# Caught On Tape: Predicting Institutional Ownership With Order Flow*

John Y. Campbell, Tarun Ramadorai and Tuomo O. Vuolteenaho†

September 2004

*We thank Peter Hawthorne, Sung Seo, and especially Allie Schwartz for excellent research assistance, and Boris Kovtunenko and Nathan Sosner for their assistance with the Spectrum dataset. We thank seminar participants at the Morgan Stanley microstructure research conference, the Oxford summer symposium, the Oxford econometrics lunch workshop and the European Finance Association for many useful comments. Thanks go to Pablo Casas-Arce, Soeren Hvikdjaer, Pete Kyle, Narayan Naik, Venkatesh Panchapagesan, Jos van Bommel, Joshua White and Pradeep Yadav for useful discussions. This material is based upon work supported by the National Science Foundation under Grant No. 0214061 to Campbell, and by Morgan Stanley and Co. under its Microstructure Research Program.

†Campbell and Vuolteenaho are at Harvard University, Department of Economics, Littauer Center, Cambridge, MA 02138, USA, john_campbell@harvard.edu and t_vuolteenaho@harvard.edu, and the NBER. Ramadorai is at the University of Oxford, Said Business School, Park End St., Oxford OX1 1HP, UK, tarun.ramadorai@said-business-school.oxford.ac.uk.

# Caught On Tape: Predicting Institutional Ownership With Order Flow

September 2004

**Abstract**

Many questions about institutional trading can only be answered if one can track institutional equity ownership continuously. However, these data are only available on quarterly reporting dates. We infer institutional trading behavior from the "tape," the Transactions and Quotes database of the New York Stock Exchange, by regressing quarterly changes in reported institutional ownership on quarterly buy and sell volume in different trade size categories. Our regression method predicts institutional ownership significantly better than the simple cutoff rules used in previous research. We also find that total buy (sell) volume predicts increasing (decreasing) institutional ownership, consistent with institutions demanding liquidity in aggregate. Furthermore, institutions tend to trade in large or very small sizes: buy (sell) volume at these sizes predicts increasing (decreasing) institutional ownership, while the pattern reverses at intermediate trade sizes that appear favored by individuals. We then explore changes in institutional trading strategies. Institutions appear to prefer medium size trades on high volume days and large size trades on high volatility days.

## 1. Introduction

How do institutional investors trade in equity markets? Do they hold stocks that deliver high average returns? Do they arbitrage irrationalities in individual investors' responses to information? Are they a stabilizing or destabilizing influence on stock prices? These questions have been the focus of a large and recent body of empirical literature.

Lakonishok, Shleifer, and Vishny (1992), Grinblatt, Titman, and Wermers (1995), Wermers (1999, 2000), Nofsinger and Sias (1999), and Grinblatt and Keloharju (2000a, b) show that quarterly increases in institutional ownership and quarterly stock returns are contemporaneously correlated. Several studies investigate this relationship further, and find evidence that short-term expected returns are higher (lower) for stocks that have recently been subject to significant institutional buying (selling).[1] Some authors, notably Lakonishok, Shleifer, and Vishny (1992), suggest that institutional investors follow simple price-momentum strategies that push stock prices away from fundamental values. This is disputed by others, such as Cohen, Gompers, and Vuolteenaho (2002), who find that institutions are not simply following price-momentum strategies; rather, they sell shares to individuals when a stock price increases in the absence of any news about underlying cash flows.

One limitation of this literature is that it is difficult to measure changes in institutional ownership as they occur. While some countries, such as Finland, do record institutional ownership continuously, in the United States institutional positions are reported only quarterly in 13-F filings to the Securities and Exchange Commission. A quarterly data frequency makes it hard to say whether institutions are reacting to stock price movements or causing price movements, and makes it impossible to measure institutional responses to high-frequency news such as earnings announcements.

To measure institutional trading at higher frequencies, some authors have looked at data on equity transactions, available on the New York Stock Exchange Trade and Quotes (TAQ) database. Most transactions can be identified as buy orders or sell orders using the procedure

[1] See Daniel, Grinblatt, Titman, and Wermers (1997), Chen, Jegadeesh, and Wermers (2000), and Gompers and Metrick (2001), among others.

of Lee and Ready (1991), which compares the transaction price to posted bid and ask quotes.

A more difficult challenge is to identify orders as coming from institutions or individuals. A common procedure is to label orders above some upper cutoff size as institutional, and those below a lower cutoff size as individual. Trades at intermediate sizes remain unclassified. Lee and Radhakrishna (2000) evaluate several alternative cutoff rules by applying them to the TORQ data set, a sample of trades with complete identification of market participants. They find, for example, that upper and lower cutoffs of $20,000 and $2,500 are most effective at accurately classifying trades in small stocks. Unfortunately the TORQ data set includes only 144 stocks over a three-month period in 1994 and it is not clear that these results apply more generally or in more recent data.

In this paper we develop a new method for inferring high-frequency institutional trading behavior. Our method combines two datasets that in the past have been used separately in analyses of investor behavior. The TAQ database gives us trade-by-trade data pertaining to all listed stocks on the NYSE and AMEX, NASDAQ national market system, and small cap stocks, beginning in 1993. We restrict the current analysis to stocks traded on the NYSE and AMEX. TAQ is essentially the "tape", recording transactions prices and quantities of every trade conducted on these exchanges. We match TAQ to the Spectrum database. Spectrum records the SEC mandated 13-F filings of large institutional investors, providing quarterly snapshots of institutional holdings. Finally, we use the cumulative quarterly trades recorded on the "tape" to predict institutional holdings in Spectrum. By regressing changes in institutional ownership on cumulative trades of different sizes, we find the best function mapping trade size to institutional behavior. This function can be used to track institutional trading on a daily or intra-daily basis.

