probability threshold for both the SIP Model and the OL Model. Likewise, the SIP Model generally performs poorly. While the SIP Model achieves positive returns of 0.074% per day at a threshold of 65%, as with Panel A, these returns are generated by just 17 trades per day. Imposing a higher threshold produces negative or zero returns as the number of daily trades declines to zero.

In contrast, the performance of the OL Model is notably stronger for Group 5 stocks. In particular, the OL Model produces average daily returns of 1.42% at a threshold of 65%, which are generated by approximately 461 trades per day. Moreover, the model results in positive daily trades even when we impose thresholds as high as 90%. The fact that only the OL Model produces probabilities of this magnitude is consistent with the greater informational content of odd lot quotes, especially in higher priced stocks where odd lot quotes are likely to differ from round lot quotes.

Lastly, while Figure 14 suggests that average returns from the OL Model are greatest when the probability threshold is set to 70%, it says little about the risk associated with the trading strategy. Therefore, in Figure 15 we additionally present the distribution of daily returns across the sixty-one trading days for the SIP Model and the OL Model using this threshold. As shown in the figure, daily returns for the SIP Model are zero or negative for 48 of the 61 trading days (78.7%). In contrast, daily returns for the OL Model are negative on just 9 of the 61 trading days (14.8%), and positive for the remaining 52 days. Moreover, with the exception of a single day where the daily return was roughly -6%, losses across these 9 days ranged from just -.51% to -1.8%.

5. Is redefining round lots the answer? The SEC and NMS II

Our results demonstrate how odd lot quotes constitute a "market inside the market", providing advantages to those market participants able to observe and access such quotes. The issues we identify here are part of a larger set of issues falling under the general rubric of the collection, consolidation and dissemination of equity market data. What data should be on the consolidated tape, how should it be recorded, who should provide it, what should it cost – these are all questions at the heart of how the U.S. equity markets should operate. The SEC has addressed these questions in Reg NMS II, a new regulatory framework that, while adopted, may be years away from implementation.[22] Much of this proposed framework is beyond our focus here, but the new rules affecting the new definition of a round lot are certainly in our purview. In this section, we investigate how well the SEC's proposed definition addresses the problems posed by odd lot quotes.

As noted earlier, the SEC framework moves away from a standard 100 share round lot definition applied to all (or all but Berkshire Hathaway and a handful of other stocks) listed equities to a price-dependent metric.[23] For stocks with prices at or below $250.00, a round lot stays at 100 shares; for stock prices greater than $250.00 and up to $1000.00, a round lot would be 40 shares; for stocks above $1000.00 up to $10,000.00, 10 shares; and for stocks above that level, 1 share. A second relevant rule change involves the calculation of an exchange's best bid or offer (BBO). The new approach would allow aggregation of some odd lot limit orders to influence the BBO and, therefore, the NBBO.

[22] See SEC Press Release 2020-211, Dec. 9, 2020, "SEC Adopts Rules to Modernize Key Market Infrastructure Responsible for Collecting, Consolidating, and Disseminating Market Data".
[23] In addition to Berkshire Hathaway, the following five companies have elected to use a round lot definition of less than 100 shares for trades in their common equity: Seaboard Corporation, Biglari Holdings Inc., AMCON Distributing Co., Markel Corporation, and NVR, Inc.

We use a counter-factual approach to investigate how these rule changes would have affected the incidence of odd lot quotes over our January – March 2021 sample period. To be sure, the quotes we observe during the sample period reflect traders' behavior under the current regulatory regime, and this behavior might have been different had the new round lot rules applied. Nonetheless, imposing the new rules on the quotes we observe during the sample period allows us an opportunity to examine how much the new rules would reduce odd lot quotes in an ideal environment where quoting behavior was orthogonal to the definition of a round lot.

To conduct our counter-factual analysis, we first calculate what would have been the top of book quote for each venue under the new rules. The SEC's new framework calculates the best bid and offer prices available at an exchange as the least aggressive prices that would satisfy all displayed orders in an amount equal to a round lot. To see how this works, we present a simple example in Table 9 for Nasdaq. In the table we use the quote data on Nasdaq for AMZN at 13:25:00 on March 31, 2021, which we previously discussed in Figure 3. Under the SEC's proposal, the new round lot for AMZN on this day would have been 10 shares because its average stock price for the prior month was between $1000 and $10,000. The order book for AMZN is composed of limit orders wishing to buy (the bid side) and limit orders wishing to sell (the ask side).[24] Looking at the first three columns of Table 9, we find bids at Nasdaq for varying prices between $3112.59 down to $3110.67. Note that the bid at $3110.67 is for 100 shares and that every other bid is for a smaller amount. Under the current rules, the best bid (the top of book) for Nasdaq would be priced at the 100-share order at $3110.67. Under the new rules, the best bid can be determined by walking down the book and aggregating the orders until they constitute at least the new round lot level of 10. Starting at $3112.59, we reach this level at $3111.81. The top of book for Nasdaq would thus be a bid price of $3111.81 with a depth of 23 shares, as shown in columns 4 and 5.

Given these SEC changes, the "inside market" will be reduced, but with only 4 round lot categories the effects will differ depending upon a stock's price. To determine this, we calculated how this approach would have affected the incidence of better odd lot prices for the 1751 sample stocks using data from January – March 2021. Under the new rules, a stock's round lot definition will vary based on its prior month's stock price; therefore, we assign stocks

[24] Recall that market orders are matched with the best orders on the book, so a market order to buy is matched with the lowest ask and a market order to sell is matched against the highest bid.

to a round lot category on a month-by-month basis. For each of the three months, an average of 1636 stocks retained the 100-share round lot level, 107 would have a 40-share round lot, and 8 would have a 10-share round lot (no stock in our sample had a 1 stock round lot). For each stock, we next calculate each exchange's BBO using the new round definitions for each stock, and we calculate the new NBBO by taking the best BBO from across all exchanges. Lastly, we evaluate all end of minute quotes as to whether an odd lot bid (ask) is superior to this new NBBO. The incidence of better odd lot pricing is the average percent of these end of minute observations for each stock.

Table 10 provides results on the incidence of better odd lot pricing relative to the NBBO with the new definitions and the old. As expected, there is no change for the stocks in Panel A which faced no change in the odd lot definition. This is not the case in Panels B (40 share round lot) where the incidence of better odd lot pricing falls an average of 4.8% for bids and 4.9% for offers or in Panel C where it falls 21.7% for bids and 22.6% for offers. By this simple metric, the new policy has reduced the inside market. However, in both categories the incidence of better odd lot prices remains substantial. For the 40-share category, 35.1% (33.2%) of the time there exists a better bid (ask) than the NBBO price; the 10-share category has a better bid 38.8% of the time and a better ask 37% of the time. Figure 16 plots the full distribution of the change in rates across all stocks in the 40-share category and in the 10-share category. The SEC's approach of redefining what constitutes a round lot appears to make headway on the inside market issue, but a substantial problem remains.

6. Conclusions

What determines the "best" trading price available in the market, how limit orders for small quantities fit into the quote montage, and who gets to see this information are all issues that greatly influence the efficiency and fairness of the U.S. equity markets. We find that current market practices relating to odd lot quotes result in a large "inside" market where for many stocks better prices routinely exist relative to the NBBO. We provide strong evidence that being able to see these odd lot quotes provides valuable information to traders with access to proprietary data feeds. We also show that the SEC's approach of changing the definition of a round lot reduces, but does not ameliorate, the high incidence of superior odd lot quotes within the NBBO.

What then to do about odd lot quotes? One approach is to simply make odd lot quotation data cheaper to acquire. This is effectively the approach taken in Reg NMS II, which seeks to promote a competitive market for exchange data. In particular, odd lot quotes that remain despite the new round lot definitions will be part of "core" data that exchanges must distribute to "competing consolidators." As envisioned by Reg NMS II, these new market participants will compete in the offering of market data products, such as by providing a lower-priced product that resembles the current SIP quote data (e.g., a product that provides simply the exchange BBOs) as well as more expensive products that include more elaborate quote data, such as odd lots and depth of book data. In this regard, Reg NMS II may represent a continuation of a "market inside a market", particularly for higher priced stocks where our results show that odd lots quotes will continue to be substantial.

To the extent a policy concern is minimizing the information asymmetries created by the market inside a market, our results would thus point to alternative interventions that would ensure that odd lot information is incorporated into any quote data product. For instance, the Nasdaq market has advocated for changing the tick size from the current one-penny "one size fits all" approach to a variable tick size that differs with the share price. This larger tick would effectively reduce the number of odd lot quotes by shrinking the possible quote price points. Mackintosh [2021] argues, for example, that a 4-cent minimum tick would be the appropriate level for Amazon. While this may prove a useful construct, the recent SEC tick size pilot (which was generally viewed as a failure) suggests that varying tick levels across stocks can be problematic. If the policy objective is to level the playing field, our analysis suggests a different approach, one that involves refining the round lot definitions by relying on machine learning. In particular, it would seem that the optimal round lot is one in which knowing the remaining odd lot data provides no incremental value to market participants. We hope to investigate this approach in future research.

Perhaps the biggest puzzle is why firms do not solve this problem themselves by splitting their stocks to lower price levels. The era of ultra-high stock prices is a new phenomenon. While Berkshire-Hathaway's stock crossed the $1000 level in 1983, only 3 other stocks were above that level in 2013 (Google, Priceline.com, and Seaboard).[25] Weld, Michaely, Thaler, and Benartzi [2009] found that since the Great Depression the average price for a share of stock on the NYSE remained roughly constant at $35.00.[26] In our sample, the median stock price is $42.05 but the average price is $253.71, reflecting the long tail of very high-priced stocks. A variety of changes in the market such as decimal trading, smaller ticks, and the reduction of commissions have dramatically reduced transaction costs and so may have changed the "optimal" price level, but our results suggest these changes have not made the price level irrelevant. Suggestions that fractional share trading solves the high stock price problem for retail traders, or that most stocks are held indirectly in mutual funds and ETFs managed by institutional institutions who do not care about high price levels miss an important point: high priced stocks provide profit opportunities for high frequency and other sophisticated traders who know, and profit from, the inside market. A return to lower stock price levels would alleviate this problem.

[25] For discussion see https://www.kiplinger.com/article/investing/t052-c008-s002-welcome-to-the-1-000-stock-club.html See also Conroy and Harris [1999] who examine how companies traditionally used splits to keep their stock prices within a preferred range.
[26] These authors find little support for a variety of explanations for this constancy (marketability in making the stock easier for investors to buy, inducing brokers to provide liquidity, signaling) and ultimately conclude that it is due to market norms.