There is a fundamental difference between the approach in this paper, and that employed in the previous literature attempting to separate individual from institutional ownership. The best known example is the analysis of Lee and Radhakrishna (2000), in which the authors attempt to classify each trade as institutional or individual, using characteristics of the trade, such as the size of the trade in dollars or number of shares. However, this classification is done without regard to the characteristics of other trades that may be occurring in any specified interval of time. In contrast, our method combines the information provided by the entire observed set of trades to get the best overall prediction of changing institutional ownership. It is worth noting at this point that if we had higher frequency institutional ownership data than quarterly 13-F filings, say weekly data, we could still employ our method (at the weekly frequency using the weekly ownership data as our left hand side variable) to yield rich insights about the trading behavior of institutional investors.

The difference between our approach and that employed by the pre-existing literature is dramatic. If one is equipped with a correct classification scheme that gives the true probability that each trade is institutional, then one can aggregate the probability weighted trades to get the best prediction of the change in institutional ownership. In general, however, the probability that a trade is institutional depends on the entire environment and not just on the characteristics of the trade alone. This is best elucidated using an example: suppose all individuals trade in \$10,000 amounts and trade in a perfectly correlated manner (either all sells, or all buys on a particular day); assume that all institutions except one trade in \$10,000 amounts and trade in a manner that is perfectly positively correlated with other institutions and perfectly negatively correlated with individuals; finally one large institution trades in \$100,000 amounts and trades in a manner that is perfectly correlated with other institutions. In this case the probability that a \$10,000 trade is institutional, based on its own characteristics is 50 percent, and the probability that a \$100,000 trade is institutional is 100 percent. However, if we observe a \$100,000 buy, then we can infer that all the \$10,000 buys are institutional with probability 100 percent.

What this means is that the coefficients on trade size bins in a regression predicting institutional ownership may be very different from the probabilities that trades of that size are institutional. In the example above, volume occurring in trade sizes of \$100,000 should get a coefficient that is far greater than one, because it reveals the direction of all the \$10,000 institutional trades. Our paper reports regression coefficients rather than probabilities that particular trades are institutional or individual: we cannot directly infer these probabilities from our regression coefficients. Our analysis gives us a regression function with which to infer institutional ownership, rather than a rule generating probabilities of isolated trades being individual or institutional. We can map from Lee and Radhakrishna and similar rules to regression functions, but we cannot in general do the reverse operation. Using this mapping, we find that our method of inferring institutional buying and selling from the "tape" significantly outperforms the simple classification rules in previous literature. For example, a simple cut-off rule that classifies all trades over $20,000 as institutional has a negative $R^2$ when used as a predictor of the change in institutional ownership. This is in contrast to the 10 percent $R^2$ obtained by our method.

Our second finding is that institutions on average appear to demand liquidity. Across all trades (ignoring trade sizes), volume classifiable as buys predicts an increase and volume classifiable as sells predicts a decline in reported institutional ownership. These results suggest that institutions use the liquidity provided by the specialist and possibly also provided by limit orders from individuals.

Third, we find that buying at the ask and selling at the bid is more likely to be indicative of institutional buying or selling if the trade size is either very small or very large. Trades that are either under $2,000 or over $30,000 in size are very likely to be initiated by institutions, whereas intermediate size trades are relatively more likely to be by individuals.

We then go several steps further. First, we smooth the effects of trade size in our specifications by employing the exponential function of Nelson and Siegel (1987), formerly used for parsimonious yield curve modeling. The resulting specification is far less unwieldy than allowing separate regressors for each trade size bin. We explore the time stability of the parameters of the function, and find that the out-of-sample $R^2$ statistics are still much higher than those generated by simple cutoff rule based classification schemes.

We then use the methodology we develop to explore the sensitivity of the trading patterns of institutional investors to daily movements in volume, returns and volatility. We do so by incorporating daily interactions between these variables and the TAQ flows in various bins into the Nelson-Siegel specification. This generates several new and interesting findings about features of institutional trading behaviour. In brief, institutions appear to stop trading in the very smallest bins in small firms when returns are high, large institutional trades in small firms are concentrated on days when volatility is high, perhaps because institutions are particularly urgent about their transactions at such times; and medium-size institutional trades in small firms are concentrated on days when volume is high, possibly because institutions see an opportunity for stealth trading at times when liquidity is high. Several of these results hold true for the largest firms in our sample as well.

The organization of the paper is as follows. Section 2 introduces the TAQ, Spectrum and CRSP data used in the study, and conducts a preliminary data analysis. Section 3 presents and applies our method for predicting institutional ownership. In Section 4 we explore how institutional trading is affected by variation in daily returns, volume and volatility. Section 5 concludes.

## 2. Preliminary data analysis

### 2.1. CRSP data

Shares outstanding, stock returns, share codes, exchange codes and prices for all stocks come from the Center for Research on Security Prices (CRSP) daily and monthly files. In the current analysis, we focus on ordinary common shares of firms incorporated in the United States that traded on the NYSE and AMEX. Our sample begins in January 1993, and ends in December 2000. We use the CRSP PERMNO, a permanent number assigned to each security, to match CRSP data to TAQ and Spectrum data. Figure 1 shows the evolution of the number of matched firms in our data over time: as firms list or delist from the NYSE and AMEX, or move between NYSE and AMEX and other exchanges, this number changes. The maximum number of firms is 2222, in the third quarter of 1998. The minimum number of firms is 1843, in the first quarter of 1993.

In the majority of our analysis, we present results for all firms, as well as for five quintiles of firms, where quintile breakpoints and membership are determined by the market capitalization (size) of a firm at the start of each quarter. Our data are filtered carefully, as described below. After filtering, our final sample consists of 3334 firms. When sorted quarterly into size quintiles, this results in 735 firms in the largest quintile, and between 1131 and 1357 firms in the other four quintiles (these numbers include transitions of firms between quintiles), and 63,403 firm quarters in total.

### 2.2. TAQ data

The Transactions and Quotes (TAQ) database of the New York Stock Exchange contains trade-by-trade data pertaining to all listed stocks, beginning in 1993. TAQ records transactions prices and quantities of all trades, as well as a record of all stock price quotes that were made. TAQ lists stocks by their tickers. We map each ticker symbol to a CRSP PERMNO. As tickers change over time, and are sometimes recycled or reassigned, this mapping changes over time.

The TAQ database does not classify transactions as buys or sells. To classify the direction of trade, we use an algorithm suggested by Lee and Ready (1991). This algorithm looks at the price of each stock trade relative to contemporaneous quotes in the same stock to determine whether a transaction is a buy or sell. In cases where this trade-quote comparison cannot be accomplished, the algorithm classifies trades that take place on an uptick as buys, and trades that take place on a downtick as sells. The Lee-Ready algorithm cannot classify some trades, including those executed at the opening auction of the NYSE, trades which are labelled as having been batched or split up in execution, and cancelled trades. We aggregate all these trades, together with "zero-tick" trades which cannot be reliably identified as buys or sells, into a separate bin, and use this bin of unclassifiable trades as an additional input into our prediction exercise.

Lee and Radhakrishna (2000) find that the Lee-Ready classification of buys and sells is highly accurate; however it will inevitably misclassify some trades which will create measurement error in our data. Appendix 1 describes in greater detail our implementation of the Lee-Ready algorithm.

Once we have classified trades as buys or sells, we assign them to bins based on their dollar size. In all, we have 19 size bins whose lower cutoffs are $0, $2000, $3000, $5000, $7000, $9000, $10,000, $20,000, $30,000, $50,000, $70,000, $90,000, $100,000, $200,000, $300,000, $500,000, $700,000, $900,000, and $1 million. In several of our specifications below, we use buy and sell bins separately, and in others, we subtract sells from buys to get the net order flow within each trade size bin. We aggregate all shares traded in these dollar size bins to the daily frequency, and then normalize each daily bin by the daily shares outstanding as reported in the CRSP database. This procedure ensures that our results are not distorted by stock splits.

We aggregate the daily normalized trades within each quarter to obtain quarterly buy and sell volume at each trade size. The difference between these is net order imbalance or net order flow. We normalize and aggregate unclassifiable volume in a similar fashion. The sum of buy, sell, and unclassifiable volumes is the TAQ measure of total volume in each stock-quarter.

We filter the data in order to eliminate potential sources of error. We first exclude all stock-quarters for which TAQ total volume as a percentage of shares outstanding is greater than 200 percent (there are a total of 102 such stock-quarters). We then compute the standard deviation across stock-quarters of each volume measure and the net order imbalance, relative to each quarter's cross-sectional mean, and winsorize all observations that are further than 2.5 standard deviations from their cross-sectional mean. That is, we replace such outliers with the cross-sectional mean for the quarter plus or minus 2.5 standard deviations. This winsorization procedure affects between 2.50 and 3.15 percent of our data.

Figure 2 shows equal and market capitalization weighted cross-sectional means and standard deviations of TAQ total volume as a percentage of shares outstanding in each quarter, in annualized percentage points. In the early years of our sample period equal weighted total volume averaged between 60 percent and 80 percent of shares outstanding per year; this increased to between 80 percent and 100 percent in the later years of the sample. These numbers are consistent with other recent studies such as Chen, Hong and Stein (2002) and

Daves, Wansley and Zhang (2003). The equal weighted cross-sectional standard deviation ranges between 30 and 40 percent of total shares outstanding. This indicates that there is considerable cross-sectional heterogeneity in volume. Some of this cross-sectional heterogeneity can be explained by differences in the trading patterns in small and large stocks. The size-weighted average indicates that total volume as a percentage of shares outstanding has experienced a relative increase for the larger stocks in the later years of the sample. However, the size-weighted standard deviation of total volume as a percentage of shares outstanding is not dramatically different from the equal weighted standard deviation.

The differences in trading patterns across small and large stocks are summarized differently in Table I, which reports means, medians, and standard deviations across all firm-quarters, and across firm-quarters within each quintile of market capitalization. Mean total volume ranges from 53 percent of shares outstanding in the smallest quintile to 91 percent in the largest quintile. Figure 2 suggests that much of this difference manifests itself in the final years of our sample. The distribution of total volume is positively skewed within each quintile, so median volumes are somewhat lower. Nevertheless, median volumes also increase with market capitalization. This is consistent with the results of Lo and Wang (2000), who attribute the positive association between firm size and turnover to the propensity of active institutional investors to hold large stocks for reasons of liquidity and corporate control. The within-quintile annualized standard deviations (computed by multiplying quarterly standard deviation by a factor of 200, under the assumption that quarterly observations are iid) are fairly similar for stocks of all sizes, ranging from 27 percent to 33 percent.