References

Bartlett, R. P., 2021, Modernizing Odd Lot Trading, SSRN: https://ssrn.com/abstract=3849763

Bartlett, R, P., and J. McCrary, 2019. "How rigged are stock markets? Evidence from microsecond timestamps," *Journal of Financial Markets*, 45(C), pages 37-60.

Bartlett, R, P., and J. McCrary, 2020. Subsidizing Liquidity with Wider Ticks: Evidence from the Tick Size Pilot Study, *Journal of Empirical Legal Studies*, 17, 262-316.

Battalio, R., S. Corwin, and R. Jennings, 2016, Can Brokers Have it All? The relation between maker-taker and limit order execution quality, *Journal of Finance* (71).

Brogaard, J., B. Hagstromer, L. Norden, and R. Riordan 2015, Trading fast and slow: Colocation and liquidity. *Review of Financial Studies*, 28(12): 3407-3443.

Buti, S., Rindi, B., I. Werner, and Y. Wen, 2019, Tick Size, Trading Strategies and Market Quality, Available at SSRN: https://ssrn.com/abstract=3330807

Chao, Y., Y. Yao, and M. Ye, 2019, Why discrete price fragments U.S. stock exchanges and disperses their fee structures, *Review of Financial Studies*, 32(2), 1018 – 1034

Chen, T. and C. Guestrin, 2016, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pages 785–794 https://doi.org/10.1145/2939672.2939785

Chinco, A., A. Clark-Joseph, and M. Ye. 2018. Sparse signals in the cross-section of returns. *Journal of Finance* 74:449-492.

Cimon, D, 2021, Broker Routing Decisions in Limit Order Markets, *Journal of Financial Markets*, (54)

Conroy, R. and R. Harris, 1999, Stock Splits and Information: The role of share price, *Financial Management*, Vol. 28, No. 3, pp. 28-40

Easley, D., M. Lopez de Prado, M. O'Hara, and Z. Zhang, 2021, Microstructure in the Machine Age, *Review of Financial Studies*. 34 (7) 3316-3363

Eril, I., L Stearn, C. Tan, and M. Weisbach, 2021, Selecting Directors Using Machine Learning, *Review of Financial Studies*, 34 (7) 3226-3264.

Goldstein, I., C. Spatt, and M. Ye, 2021, Big Data in Finance, *Review of Financial Studies,* 34 (7) 3216-325.

S. Gu., B. Kelly, and D. Xiu. 2020. Empirical asset pricing via machine learning. *Review of Financial Studies*.

Lundberg, S. and S-I, Lee, 2017, A Unified Approach to Interpreting Model Predictions, NIPs Conference Proceedings.

Mackintosh, P., 2019, What happens in an odd lot world? Available at https://www.nasdaq.com/articles/what-happens-in-an-odd-lot-world-2019-06-20

Mackintosh, P., 2021, A more intelligent (tick) plan for odd lots, available at https://www.nasdaq.com/articles/a-more-intelligent-tick-plan-for-odd-lots-2021-05-13

Mullainathan, S., and Spiess J. 2017. Machine learning: An applied econometric approach. *Journal of Economic Perspectives* 31:87–106

O'Hara, M., G. Saar, and Z. Zhou, 2020, Relative Tick Size and the Trading Environment, *Review of Asset Pricing Studies*,

O'Hara, M., C Yao, and M. Ye, 2014, What's Not There: Odd Lots and Market Data," *Journal of Finance*, LXIX (6).

Philip, R. 2020. Estimating permanent price impact via machine learning. *Journal of Econometrics* 215:414-449.

Rossi, A. G. and S. P. Utkus, 2020, Who Benefits from Robo-advising? Evidence from Machine Learning. Available at SSRN: https://ssrn.com/abstract=3552671

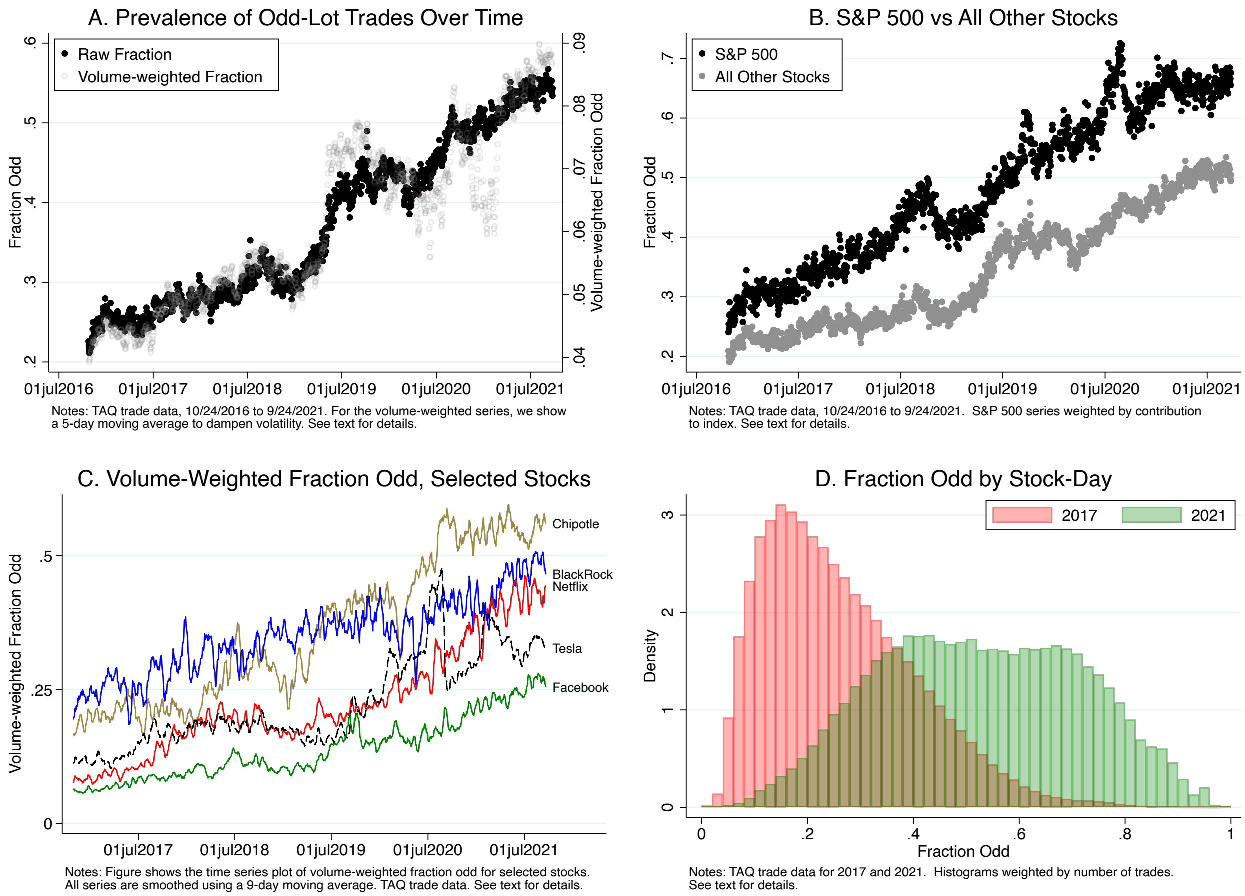


Figure 1: Odd Lot Trades in the TAQ Data, 2016-2021. Panel A presents a time series plot of the fraction of trades that are odd, both in terms of transactions (black dots, left axis) and in terms of shares (gray dots, right axis.) Panel B shows the same trend line for S&P500 stocks and non-S&P500 stocks, while Panel C shows the trend line for select high priced stocks. Panel D presents histograms showing the stock-day distribution of the fraction of trades that are odd in 2017 relative to 2020.

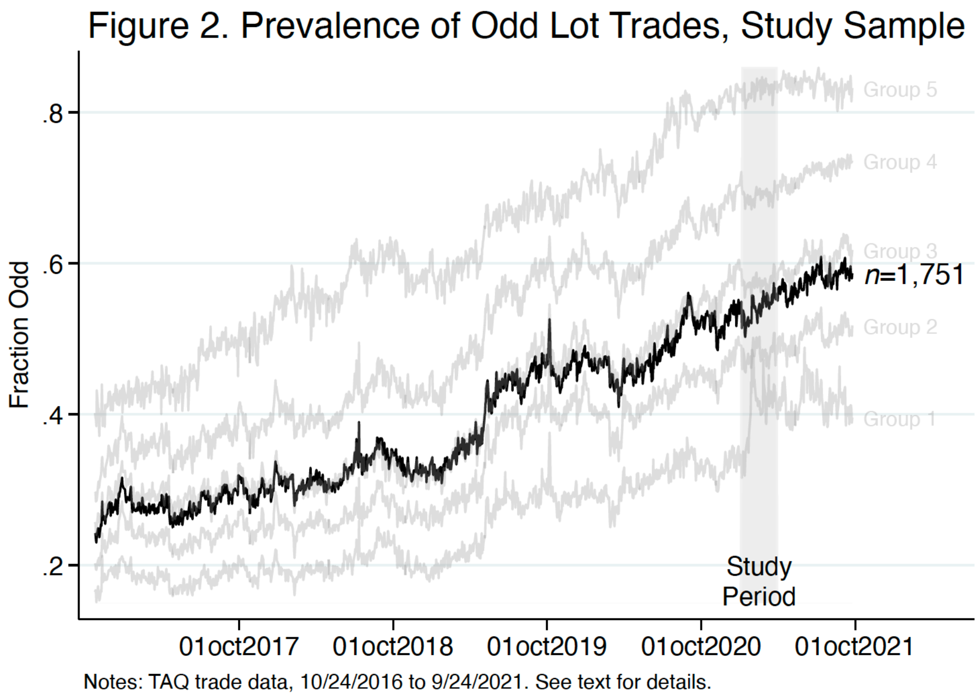


Figure 2: Prevalence of Odd Lot Trades in Study Stocks. Figure presents a time series plot of the fraction of trades that are odd for the study stocks.