Table I also reports the moments of the net order flow for each size quintile. Mean net order flow increases strongly with market capitalization, ranging from –2.1 percent for the smallest quintile to 4.5 percent for the largest quintile. This suggests that over our sample period, there has been buying pressure in large stocks and selling pressure in small stocks, with the opposite side of the transactions being accommodated by unclassifiable trades that might include limit orders.[2] This is consistent with the strong price performance of large

[2]In support of this interpretation, net order flow is strongly negatively correlated with Greene's [1995]

stocks during most of this period.

Unclassifiable volume is on average about 15 percent of shares outstanding in our data set. This number increases with firm size roughly in proportion to total volume; our algorithm fails to classify 18 percent of total volume in the smallest quintile, and 21 percent of total volume in the largest quintile. It is encouraging that the algorithm appears equally reliable among firms of all sizes. Note that the means of buy volume, sell volume, and unclassifiable volume do not exactly sum to the mean of total volume because each of these variables has been winsorized separately.

Figure 3 summarizes the distribution of buy and sell volume across trade sizes. The figure reports three histograms: for the smallest, median, and the largest quintiles of stocks. Since our trade size bins have different widths, ranging from \$1000 in the second bin to \$200,000 in the penultimate bin and even more in the largest bin, we normalize each percentage of total buy or sell volume by the width of each bin, plotting "trade intensities" rather than trade sizes within each bin. As the largest bin aggregates all trades greater than \$1 million in size, we arbitrarily assume that this bin has a width of \$5 million.

It is immediately obvious from Figure 3 that trade sizes are positively skewed, and that their distribution varies strongly with the market capitalization of the firm. In the smallest quintile of stocks almost no trades of over \$70,000 are observed, while such large trades are commonplace in the largest quintile of stocks. A more subtle pattern is that in small stocks, buys tend to be somewhat smaller than sells, while in large stocks the reverse is true.

Table II summarizes the distribution of trade sizes in a somewhat different fashion. The table reports the medians and cross-sectional standard deviations of total classifiable volume (buys plus sells) in each trade size bin for each quintile of market capitalization. The rarity of large trades in small stocks is apparent in the zero medians and tiny standard deviations for large-size volume in the smallest quintile of firms.

measure of limit order depth for all size quintiles of stocks. This measure essentially identifies a limit order execution as the quoted depth when a market order execution is accompanied by a movement of the revised quote away from the quoted midpoint.

## 2.3. Spectrum data

Our data on institutional equity ownership come from the Spectrum database, currently distributed by Thomson Financial. They have been extensively cleaned by Kovtunenko and Sosner (2003) to remove inconsistencies, and to fill in missing information that can be reconstructed from prior and future Spectrum observations for the same stock. A more detailed description of the Spectrum data is presented in Appendix 2. Again, we first filter the data by removing any observation for which the change in Spectrum recorded institutional ownership as a percentage of firm shares outstanding is greater than 100 percent (there are 28 such stock-quarters). We then winsorize these data in the same manner as the TAQ data, truncating observations that are more than 2.5 standard deviations away from each quarter's cross-sectional mean. This procedure affects 2.5 percent of our Spectrum data.

Table I reports the mean, median, and standard deviation of the change in institutional ownership, as a percentage of shares outstanding. Across all firms, institutional ownership increased by an average of 0.6 percent per year, but this overall trend conceals a shift by institutions from small firms to large and especially mid-cap firms. Institutional ownership fell by 1.3 percent per year in the smallest quintile but rose by 1.7 percent per year in the median quintile and 0.8 percent per year in the largest quintile.

On average, then, institutions have been selling smaller stocks and buying larger stocks. This corresponds nicely with the trade intensity histograms in Figure 3, which show that the smallest stocks tend to have larger-size sales than buys, while the largest stocks have larger-size buys than sells. If institutions more likely trade in large sizes, we would expect this pattern. The behavior of mid-cap stocks is however anomalous in that these stocks have larger-size sales than buys despite their growth in institutional ownership.

We now turn to our regression methodology for predicting institutional ownership.

## 3. Predicting institutional ownership

### 3.1. Regression methodology

In the market microstructure literature, institutional trading behavior has generally been identified using a cutoff rule. Trades above an upper cutoff size are classified as institutional, trades below a lower cutoff size are classified as individual, and intermediate-size trades are unclassified. Lee and Radhakrishna (2000) evaluate alternative cutoff rules using the TORQ data set. As an example of their findings, they recommend an upper cutoff of $20,000 in small stocks. 84 percent of individual investors' trades are smaller than this, and the likelihood of finding an individual initiated trade larger than this size is 2 percent.

Our methodology refines and extends the idea of using an optimally chosen cutoff rule. We match the TAQ data at a variety of trade sizes to the Spectrum data for a broad cross-section of stocks, over our entire sample period. That is, we use the intra-quarter tape to predict institutional ownership at the end of the quarter. Our predictive regression combines information from various trade size bins in the way that best explains the quarterly changes in institutional ownership identified in Spectrum.