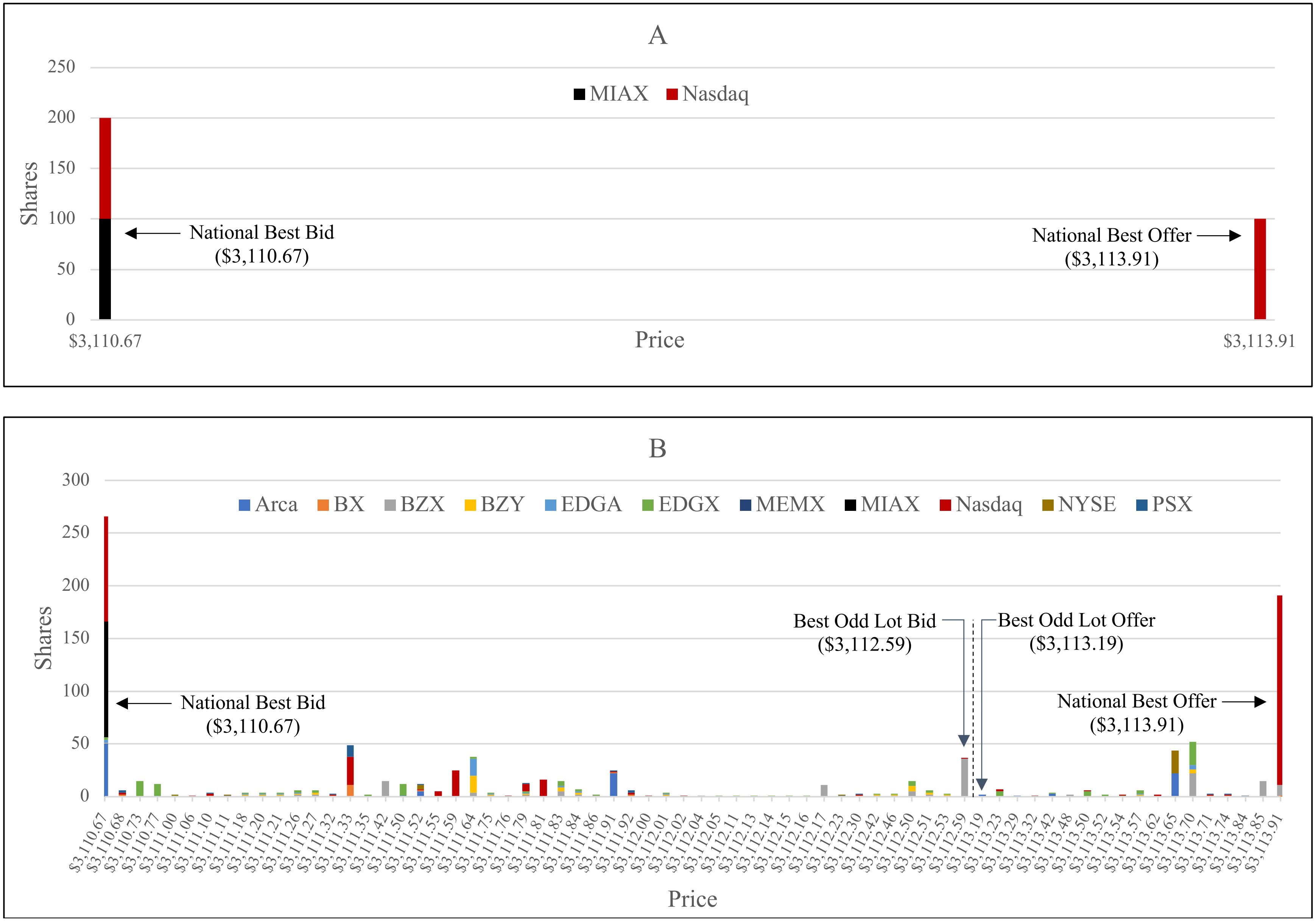


Figure 3: TAQ Order Book at NBBO vs. May Street Order Book at NBBO. Panel A shows the NBBO for AMZN at 13:25:00 on 3/31/21, as observed using TAQ Data. Panel B shows the NBBO and all better-priced odd lot quotes for AMZN at 13:25:00 on 3/31/21, as observed using May Street Data.

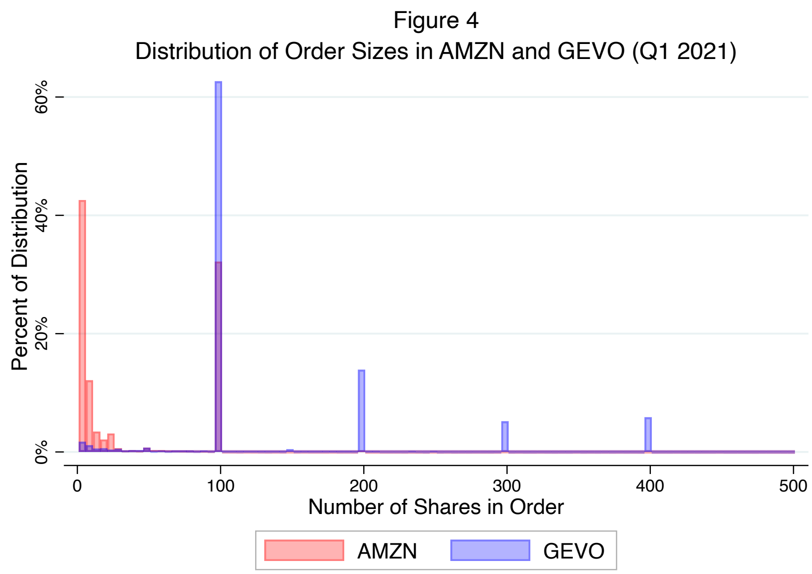


Figure 4: Distribution of Order Sizes in AMZN and GEVO (Q1 2021). Data represent all messages to add liquidity across 16 stock exchanges for AMZN and GEVO between 1/4/21 and 3/31/21 for order sizes that were less than 500 shares. The histogram covers 57,523,401 messages for AMZN and [9,706,139] messages for GEVO.

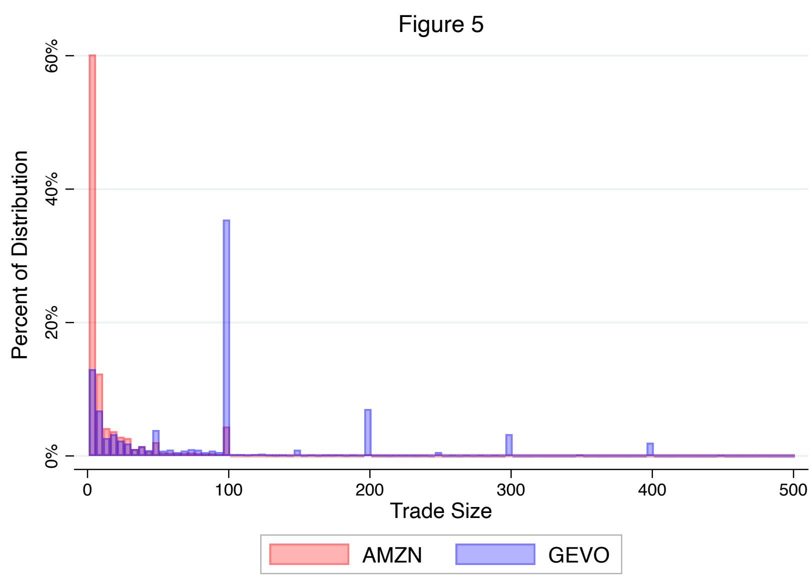


Figure 5: Distribution of Trade Sizes in AMZN and GEVO. Distribution of trades by size for Q1 2021 for AMZN and GEVO, excluding trades for 500 shares or more. The figure includes 10,777,527 trades in AMZN and 6,351,743 trades in GEVO.

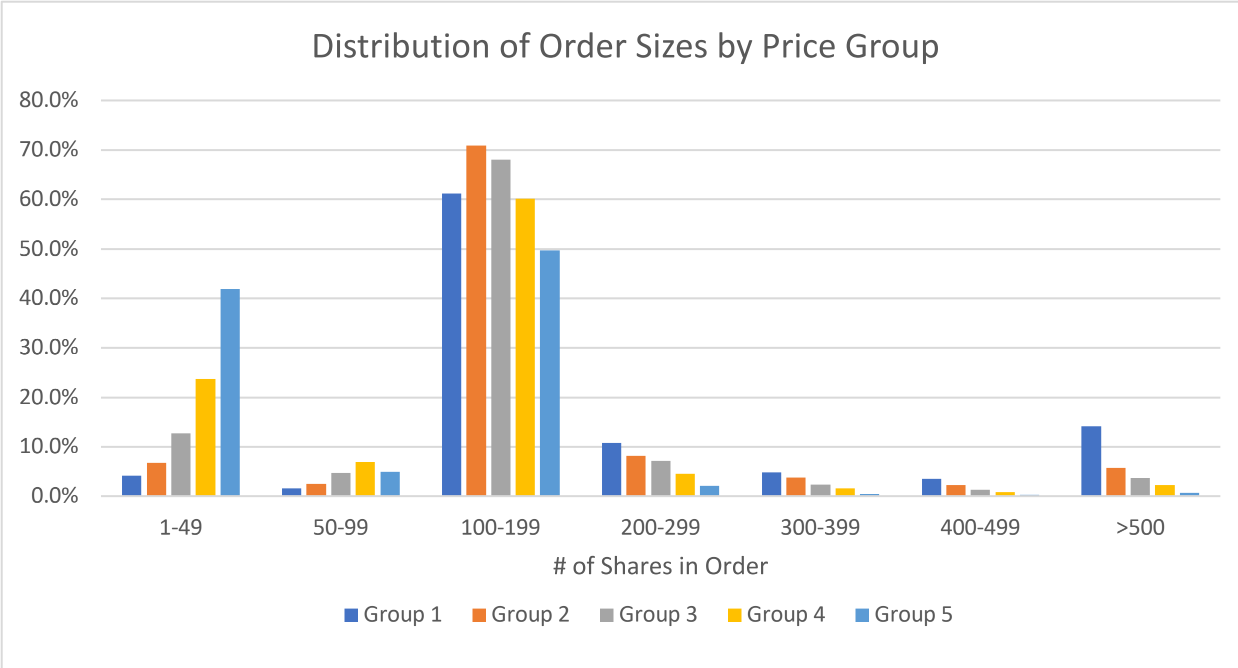


Figure 6: Distribution of Order Sizes by Price Group (Q1 2021). Data represent all messages to add liquidity across 16 stock exchanges for all sample stocks between 1/4/21 and 3/31/21 for order sizes that were less than 500 shares.

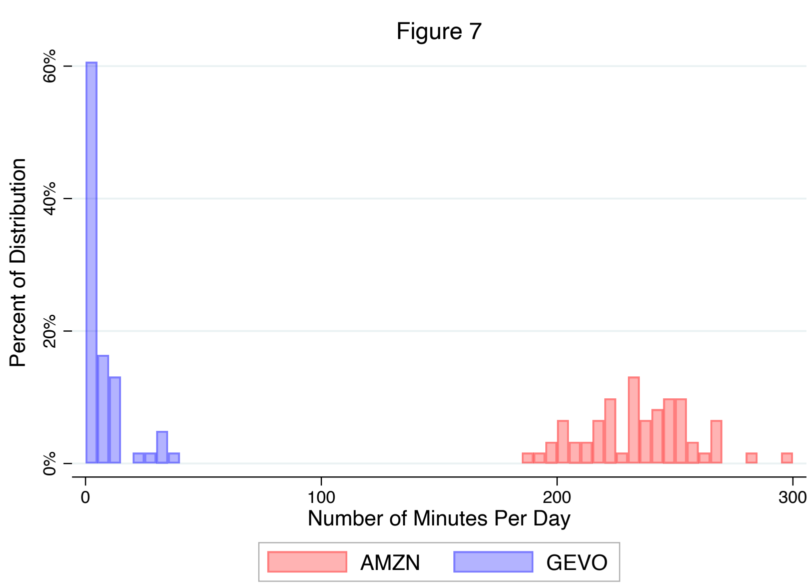


Figure 7: Minutes Per Day Where OL Bid>SIP Bid in AMZN and GEVO. Distribution of the number of minutes for each day in the sample where the OL Bid was better than the SIP bid for AMZN and GEVO. The book was evaluated at the end of each minute of the trading day, so there are a total of 390 observed minutes per day.

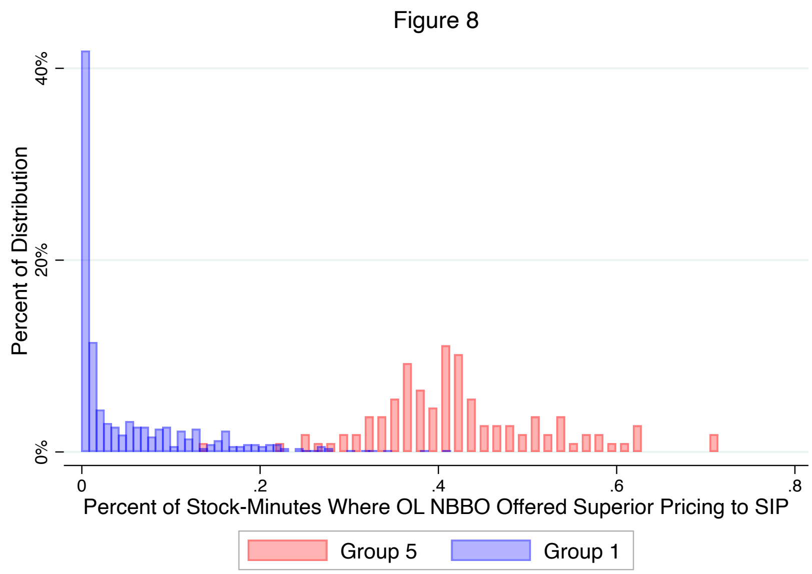


Figure 8: Rate at which Odd Lot NBBO Offered Superior Pricing - Group 1 vs. 5.
Distribution of stock-minutes where the price of the best displayed odd lot bid was higher than the price of the SIP bid across all stock-minutes for Group 1 and Group 5 stocks

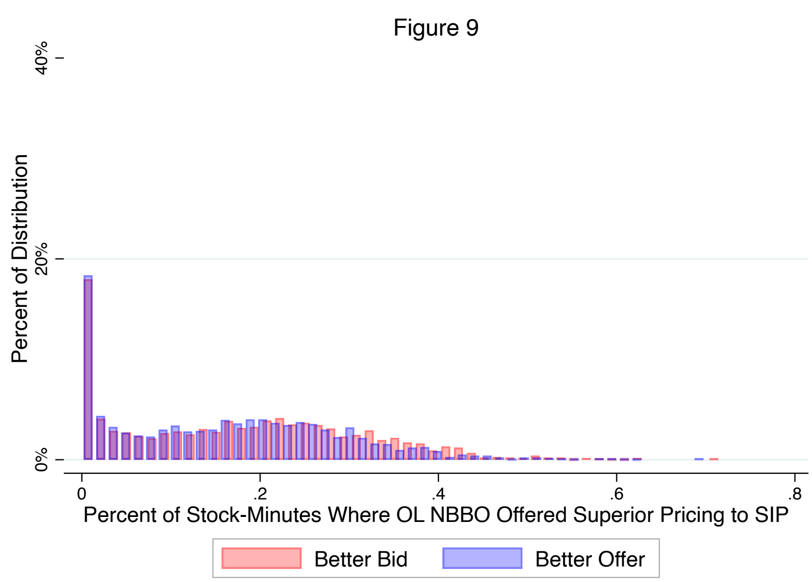
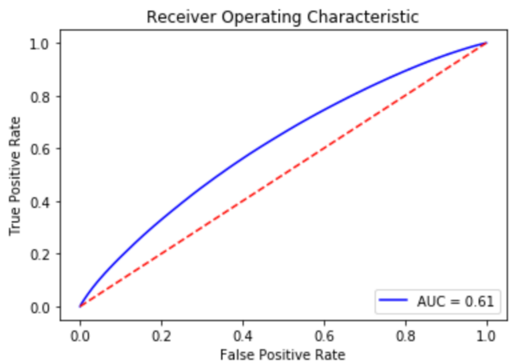
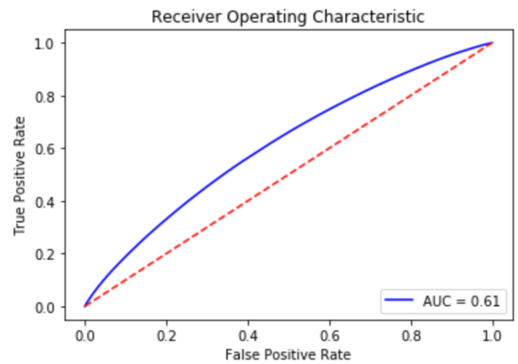


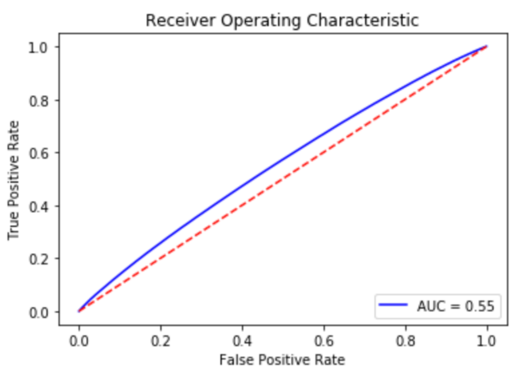
Figure 9: Rate at which Odd Lot NBBO Offered Superior Pricing – Full Sample.
Distribution of stock-minutes where the price of the best displayed odd lot bid was higher than the price of the SIP bid across all stock-minutes in the sample



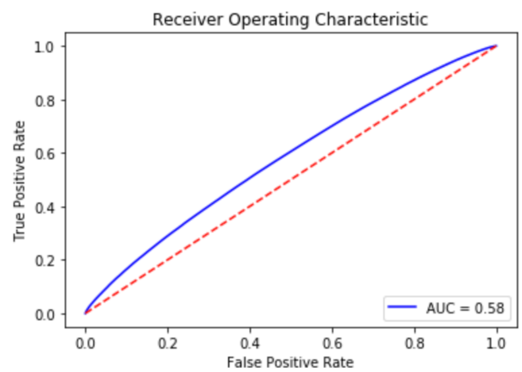
A. SIP Model (Group 1)



B. OL Model (Group 1)



C. SIP Model (Group 5)



D. OL Model (Group 5)

Figure 10: Receiver operating characteristic (ROC) curves of SIP Model vs OL Model. Area under the ROC curve showing the predictive performance of the SIP Model versus the OL Model using the test data for two SIP Models and two OL Models. Panel A presents results for the SIP Model using Group 1 stocks. Panel B presents results for the OL Model using Group 1 stocks. Panel C presents results for the SIP Model using Group 5 stocks. Panel D presents results for the OL Model using Group 5 stocks.

Figure 11: SHAP Values for Chipotle OL Model Prediction. Illustration of the SHAP Values for all features in the OL Model for the model's prediction for Chipotle at the beginning of 09:34:00 on January 8, 2020.

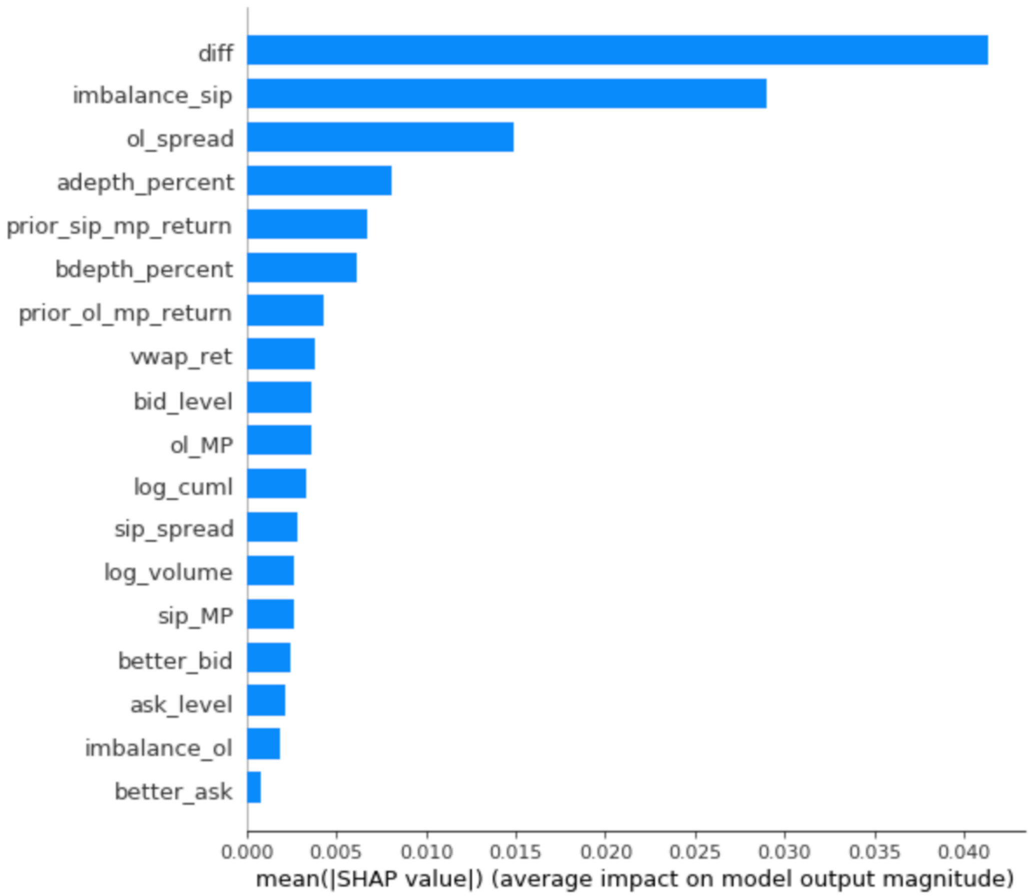


Figure 12: Average SHAP Values for Features in the Group 5 Training Data. The figure summarizes the mean of the absolute value of all SHAP values calculated across 1,916,000 observations in the training data for Group 5 stocks.