We begin with extremely simple regressions that ignore the information in trade sizes. Writing $Y_{it}$ for the share of firm $i$ that is owned by institutions at the end of quarter $t$, $U_{it}$ for unclassifiable trading volume, $B_{it}$ for total buy volume, and $S_{it}$ for total sell volume in stock $i$ during quarter $t$ (all variables are expressed as percentages of the end-of-quarter $t$ shares outstanding of stock $i$), we estimate

$$\Delta Y_{it} = \alpha + \phi Y_{i,t-1} + \rho \Delta Y_{i,t-1} + \beta_U U_{it} + \beta_B B_{it} + \beta_S S_{it} + \varepsilon_{it} \tag{3.1}$$

This regression tells us how much of the variation in institutional ownership can be explained simply by the upward drift in institutional ownership of all stocks (the intercept coefficient $\alpha$), short and long-run mean-reversion in the institutional share for particular stocks (the autoregressive coefficients $\phi$ and $\rho$), and the total unclassifiable, buy, and sell volumes during the quarter (the coefficients $\beta_U$, $\beta_B$, and $\beta_S$). An even simpler variant of this regression restricts the coefficients on buy and sell volume to be equal and opposite, so that the explanatory variable becomes net order flow $F_{it} = B_{it} - S_{it}$ and we estimate

$$\Delta Y_{it} = \alpha + \phi Y_{i,t-1} + \rho \Delta Y_{i,t-1} + \beta_U U_{it} + \beta_F F_{it} + \varepsilon_{it} \tag{3.2}$$

We also consider variants of these regressions in which the intercept $\alpha$ is replaced by time dummies that soak up time-series variation in the institutional share of the stock market as a whole. In this case the remaining coefficients are identified purely by cross-sectional variation in institutional ownership, and changes in this cross-sectional variation over time. Standard errors in all cases are computed using the delete-cross-section jackknife methodology of Shao and Wu (1989) and Shao (1989). The jackknife estimator, besides being nonparametric, has the added advantage of being robust to heteroskedasticity and cross-contemporaneous correlation of the residuals.

Table III reports estimates of equation (3.1) in the top panel, and equation (3.2) in the bottom panel. Within each panel, column A restricts the lagged level of the dependent variable, the lagged change in the dependent variable and unclassifiable volume to have zero coefficients, column B restricts the lagged dependent variable, and the lagged change in the dependent variable, column C restricts only the lagged change in the dependent variable, and column D is unrestricted. Columns E, F, and G repeat these specifications including time dummies rather than an intercept. The results are remarkably consistent across all specifications. On average, buy volume gets a coefficient of about 0.37 and sell volume gets a coefficient of about –0.46. This suggests that institutions tend to use market orders, buying at the ask and selling at the bid or buying on upticks and selling on downticks, so that their orders dominate classifiable volume. The larger absolute value of the sell coefficient indicates that institutions are particularly likely to behave in this way when they are selling. The autoregressive coefficients are negative, and small but precisely estimated, telling us that there is statistically detectable mean-reversion in institutional ownership, at both short and long-run horizons.

The coefficient on unclassifiable volume is small and only marginally significant when buys and sells are included separately in equation (3.1), but it becomes significantly negative when buys and sells are restricted to have equal and opposite coefficients in equation (3.2). To understand this, note that a stock with an equal buy and sell volume is predicted to have declining institutional ownership in the top panel of Table III. The net flow regression in the bottom panel cannot capture this effect through the net flow variable, which is identically zero if buy and sell volume are equal. Instead, it captures the effect through a negative coefficient on unclassifiable volume, which is correlated with total volume.

Table IV repeats the unrestricted regressions incorporating time dummies, for the five quintiles of market capitalization. The main result here is that the coefficients on buys, sells, and net flows are strongly increasing in market capitalization. Evidently trading volume is more informative about institutional ownership in large firms than in small firms. The explanatory power of these regressions is U-shaped in market capitalization, above eight percent for the smallest firms, above 10 percent for the largest quintile, and around six percent for the median size firms. This is consistent with the fact, reported in Table II, that institutional ownership has the greatest cross-sectional volatility in mid-cap firms.

### 3.2. The information in trade size

The above summary regressions ignore the information contained in trade size. We now generalize our specification to allow separate coefficients on buy and sell volume in each trade size bin:

$$\Delta Y_{it} = \alpha + \phi Y_{i,t-1} + \beta_U U_{it} + \sum_Z \beta_{BZ} B_{Zit} + \sum_Z \beta_{SZ} S_{Zit} + \varepsilon_{it}, \tag{3.3}$$

where $Z$ indexes trade size. In the case where we use net flows rather than separate buys and sells, the regression becomes

$$\Delta Y_{it} = \alpha + \phi Y_{i,t-1} + \beta_U U_{it} + \sum_Z \beta_{FZ} F_{Zit} + \varepsilon_{it}. \tag{3.4}$$

Table V estimates equation (3.4) separately for each quintile of market capitalization, replacing the intercept $\alpha$ with time dummies. It is immediately apparent that the coefficients tend to be negative for smaller trades and positive for larger trades, consistent with the intuition that order flow in small sizes reflects individual buying while order flow in large sizes reflects institutional buying. There is however an interesting exception to this pattern. Extremely small trades of less than $2,000 have a significantly positive coefficient in the smallest three quintiles of firms, and in all quintiles have a coefficient that is much larger than that for somewhat larger trades. This is consistent with several possibilities. Institutions might break trades into extremely small sizes when they are "stealth trading" (trying to conceal their activity from the market), or institutions are likely to engage in "scrum trades" to round off extremely small equity positions.[3] Another possibility is that institutions may put in tiny "iceberg" trades to test the waters before trading in larger sizes. It could also be the case that these trades are in fact by individuals, but they are correlated with unobserved variables (such as news events). This could generate unclassifiable volume from institutions in a direction consistent with small trades.