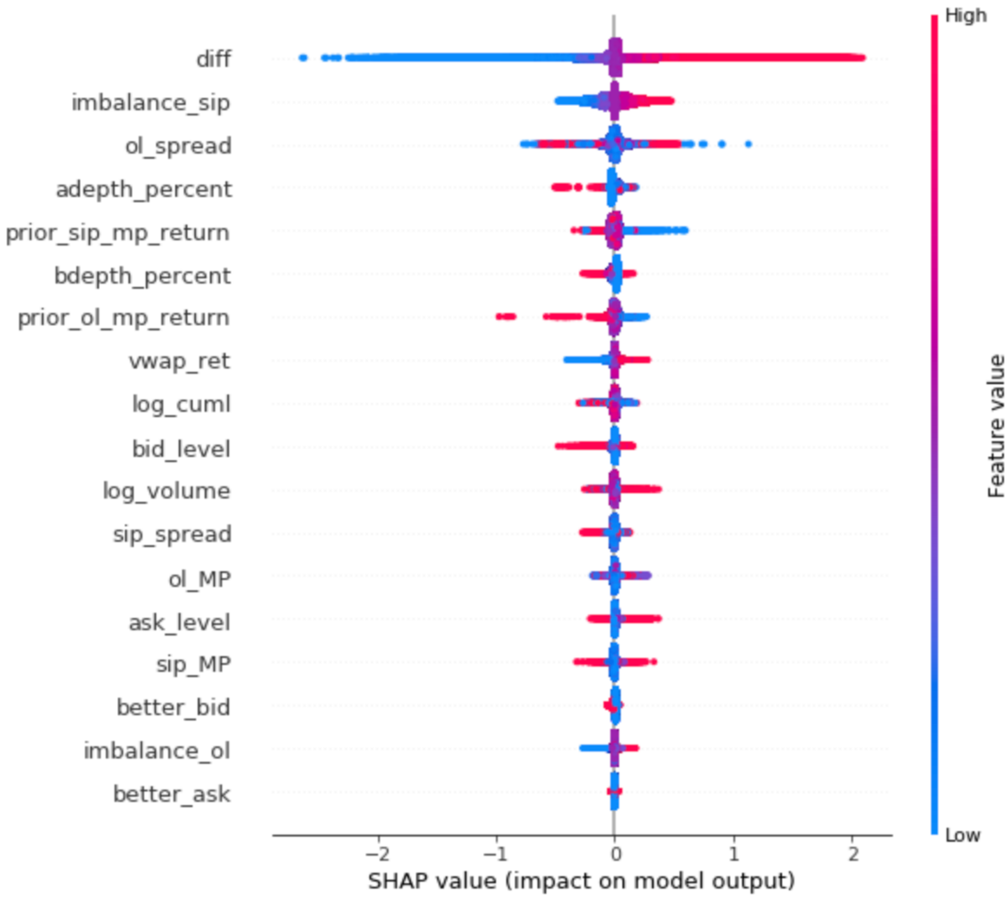
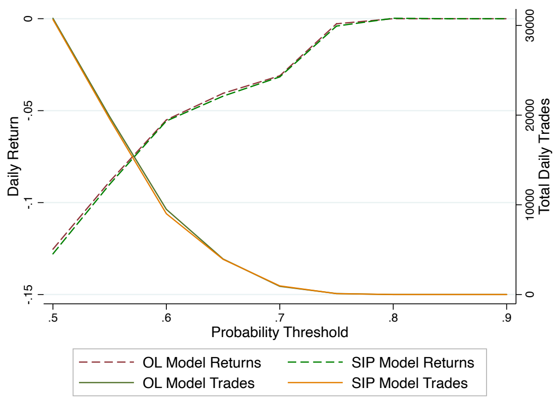
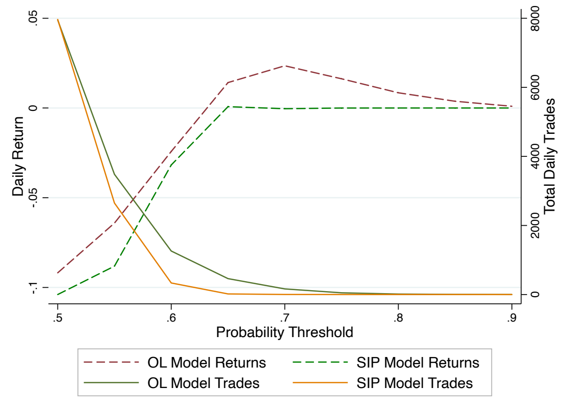


Fig. 13: Distribution of SHAP Values for All Features of Group 5 Training Data – OL Model. The plot depicts the relative importance, impact, and contribution of different features in the Group 5 training data for the OL model. Positive SHAP values indicate a positive classification. The summary plot combines feature importance with feature effects. The features on the y-axis are ordered according to their importance (top=high importance). Each point on the summary plot is a SHAP value for a feature and an instance (i.e., a single stock-day-minute) in the training dataset. The position of each point on the x-axis shows the impact that the feature has on the classification model's prediction for a given instance. The color represents the high (red) to low (blue) values of the feature.



A



B

Fig. 14: Daily Returns for Group 1 vs. Group 5 Stocks. Average daily returns from a trading strategy that uses the SIP Model and OL Model to execute a round trip trade that begins with a purchase at the commencement of a stock-minute and ends with a sale at $VWAP_{t+1}$. Panel A presents returns for Group 1 stocks. Panel B presents returns for Group 5 stocks.

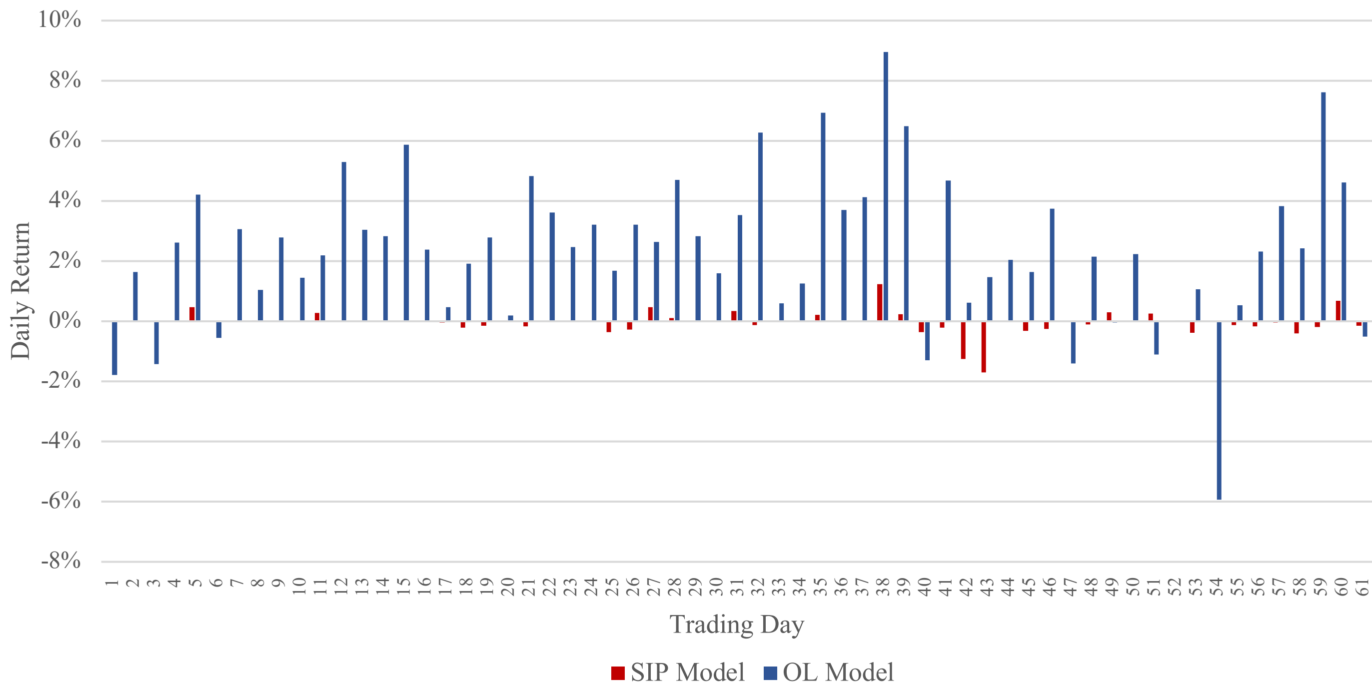
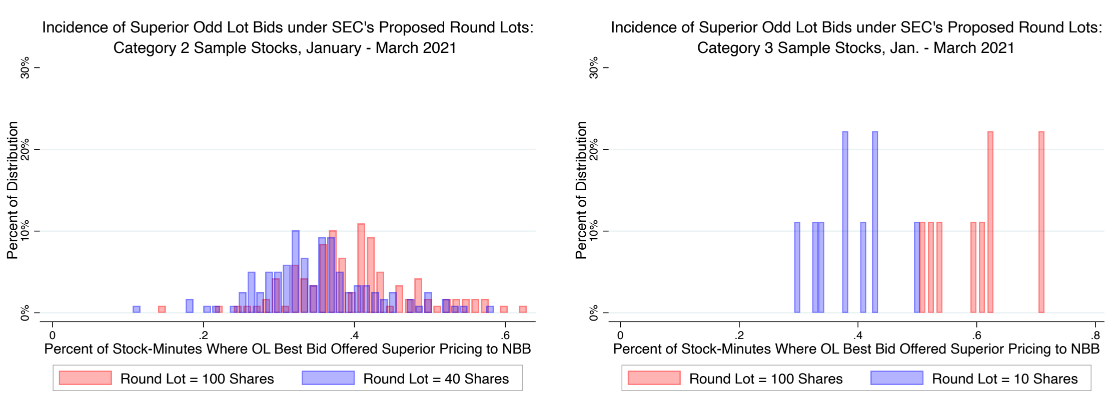


Fig. 15: Distribution of Daily Returns with a 70% Probability Threshold: SIP Model vs. OL Model. The distribution of daily returns across the sixty-one trading days for the SIP Model compared to the OL Model when the models use a probability threshold of 70% to predict a positive label.



A B

Figure 16: **Effect on Sample Securities of Proposed Round Lot Definitions.** Panel A presents a histogram showing the average percentage of minutes with superior odd lot bids (relative to the NBBO) for each stock in proposed category 2 (round lot=40 shares) assuming the proposed definition applied during the sample period, as compared to the results under the current round lot definition. Panel B presents the same figure for each stock in proposed category 3 (round lot=10 shares).

Table 1: Summary Statistics

This table presents summary statistics for the sample securities. *VWAP* is calcualted for each stock as the mean of the stock's volume-weighted average daily price during December 2020. All the other variables are based on end-of-minute (EOM) quotation data from May Street (Panel A) and trade data from TAQ (Panel B) during the sample period (January 1, 2021 - March 31, 2021). Other than VWAP, all variables are averaged over the day for each stock in the sample, and the reported statistics are the mean of all stock-days during the sample period for five, price-based groups of stocks: Group 1 (VWAP <= $20), Group 2 (VWAP > $20 & <=$50), Group 3 (VWAP >$50 & <=$100), Group 4 (VWAP>$100 & <=$250) and Group 5 (VWAP>$250). SIP Spread is the difference between the national best offer and the national best bid reported by the SIPs. OL Spread is the difference between the national best offer and the national best bid reported across all exchanges using May Street's data. Depth at Top of SIP Bid (Offer) is the dollar value of depth at national best bid (offer) reported by the SIPs. Depth at Top of OL Bid (Offer) is the dollar value of depth at the national best bid (offer) for any EOM observation where the national best bid (offer) reported by Maystreet is an odd lot that is superior to the national best bid (offer) reported by the SIPs. Depth at Top of SIP Bid (Offer) is the dollar value of depth at national best bid (offer) reported by the SIPs. Total Depth at OL Bids (Offer) is the aggregate dollar value of all bids (offers) having superior prices to the SIP Bid (Offer).

	Group 1	Group 2	Group 3	Group 4	Group 5	Full Sample
N	504	478	361	299	109	1,751
VWAP						
mean	10.16	33.74	72.67	155.98	507.06	85.32
median	10.47	33.12	71.48	145.44	353.32	42.05
SIP Spread						
mean	0.0246	0.0663	0.1343	0.2478	0.9183	0.1537
median	0.0115	0.0490	0.1081	0.1852	0.6113	0.0617
OL Spread						
mean	0.0217	0.0576	0.1166	0.2050	0.6769	0.1241
median	0.0109	0.0428	0.0925	0.1527	0.4789	0.0531
Depth at Top of SIP Bid ($)						
mean	66,974.12	40,029.85	38,852.74	56,866.39	136,238.41	56,393.49
median	23,641.01	22,076.35	31,562.15	51,036.11	101,759.48	32,631.26
Depth at Top of SIP Offer ($)						
mean	67,297.23	40,178.10	38,909.19	56,698.01	134,191.54	56,378.92
median	23,937.03	22,424.64	32,219.62	51,652.56	102,258.43	33,070.40
Depth at Top of OL Bid ($)						
mean	2,730.67	3,463.79	2,864.76	3,638.54	8,578.40	3,521.50
median	417.10	938.80	1,684.55	3,134.01	6,863.56	1,366.53
Depth at Top of OL Offer ($)						
mean	2,541.12	3,455.74	2,855.15	3,820.76	8,999.64	3,537.69
median	454.88	1,010.83	1,770.36	3,308.28	7,147.61	1,464.17
Total Depth at OL Bids ($)						
mean	7,186.55	8,350.30	5,357.19	6,379.33	23,147.68	8,035.70
median	499.75	1,208.22	2,419.09	5,335.48	14,089.95	2,031.33
Total Depth at OL Offers ($)						
mean	6,569.37	7,134.90	5,109.81	6,305.87	23,138.50	7,473.33
median	526.66	1,258.72	2,447.76	5,346.21	14,269.21	2,082.96

Table 2: Distribution of Order Sizes for All Orders Submitted Across Exchanges
This table presents statistics for the distribution of order sizes for 32.3 billion orders submitted for sample securities during the sample period. Panel A presents the distribution of order sizes across the 5 price-based group of stocks. Panel B presents the distribution of order sizes across exchanges. Note: Panel B includes data for only 15 of the 16 stock exchanges in operation because the Long Term Stock Exchange did not process any messages during the sample period.

Panel A: Distribution of Order Sizes by Group

	No. of Shares in Order						
Group	1-49	50-99	100-199	200-299	300-399	400-499	>500
1	4.1%	1.5%	61.2%	10.7%	4.8%	3.5%	14.1%
2	6.8%	2.5%	70.9%	8.2%	3.8%	2.3%	5.7%
3	12.7%	4.7%	68.1%	7.2%	2.4%	1.4%	3.6%
4	23.7%	6.9%	60.2%	4.5%	1.6%	0.9%	2.3%
5	41.9%	5.0%	49.7%	2.1%	0.5%	0.2%	0.6%
All	12.4%	3.6%	64.4%	7.7%	3.2%	2.1%	6.7%

Panel B: Distribution of Order Sizes Across Exchanges for Group 5 Stocks

	No. of Shares in Order						
Exchange	1-49	50-99	100-199	200-299	300-399	400-499	>500
AMEX	5%	3%	42%	16%	8%	5%	22%
ARCA	10%	3%	67%	9%	3%	2%	5%
BATS	11%	3%	67%	8%	3%	2%	5%
BATSY	14%	2%	76%	2%	1%	1%	2%
BX	35%	2%	49%	5%	2%	1%	5%
CHX	1%	0%	80%	10%	7%	0%	2%
EDGA	16%	2%	79%	1%	0%	0%	1%
EDGX	19%	4%	52%	9%	3%	3%	10%
IEXDEEP	0%	0%	44%	21%	14%	8%	13%
NASDAQ	17%	5%	62%	6%	3%	2%	5%
MEMX	2%	0%	88%	4%	2%	1%	3%
MIAX	10%	14%	75%	0%	0%	0%	0%
NSX	5%	2%	87%	2%	1%	0%	2%
NYSE	10%	3%	64%	8%	3%	2%	10%
PSX	10%	1%	60%	11%	4%	3%	11%
All	9.7%	3.3%	64.0%	7.9%	3.3%	2.3%	9.5%

Table 3: Incidence of Superior Odd Lot Prices at the NBBO

This table presents statistics for the rate that the National Best Bid and the National Best Offer calculated using odd lot quotes (the OL NBBO) offered superior pricing to that displayed by the SIPs (the SIP NBBO) for each stock within the sample. *Better Bid* is the average percent of end-of-minute (EOM) observerations for a stock when the OL best bid was greater than the SIP best bid. *Better Offer* is the average percent of EOM observerations for a stock when the OL best offer was less than the SIP best offer. Reported statistics are the mean across all stock-days within each price-based group of stocks. Panel A reports statistics for the best bid; Panel B reports statistics for the best offer.

Panel A: Bids

Group	N	mean	sd	p25	p50	p75
1	504	0.0516	0.0803	0.0026	0.0103	0.0718
2	478	0.1414	0.1098	0.0513	0.1282	0.2051
3	361	0.2236	0.1198	0.1385	0.2128	0.2974
4	299	0.2955	0.1139	0.2128	0.2872	0.3692
5	109	0.4201	0.1343	0.3256	0.4103	0.5051
All	1,751	0.1771	0.1513	0.0359	0.1564	0.2769

Panel B: Offers

Group	N	mean	sd	p25	p50	p75
1	504	0.0429	0.0674	0.0026	0.0103	0.0590
2	478	0.1263	0.0979	0.0462	0.1128	0.1846
3	361	0.2033	0.1069	0.1282	0.1949	0.2692
4	299	0.2789	0.1055	0.2026	0.2692	0.3462
5	109	0.4056	0.1323	0.3077	0.3923	0.4897
All	1,751	0.1625	0.1419	0.0308	0.1410	0.2538

Table 4: Predictors of Superior Odd Lot Prices

This table reports regression estimates of stock-level predictors for when the OL NBBO improves on the SIP NBBO. The unit of observation is the stock-day end-of-minunte (EOM) for all stocks in the sample, and estimates are for a linear probability model that the NBB or NBO calculated using odd lots was superior to the NBB or NBO reported by the SIP. *Log_price* is the natural log of the value-weighted average price for the minute. *Log_spread* is the natural log of the quoted spread of the SIP NBBO for the minute. *Group* is an indicator variable for the the sized-based group assigned to the stock. *Log_volume* is the natural log of the volume of shares transacted witin the minute. *Return* is the return for the minute. *Volatility* is the standard deviation of Return across all EOM observations in a day. Stock-clustered standard errors are in brackets. *,**, and *** indicate significance levels of 10%, 5%, and 1%.

	(1)	(2)	(3)	(4)
DV	Better Bid	Better Offer	Better Bid	Better Offer
log_spread	0.0958***	0.0843***	0.0852***	0.0715***
	[0.00170]	[0.00150]	[0.00578]	[0.00484]
Group 2 X log_spread			0.0199***	0.0200***
			[0.00595]	[0.00499]
Group 3 X log_spread			0.00801	0.0120**
			[0.00615]	[0.00525]
Group 4 X log_spread			0.00589	0.0111**
			[0.00630]	[0.00540]
Group 5 X log_spread			0.0279***	0.0328***
			[0.00753]	[0.00668]
log_price	-0.0109*	-0.0183***	-0.0239***	-0.0169**
	[0.00649]	[0.00549]	[0.00805]	[0.00700]
Group 2 X log_price			0.0420***	0.0176*
			[0.0106]	[0.00907]
Group 3 X log_price			0.0380**	-0.0136
			[0.0160]	[0.0121]
Group 4 X log_price			0.0232	0.0071
			[0.0184]	[0.0150]
Group 5 X log_price			0.0754**	0.0164
			[0.0370]	[0.0259]
log_volume	-0.000215	-0.000762***	0.000254	3.99E-05
	[0.000137]	[0.000124]	[0.000195]	[0.000165]
Group 2 X log_volume			-0.00151***	-0.00182***
			[0.000268]	[0.000235]
Group 3 X log_volume			-0.00149***	-0.00190***
			[0.000307]	[0.000285]
Group 4 X log_volume			0.00117***	0.000107
			[0.000385]	[0.000338]
Group 5 X log_volume			0.00391***	0.00339***
			[0.000674]	[0.000625]
volatility	-0.135***	0.104***	0.0840***	0.156***
	[0.0296]	[0.0259]	[0.0260]	[0.0290]
Group 2 X log_volatility			-0.450***	-0.140**
			[0.0772]	[0.0640]
Group 3 X log_volatility			-0.895***	-0.203
			[0.0982]	[0.126]
Group 4 X log_volatility			-1.160***	-0.315**
			[0.193]	[0.137]
Group 5 X log_volatility			-1.185***	-0.147
			[0.240]	[0.268]
return	-0.136***	0.156***	-0.0420***	0.0264***
	[0.00740]	[0.00955]	[0.00624]	[0.00598]
Group 2 X log_return			-0.0816***	0.133***
			[0.0146]	[0.0142]
Group 3 X log_return			-0.192***	0.297***
			[0.0229]	[0.0229]
Group 4 X log_return			-0.510***	0.644***
			[0.0366]	[0.0425]
Group 5 X log_return			-0.726***	0.960***
			[0.0760]	[0.0762]
Constant	0.526***	0.496***	0.447***	0.464***
	[0.0222]	[0.0190]	[0.0265]	[0.0203]
Date FE	Y	Y	Y	Y
Stock FE	Y	Y	Y	Y
Observations	40,386,321	40,386,321	40,386,321	40,386,321
R-squared	0.148	0.136	0.148	0.136
# of Stocks	1736	1736	1736	1736

Table 5: Available Depth Inside the SIP NBBO

This table reports statistics for odd lot bid depth (OL Bid Depth) or odd lot Offer Depth (OL Offer Depth), defined as the cumulative depth at either the bid or the offer at odd lot prices that are inside the SIP Bid or SIP Offer, divided by the depth at the SIP Bid or SIP Offer. Panel A presents OL Bid Depth and OL Offer Depth across all observed end-of-minutes (EOM) within the sample. Panel B presents OL Bid Depth and OL Offer Depth across EOMs for which the OL NBBO offered superior pricing to the SIP NBBO.