These results are illustrated graphically in Figure 4. Figure 4 standardizes the net flow coefficients, for the smallest, median, and largest quintiles, subtracting their mean and dividing by their standard deviation so that the set of coefficients has mean zero and standard deviation one. The standardized coefficients are then plotted against trade size. In all cases the trough for trade sizes between $2,000 and $30,000 is clearly visible. Consistent with the results for net flows, it turns out that the pattern of coefficients for the case where buys and sells are included separately in the trade-size regression shows a trough and subsequent hump for buy coefficients, and a hump and subsequent trough for sell coefficients.

The information in trade sizes adds considerable explanatory power to our regressions. Comparing the second panel in Table IV with Table V, the $R^2$ statistics increase from 8.1 percent to 9.7 percent in the smallest quintile, from 5.7 percent to 12.1 percent in the median

[3] Chakravarty (2001) presents an in-depth analysis of stealth trading (defined, consistently with Barclay and Warner (1993) as the trading of informed traders that attempt to pass undetected by the market maker). He shows that stealth trading (i.e., trading that is disproportionately likely to be associated with large price changes) occurs primarily via medium-sized trades by institutions of 500-9,999 shares. This runs contrary to our result here.

quintile, and from 10.2 percent to 13.9 percent in the largest quintile (all $R^2$ statistics are computed after time specific fixed effects are removed). The corresponding numbers for the trade-size regressions incorporating buys and sells separately are: $R^2$ statistics increase from 8.2 to 11.9 percent in the smallest quintile, from six percent to 13.4 percent in the median quintile, and from 10.5 percent to 14.5 percent in the largest quintile. Of course, these $R^2$ statistics remain fairly modest, but it should not be surprising that institutional trading activity is hard to predict given the incentives that institutions have to conceal their activity, the considerable overlap between the trade sizes that may be used by wealthy individuals and by smaller institutions, and the increasing use of internalization and off-market matching of trades by institutional investors.

Table VI shows that our regressions are a considerable improvement over the naive cutoff approach used in the previous market microstructure literature. The cutoff model can be thought of as a restricted regression where buys in sizes above the upper cutoff get a coefficient of plus one, buys in sizes below the lower cutoff get a coefficient of minus one, and buys in intermediate sizes get a coefficient of zero. We estimate variants of this regression in Table VI, allowing greater flexibility in successive specifications. In all cases, to present a fair comparison with our method, we allow free coefficients on both the lagged level and lagged change in institutional ownership on the right hand side of each regression. When the coefficient restrictions implied by the naive approach are imposed, we find that the $R^2$ statistic in most cases is negative. In fact, the $R^2$ statistic given the restrictions on the flows above and below the cutoffs is positive only twice for the two smallest size quintiles, and maximized at 4.8 percent, 5.4 percent and 9.8 percent for the median, fourth and largest quintiles respectively. In the $R^2$ comparison, we move progressively closer to our own method, finally allowing freely estimated coefficients on the cutoff values proposed by Lee and Radhakrishna. When we allow flows above and below the cutoffs to have free coefficients, the $R^2$ statistics of the regressions increase substantially but in most cases are well below those of our freely estimated regressions in Table V.

### 3.3. Smoothing the effect of trade size

One concern about the specifications (3.3) and (3.4) is that they require the separate estimation of a large number of coefficients. This is particularly troublesome for small stocks, where large trades are extremely rare: the coefficients on large-size order flow may just reflect a few unusual trades. One way to handle this problem is to estimate a smooth function relating the buy, sell, or net flow coefficients to the dollar bin sizes. We have considered polynomials in trade size, and also the exponential function suggested by Nelson and Siegel (1987) to model yield curves. We find that the Nelson and Siegel method is well able to capture the shape suggested by our unrestricted specifications. For the net flow equation, the method requires estimating a function $\beta(Z)$ that varies with trade size $Z$, and is of the form:

$$\beta(Z) = b_0 + (b_1 + b_2)\,[1 - e^{-Z/\tau}]\frac{\tau}{Z} - b_2 e^{-Z/\tau}. \tag{3.5}$$

Here $b_0, b_1, b_2$, and $\tau$ are parameters to be estimated. The parameter $\tau$ is a constant that controls the speed at which the function $\beta(Z)$ approaches its limit $b_0$ as trade size $Z$ increases. We estimate the function using nonlinear least squares, searching over different values of $\tau$, to select the function that maximizes the $R^2$ statistic:

$$\Delta Y_{it} = \alpha + \phi Y_{i,t-1} + \beta_U U_{it} + b_0 \sum_Z F_{Zit} + b_1 \sum_Z g_1(Z) F_{Zit} + b_2 \sum_Z g_2(Z) F_{Zit} + \varepsilon_{it}, \tag{3.6}$$

where $g_1(Z) = \frac{\tau}{Z}(1 - e^{-Z/\tau})$ and $g_2(Z) = \frac{\tau}{Z}(1 - e^{-Z/\tau}) - e^{-Z/\tau}$.

Table VII presents the coefficients from estimates of equation (3.6). The $R^2$ statistics from estimating the Nelson-Siegel specification are slightly lower than the ones shown in Table V at nine percent for the smallest quintile of stocks, 10.7 percent for the median quintile, and 12.6 percent for the largest quintile. The statistical significance of the estimated parameters is quite high, giving us some confidence in the precision of our estimates of the implied trade-size specific coefficients.

Figure 5 plots the trade-size coefficients implied by estimating (3.6). The pattern of coefficients in Figure 4 is accentuated and clarified. As before, the figure standardizes the net flow coefficients, subtracting their mean and dividing by their standard deviation so that the set of coefficients has mean zero and standard deviation one. Figure 6 presents buy and sell coefficients estimated using an analogous Nelson-Siegel specification. Again, the shapes that appear in the two panels are consistent with our results from the specification estimated in Table V, that allows separate coefficients for each trade size bin.