Panel A: Unconditional Depth Inside the SIP NBBO

	Best Bid:					*Best Offer:*				
Group	mean	sd	p25	p50	p75	mean	sd	p25	p50	p75
1	0.005	0.012	0.000	0.001	0.005	0.005	0.010	0.000	0.001	0.004
2	0.016	0.017	0.004	0.011	0.022	0.014	0.015	0.004	0.011	0.020
3	0.030	0.023	0.014	0.025	0.040	0.027	0.020	0.013	0.023	0.036
4	0.050	0.030	0.028	0.044	0.065	0.047	0.028	0.027	0.041	0.060
5	0.088	0.051	0.053	0.076	0.108	0.084	0.049	0.051	0.073	0.106
All	0.026	0.032	0.003	0.016	0.037	0.024	0.030	0.003	0.015	0.034

Panel B: Depth Inside the SIP NBBO, Conditional on Superior OL Quotes

	Best Bid:					*Best Offer:*				
Group	mean	sd	p25	p50	p75	mean	sd	p25	p50	p75
1	0.161	0.267	0.028	0.072	0.139	0.160	0.266	0.028	0.074	0.140
2	0.124	0.136	0.067	0.100	0.139	0.126	0.135	0.068	0.102	0.142
3	0.130	0.069	0.092	0.123	0.158	0.129	0.066	0.092	0.122	0.157
4	0.161	0.056	0.122	0.155	0.192	0.162	0.055	0.124	0.156	0.193
5	0.200	0.069	0.153	0.190	0.234	0.200	0.069	0.154	0.190	0.233
All	0.146	0.160	0.075	0.118	0.168	0.146	0.158	0.076	0.120	0.169

Table 6: Predictors of Depth Inside the SIP NBBO

This table reports regression estimates of stock-level predictors for odd lot bid depth (OL Bid Depth) or odd lot Offer Depth (OL Offer Depth), defined as the cumulative depth at either the bid or the offer at odd lot prices that are inside the SIP Bid or SIP Offer, divided by the depth at the SIP Bid or SIP Offer. The unit of observation is the stock-day-EOM for all stocks in the sample, and the dependent variable is OL Bid Depth or OL Offer Depth as of the EOM for each stock-day. *Log_price* is the natural log of the value-weighted average price for the minute. *Log_spread* is the natural log of the average quoted spread of the SIP NBBO for the minute. *Group* is an indicator variable for the the sized-based group assigned to the stock. *Log_volume* is the natural log of the volume of shares transacted witin the minute. *Return* is the return for the minute. *Volatility* is the standard deviation of Return across all EOM observations in a day. Stock-clustered standard errors are in brackets. *,**, and *** indicate significance levels of 10%, 5%, and 1%.

	(1)	(2)	(3)	(4)
DV:	OL Bid Depth	OL Offer Depth	OL Bid Depth	OL Offer Depth
log_spread	0.0246***	0.0221***	0.0143***	0.0123***
	[0.000559]	[0.000506]	[0.00107]	[0.000915]
Group 2 X log_spread			0.00640***	0.00615***
			[0.00116]	[0.000988]
Group 3 X log_spread			0.0103***	0.00941***
			[0.00125]	[0.00108]
Group 4 X log_spread			0.0191***	0.0183***
			[0.00148]	[0.00130]
Group 5 X log_spread			0.0435***	0.0418***
			[0.00377]	[0.00337]
log_price	-0.00748***	-0.00785***	-0.0025	-0.00201
	[0.00249]	[0.00209]	[0.00372]	[0.00315]
Group 2 X log_price			0.00302	0.00105
			[0.00393]	[0.00334]
Group 3 X log_price			0.00187	-0.00402
			[0.00537]	[0.00363]
Group 4 X log_price			-0.00938	-0.00588
			[0.00583]	[0.00460]
Group 5 X log_price			-0.00836	-0.0255***
			[0.0109]	[0.00905]
log_volume	0.000998***	0.000910***	0.000414***	0.000387***
	[4.06e-05]	[3.69e-05]	[3.80e-05]	[3.28e-05]
Group 2 X log_volume			0.000442***	0.000426***
			[5.61e-05]	[5.03e-05]
Group 3 X log_volume			0.000756***	0.000684***
			[8.18e-05]	[7.00e-05]
Group 4 X log_volume			0.00142***	0.00123***
			[0.000115]	[0.000102]
Group 5 X log_volume			0.00239***	0.00223***
			[0.000250]	[0.000238]
volatility	-0.0387***	-0.000482	0.0186***	0.0281***
	[0.00667]	[0.00614]	[0.00472]	[0.00515]
Group 2 X log_volatility			-0.0493***	-0.00979
			[0.0158]	[0.0150]
Group 3 X log_volatility			-0.180***	-0.0362
			[0.0234]	[0.0287]
Group 4 X log_volatility			-0.296***	-0.124***
			[0.0494]	[0.0346]
Group 5 X log_volatility			-0.459***	-0.371***
			[0.0895]	[0.0980]
return	-0.0420***	0.0435***	-0.00628***	0.00434**
	[0.00237]	[0.00295]	[0.00150]	[0.00181]
Group 2 X log_return			-0.0255***	0.0346***
			[0.00376]	[0.00358]
Group 3 X log_return			-0.0779***	0.0848***
			[0.00655]	[0.00666]
Group 4 X log_return			-0.186***	0.199***
			[0.0116]	[0.0147]
Group 5 X log_return			-0.297***	0.376***
			[0.0278]	[0.0317]
Constant	0.125***	0.115***	0.101***	0.0988***
	[0.00903]	[0.00767]	[0.00760]	[0.00578]
Observations	40,386,321	40,386,321	40,386,321	40,386,321
R-squared	0.077	0.072	0.08	0.075
# of Stocks	1736	1736	1736	1736

Table 7: Predictors of Future Trade Prices - Regression

This table reports regression estimates of the predictors of intra-day trade prices. The unit of observation is the stock-day-minute for all stocks in Groups 1 and 5. The dependent variable is the difference between the natural log of the volume-weighted average price (VWAP) of trades between the beginning of minute *t* until the commencement of *t+1* . All independent variables are measured at the commencement of *t* . *Prior VWAP Return* is the log difference between the return on VWAP during the minute t-1 and the minute t-2. *Log Volume* is the natural log of the volume of trades observed between *t-1* and *t* . *Cumulative Volume* is the natural log of the cumulative volume of shares traded from 9:30 am through t-1. *SIP Spread* is the quoted spread based on the best bid and offer shown by the SIPs. *OL Spread* is the quoted spread based on the best bid and offer based on exchange data, which includes odd lot quotes. *SIP Midpoint* is the midpoint of the best bid and offer shown by the SIPs. *OL Midpoint* is the midpoint of the best bid and offer based on exchange data, which includes odd lot quotes. *Difference* is the difference between the natural log of OL Midpoint and the natural log of SIP Midpoint. *SIP Ibalance* is the difference between the displayed depth at the SIP bid and SIP offer. *OL Imbalance* is the difference between the displayed depth at the best bid and offer based on exchange data, which includes odd lot quotes. *Prior SIP MP Return* and *Prior OL MP Return* are the log difference between either SIP Midpoint or OL Midpoint from the start of the minute t-1 to the start of the minute t. *OL Bid Depth* and *OL Offer Depth* are the cumulative depth at either the bid or the offer at odd lot prices that are inside the SIP Bid or SIP Offer, divided by the depth at the SIP Bid or SIP Offer, respectively. *Better Bid* and *Better Offer* are indicators for whether odd lot quotes exist that are superior to the SIP Bid or SIP Offer, respectively. *Bid Level* and *Offer Level* are counts of the number of price points where superior odd lot pricing exists relative to the SIP Bid and the SIP Offer, respectively. Stock-fixed effects are included in columns (3) , (4), (7) and (8). Columns (1) - (4) present results for all Group 1 stocks, and Columns (5) - (8) present results for all Group 5 aroxka. Odd column numbers are the SIP Model, and even column numbers are the OL Model. Stock-clustered standard errors are in brackets. *,**, and *** indicate significance levels of 10%, 5%, and 1%.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Group 1				Group 5			
DV:	Return on SIP Midpoint	Return on SIP Midpoint	Return on SIP Midpoint	Return on SIP Midpoint	Return on SIP Midpoint	Return on SIP Midpoint	Return on SIP Midpoint	Return on SIP Midpoint
Prior VWAP Return	0.0186***	0.0184***	0.0186***	0.0184***	0.0394***	0.0291***	0.0394***	0.0291***
	[0.00177]	[0.00176]	[0.00177]	[0.00176]	[0.00286]	[0.00258]	[0.00286]	[0.00258]
Log Volume	4.67e-06***	4.62e-06***	7.80e-06***	7.63e-06***	-6.52E-07	-6.99E-07	1.02E-06	1.16e-06*
	[9.45e-07]	[9.40e-07]	[9.60e-07]	[9.52e-07]	[7.23e-07]	[6.90e-07]	[7.31e-07]	[6.81e-07]
Cumulative Volume	-4.13e-06***	-3.05e-06*	8.65e-06***	1.03e-05***	-7.75E-07	-3.07e-06**	-3.89E-07	-2.25E-06
	[1.56e-06]	[1.72e-06]	[1.74e-06]	[2.12e-06]	[1.52e-06]	[1.28e-06]	[1.87e-06]	[1.60e-06]
SIP Spread	-0.000473***	-0.00179***	-0.000257	-0.00159***	-5.81e-06*	-2.75E-06	-1.33e-05**	-6.24E-06
	[0.000161]	[0.000570]	[0.000189]	[0.000562]	[3.33e-06]	[3.62e-06]	[5.14e-06]	[3.80e-06]
OL Spread		0.00154***		0.00163***		-1.38e-05**		-1.90e-05**
		[0.000482]		[0.000479]		[6.72e-06]		[7.49e-06]
SIP Midpoint	1.06e-06**	-0.00384***	-1.63e-06***	-0.00389***	1.04e-08*	-0.000117***	1.09E-09	-0.000117***
	[4.80e-07]	[0.00101]	[4.50e-07]	[0.00100]	[5.49e-09]	[2.86e-05]	[3.76e-08]	[2.88e-05]
OL Midpoint		0.00384***		0.00389***		0.000117***		0.000117***
		[0.00101]		[0.00100]		[2.86e-05]		[2.88e-05]
SIP Imbalance	2.56e-09***	2.56e-09***	2.58e-09***	2.58e-09***	1.32e-07***	1.08e-07***	1.32e-07***	1.08e-07***
	[8.06e-10]	[8.06e-10]	[8.14e-10]	[8.14e-10]	[1.54e-08]	[1.32e-08]	[1.55e-08]	[1.32e-08]
OL Imbalance		-4.68E-09		-4.61E-09		4.43e-07***		4.42e-07***
		[1.94e-08]		[1.94e-08]		[1.33e-07]		[1.33e-07]
Prior SIP MP Return	-0.00942***	-0.00961	-0.00942***	-0.00832	-0.0319***	-0.0193**	-0.0320***	-0.0195**
	[0.00239]	[0.0229]	[0.00239]	[0.0229]	[0.00367]	[0.00794]	[0.00368]	[0.00788]
Prior OL MP Return		0.00217		0.0009		0.0126		0.0127
		[0.0231]		[0.0230]		[0.00821]		[0.00814]
Difference		0.222***		0.223***		0.509***		0.508***
		[0.0688]		[0.0687]		[0.0297]		[0.0298]
OL Bid Depth		4.59e-05***		4.60e-05***		-4.78e-05***		-4.60e-05***
		[1.58e-05]		[1.58e-05]		[1.23e-05]		[1.21e-05]
OL Offer Depth		-3.69e-05***		-3.68e-05***		5.01e-05***		5.18e-05***
		[1.18e-05]		[1.18e-05]		[1.38e-05]		[1.37e-05]
Better Bid		-0.000138***		-0.000129***		-7.92e-05***		-8.04e-05***
		[3.33e-05]		[3.37e-05]		[5.41e-06]		[5.43e-06]
Better Offer		8.51e-05***		9.09e-05***		7.12e-05***		7.00e-05***
		[2.85e-05]		[2.74e-05]		[5.43e-06]		[5.41e-06]
Bid Level		9.62e-05***		9.85e-05***		1.91E-07		2.11E-07
		[3.23e-05]		[3.23e-05]		[7.11e-07]		[7.26e-07]
Offer Level		-3.37E-05		-3.17E-05		5.20E-07		4.61E-07
		[2.42e-05]		[2.46e-05]		[6.77e-07]		[6.67e-07]
Constant	4.40E-06	-1.58E-05	-0.000171***	-0.000202***	1.63E-05	4.92e-05***	1.10E-05	4.73e-05**
	[2.06e-05]	[2.35e-05]	[2.96e-05]	[3.56e-05]	[1.91e-05]	[1.70e-05]	[2.79e-05]	[2.25e-05]
Stock FE	N	N	Y	Y	N	N	Y	Y
Observations	11,373,506	11,373,128	11,373,506	11,373,128	2,867,534	2,867,150	2,867,534	2,867,150
R-squared	0.002	0.004	0.003	0.004	0.005	0.026	0.005	0.026
# of Stocks	496	496	496	496	122	122	122	122