The parsimony of equation (3.6) is extremely useful, in that it permits a relatively straightforward investigation of changes in the functional form over time. This allows us to investigate the time stability of our regression coefficients, and to compare the out of sample forecasting power of our method to the $R^2$ statistics implied by the Lee-Radhakrishna method. The last two rows in Table VI show the implied $R^2$ statistics generated by the out of sample forecasts generated by the Nelson-Siegel specification. We first estimate the Nelson-Siegel specification over the first half of the entire sample, from the first quarter of 1993 until the final quarter of 1996. We then fix the coefficients and calculate the out-of-sample $R^2$ over the entire second half of the full sample period, from the first quarter of 1997 until the final quarter of 2000. In all cases, our implied out-of-sample $R^2$ are higher than the restricted coefficient estimates implied by Lee-Radhakrishna, though in some cases less than the $R^2$ statistics generated when free coefficients are allowed on the implied cutoff points.

We also compute out of sample $R^2$ statistics in a more sophisticated manner. We begin by using the first quarter of the entire sample period, from the first quarter of 1993 until the final quarter of 1994, and construct an implied fitted value for the first quarter of 1995 using the parameters estimated over the earlier period. We then re-estimate the Nelson-Siegel function each period, progressively forecasting one period ahead, each period. The implied out of sample $R^2$ statistics from this procedure are presented in the last row of Table VI. For the two smallest size quintiles of stocks, these are higher than any $R^2$ statistic generated by the Lee-Radhakrishna method, including the unrestricted cutoff coefficients specification. For the three largest size quintiles of stocks, the $R^2$ statistics are higher than all but the unrestricted cutoff coefficients specification.

The use of the functional form (3.6) also gives us an uncomplicated way to explore the interaction of institutional trading strategies with firm characteristics and market conditions. This is the topic of the next section.

# 4. Institutional Trading, Returns, Volume and Volatility

There has been little investigation of changes in the trading strategies of institutional investors in response to movements in variables such as total volume, returns and volatility. We attempt to shed light on these questions by augmenting the Nelson-Siegel functional form to incorporate interactions between these variables and the flows in various trade size bins.

## 4.1. Estimating Interaction Effects

The interaction variables we employ are: daily volume (measured as a fraction of total shares outstanding to normalize for stock splits); daily volatility (measured as the absolute value of returns); daily returns; and average daily quoted depth (measured as the average of depth at the bid and depth at the ask across all quotes each day as reported in the TAQ data set, and normalized by daily shares outstanding).

We consider a variation of the Nelson-Siegel function (3.5) which varies with trade size $Z$, as well as with an interaction variable represented by $\nu$. We separately estimate the specification for each one of the interaction variables independently, and do not in this analysis consider simultaneous movements in the interaction variables.

Here, the subscript $d$ indicates the daily frequency:

$$\beta(Z, \nu_{id}) = b_{01} + b_{02}\nu_{id} + (b_{11} + b_{12}\nu_{id} + b_{21} + b_{22}\nu_{id})\,[1 - e^{-Z/\tau}]\frac{\tau}{Z} - (b_{21} + b_{22}\nu_{id})e^{-Z/\tau} \quad (4.1)$$

Note here that we do not allow the parameter $\tau$ to vary with $\nu_{id}$, as a simplification.

As before, $g_1(Z) = \frac{\tau}{Z}(1 - e^{-Z/\tau})$ and $g_2(Z) = \frac{\tau}{Z}(1 - e^{-Z/\tau}) - e^{-Z/\tau}$. Armed with the parameters of function (4.1), we can evaluate the function at various levels of $\nu_{id}$, providing comparative statics on changes in institutional trading patterns with movements in volume, volatility and returns.

In order to estimate these parameters, we consider a daily version of specification (3.6). For the moment, we disregard the inclusion of the lagged level of institutional ownership (we can subsequently condition on it in the quarterly estimation):

$$\begin{aligned}\Delta Y_{id} = \alpha + \rho \Delta Y_{i,d-1} + \beta_U U_{id} + \beta_{U\nu} \nu_{id} U_{id} + \beta_\nu \nu_{id} \\ + b_{01} \sum_Z F_{Zid} + b_{02} \sum_Z \nu_{id} F_{Zid} + b_{11} \sum_Z g_1(Z) F_{Zid} \\ + b_{12} \sum_Z g_1(Z) \nu_{id} F_{Zid} + b_{21} \sum_Z g_2(Z) F_{Zid} + b_{22} \sum_Z g_2(Z) \nu_{id} F_{Zid} + \varepsilon_{id} \quad (4.2)\end{aligned}$$

We can then aggregate this daily function up to the quarterly frequency, ($q$ represents the number of days in a quarter, and as before, $t$ indicates the quarterly frequency), resulting in:

$$\begin{aligned}\Delta Y_{it} = q\alpha + \rho \Delta Y_{i,t-1} + \beta_U U_{it} + \beta_{U\nu} \left( \sum_{d=1}^{q} \nu_{id} U_{id} \right) + \beta_\nu \nu_{it} \\ + b_{01} \sum_Z F_{Zit} + b_{02} \sum_Z \left( \sum_{d=1}^{q} \nu_{id} F_{Zid} \right) + b_{11} \sum_Z g_1(Z) F_{Zit} \\ + b_{12} \sum_Z g_1(Z) \left( \sum_{d=1}^{q} \nu_{id} F_{Zid} \right) + b_{21} \sum_Z g_2(Z) F_{Zit} + b_{22} \sum_Z g_2(Z) \left( \sum_{d=1}^{q} \nu_{id} F_{Zid} \right) + \varepsilon_{it} \quad (4.3)\end{aligned}$$