Table 8: Predictors of Future Trade Prices - ML Classification

This table presents the results of a supervised machine learning classifier that predicts whether the VWAP of trades between the beginning of minute *t* until the commencement of *t+1* is greater than the SIP MP observed at the commencement of minute t. Panel A provides summary statistics of the classification label ($VWAP_{t+1}$ > SIP $Midpoint_t$) and all features, which were measured at the commencment of minute *t*. Column 1 summarizes the SIP Model, and Column 2 summarizes the OL Model. Panel B presents summary performance statistics for when the two models are applied to the Group 1 stocks (columns 1 and 2) and Group 5 stocks (columns 3 and 4). Results are from an XGBoost model that was separately trained on the SIP Model and the OL Model for both Group 1 stocks and Group 5 stocks. Reported statistics represent the models' performance on the test data for all four models. Panel C presents the average daily return from an investment strategy relied on the predict classification to buy (sell) each security in the test data at the beginning of each minute t and to sell (buy) at the end of the minute.

Panel A

	(1)	(2)
	Baseline SIP Model	Odd Lot Model
Label:		
$VWAP_{t=1}$ > SIP $Midpoint_{t=0}$	Included	Included
Features at t=0:		
Log Volume	Included	Included
Cumulative Volume	Included	Included
Prior VWAP Return	Included	Included
SIP Spread	Included	Included
OL Spread		Included
SIP Midpoint	Included	Included
OL Midpoint		Included
SIP Imbalance	Included	Included
OL Imbalance		Included
Prior SIP MP Return	Included	Included
Prior OL MP Return		Included
Difference		Included
OL Bid Depth		Included
OL Offer Depth		Included
Better Bid		Included
Better Offer		Included
Bid Level		Included
Offer Level		Included

Panel B

	(1)	(2)	(3)	(4)
	Group 1 Stocks		Group 5 Stocks	
Measure:	Baseline SIP Model	Odd Lot Model	Baseline SIP Model	Odd Lot Model
Accuracy	58%	58%	54%	55%
Precision	58%	58%	54%	56%
Recall	58%	59%	56%	58%
F1	58%	58%	55%	57%
ROC	61%	61%	55%	58%

Table 9. Constructing Top of Book Data Under SEC Proposal

This table illustrates how the depth of book data from May Street is converted into the top-of-book quote under the SEC's proposed round lot definition. Columns 1 -3 show the depth of book for bids for AMZN as of 13:25:00 on 3/31/21 displayed on Nasdaq. Under the SEC's proposed round lot definition, AMZN's stock price for Februrary 2021 would have made it a "category 3" security with a round lot of 10 shares for March 2021. The SEC's proposed rule requires calculating the best bid and offer prices as the least aggressive prices that would satisfy all displayed orders in an amount equal to a round lot. Thus, Columns 4 and 5 show thatthe Nasdaq round lot bid would be $3,111.81 x 23 because this is the least aggressive bid price for which there are at least 10 shares that would be satified by buying AMZN at this price.

(1)	(2)	(3)	(4)	(5)
Bids Displayed on Nasdaq's Proprietary Data Feeds			Top of Book Bid Data Under Proposed Round Lot	
Level	Bid Price	Bid Size	Bid Price	Bid Size
1	3112.59	1	3111.81	23
2	3112.3	1		
3	3112.02	1		
4	3112	1		
5	3111.92	2		
6	3111.91	1		
7	3111.81	16		
8	3111.79	7		
9	3111.76	1		
10	3111.59	25		
11	3111.55	5		
12	3111.52	1		
13	3111.33	27		
14	3111.32	1		
15	3111.1	2		
16	3111.06	1		
17	3110.68	2		
18	3110.67	100		

Table 10: Effect of SEC Proposed Round Lots on Incidence of Odd Lot Pricing, Jan. 2021 - March 2021

This table compares how the SEC's proposed round lot definitions would have changed the rate at which odd lot quotes would have provided superior pricing to the NBBO during the sample period across sample securities. For each stock in a proposed round lot category, all end-of-minute (EOM) quotes were evaluated for whether an odd lot bid or offer was displayed whose price was superior to that of the best bid or offer calculated using round lot quotes. Incidence of Superior Prices - Current Rule is the average percent of EOM observations for each stock in a round lot category when an odd lot bid was superior to the best round lot bid or offer where a round lot is 100 shares. Incidence of Superior Prices - Proposed Rule is the average percent of EOM observerations for each stock in a round lot category when an odd lot bid or offer was superior to the best round lot bid or offer where a round lot is equal to the SEC's proposed round lot size. Panel A presents results for sample stocks where the proposed round lot size would remain at 100 shares. Panel B presents results for sample stocks where the round lot size would decrease from 100 shares to 40 shares. Panel C presents results for sample stocks where the round lot size would decrease from 100 shares to 10 shares.

Panel A: Sample Stocks Where Proposed Round Lot = 100 Shares

Bid	Mean Stocks Per Month	mean	sd	p25	p50	p75
Incidence of Superior Prices - Current Rule	1636	16.1%	12.3%	3.6%	16.2%	25.3%
Incidence of Superior Prices - Proposed Rule	1636	16.1%	12.3%	3.6%	16.2%	25.3%
Difference		0.0%	0.0%	0.0%	0.0%	0.0%
Offer						
Incidence of Superior Prices - Current Rule	1636	14.6%	11.3%	3.1%	14.7%	23.3%
Incidence of Superior Prices - Proposed Rule	1636	14.6%	11.3%	3.1%	14.7%	23.3%
Difference		0.0%	0.0%	0.0%	0.0%	0.0%

Panel B: Sample Stocks Where Proposed Round Lot = 40 Shares

Bid	Mean Stocks Per Month	mean	sd	p25	p50	p75
Incidence of Superior Prices - Current Rule	107	39.9%	8.1%	35.5%	39.1%	43.5%
Incidence of Superior Prices - Proposed Rule	107	35.1%	7.9%	30.2%	34.5%	38.7%
Difference		-4.8%	-0.3%	-5.3%	-4.6%	-4.8%
Offer						
Incidence of Superior Prices - Current Rule	107	38.1%	8.0%	32.4%	37.2%	43.8%
Incidence of Superior Prices - Proposed Rule	107	33.2%	7.8%	27.2%	32.1%	39.1%
Difference		-4.9%	-0.2%	-5.1%	-5.1%	-4.7%

Panel C: Sample Stocks Where Proposed Round Lot = 10 Shares

Bid	Mean Stocks Per Month	mean	sd	p25	p50	p75
Incidence of Superior Prices - Current Rule	8	60.5%	7.5%	54.1%	60.9%	61.9%
Incidence of Superior Prices - Proposed Rule	8	38.8%	6.4%	33.9%	37.9%	42.7%
Difference		-21.7%	-1.2%	-20.2%	-23.0%	-19.3%
Offer						
Incidence of Superior Prices - Current Rule	8	59.6%	8.4%	57.7%	60.3%	62.8%
Incidence of Superior Prices - Proposed Rule	8	37.0%	7.6%	35.9%	37.7%	38.8%
Difference		-22.6%	-0.7%	-21.8%	-22.6%	-23.9%