We make an assumption here in moving from equation (4.2) to equation (4.3) that the error in measured daily institutional ownership $\varepsilon_{id}$ is uncorrelated at all leads and lags within a quarter with all of the right hand side variables in equation (4.2). This exogeneity assumption guarantees that the parameters of the daily function $b_{01}, b_{02}, b_{11}, b_{12}, b_{21}, b_{22}, \tau$ will be the same as those estimated at the quarterly frequency. Conditional on this assumption, we can estimate equation (4.3) by nonlinear least squares, selecting the function that maximizes the $R^2$ statistic. We additionally incorporate the lagged level of institutional ownership as a right hand side variable in our quarterly estimation to capture long run mean-reversion in the institutional share of holdings of particular stocks. We can then go on to recover the parameters of the daily function $\beta(Z, \nu_{id})$, and evaluate comparative statics at various levels of $\nu$. We now turn to the results from this exercise.

### 4.2. Results

Table VIII evaluates the additional explanatory power generated by estimating the individual interaction specifications (4.3) rather than equation (3.6). In all cases, we first add in $\nu_{it}$ on its own and evaluate the resulting changes to explanatory power. We then assess the subsequent marginal increase in the $R^2$ statistic from interacting $\nu_{it}$ with the trade size bins.

The first feature of note is that simply adding returns to the baseline specification generates large increases in the explained variation of changes to institutional ownership. The increase ranges between five percent for the largest size quintile of stocks, and 30 percent for the second smallest size quintile. This result mirrors the finding in the large body of literature (Lakonishok, Shleifer and Vishny (1992) and Gompers and Metrick (2001) are two notable examples) that examines the relationship between changes in quarterly institutional ownership and returns.

Second, when we add absolute returns to the baseline specification, increases in explanatory power are primarily evident in the two smallest size quintiles of stocks. Apparently movements in volatility directly affect changes in institutional ownership primarily in small stocks.

Next, we find that for all choices of $\nu_{it}$, the interactions with trade size bins turn out to be quite important. Changes in volume, volatility and returns clearly have significant impacts on the trading patterns of market participants. For the absolute return interactions, in all but the smallest quintile of stocks, at least 15 percent additional $R^2$ is generated, going as high as 29 percent for the third size quintile of stocks in our sample. For returns, the increases in $R^2$ from adding in the interaction terms are not as high, peaking at nine percent for the smallest size quintile of stocks, and smallest at around four percent for the third size quintile of stocks. While the explanatory power from including volume or average quoted depth alone is minimal, for volume, the increase in explanatory power from including the interaction variables ranges from four percent for the median size quintile to approximately seven percent for the second size quintile of stocks. For average quoted depth, the equivalent range is between two percent and 17 percent. For all three of the absolute return, volume and especially average quoted depth interactions, the increase in $R^2$ generated by the addition of the interaction terms is generally higher than the additional $R^2$ generated by adding the variable itself. Much of the explanatory power of volatility and volume for changes in institutional ownership comes from the changes that these variables generate in institutional and individual trading strategies.

On net, the total additional explanatory power generated by estimating (4.3) is substantial. This is especially true when $\nu$ is absolute returns or returns - for these two variables, the total improvement in $R^2$ ranges from 10 to 36 percent, and is generally around 30 percent.

For the purposes of comparative statics, we evaluate the function (4.1) in all cases at two different levels: the daily mean level of $\nu$ (computed in all cases as the quarterly mean divided by 63, the mean number of days in a quarter), and two daily standard deviations away from the mean. The daily standard deviations of volume, volatility and returns are calculated as the quarterly standard deviation divided by the square root of 63, under the assumption that these variables are iid at the daily frequency.

For the return, absolute return and volume interactions, the most pronounced effects of the interaction with the bin size coefficients are evident in the smallest size quintile of stocks, though the pattens are broadly similar across the other size quintiles. For the depth interaction, changes are most evident in the largest size quintile of stocks. In the interests of parsimony, we present results from these size quintiles to illustrate the changes to trading strategies with volume, volatility and returns, and specify when the results are dissimilar for the other size quintiles.

#### 4.2.1. Returns

First, $\gamma$, the coefficient on returns, is uniformly positive across size quintiles. The magnitude of the coefficient indicates that a one percent move in returns over a day is associated with a four basis point upward move in institutional holdings as a percentage of shares outstanding for a stock in the median quintile. Figure 7 shows the effect of movements in returns on the trading behaviour of institutional investors in the smallest size quintile of stocks. When returns ($\nu$) are set to their daily mean in the function (4.1), we see institutional buying in the smallest bin, and in bins larger than \$20,000 in size, as before. On days on which returns are two standard deviations above their daily mean, institutional buying in the smallest bin disappears on net. One possible interpretation of this result is that institutions stop using tiny scrum trades in small firms on high return days. Another possibility is that naive individual investors enter the market on such days and do a large amount of small-size buying. We now turn to the effects of volatility on our results.

#### 4.2.2. Volatility

The coefficient on absolute returns $\gamma$ in equation (4.3) is negative for the first two quintiles of stocks, and positive for the remaining three quintiles. However, the magnitude of $\gamma$ is very low - a one percent move in daily volatility generates a maximum of a 1.7 basis point move in institutional ownership over the day, for stocks in quintile four. Figure 7 shows the effects of movements in volatility on institutional trading in the smallest stocks. When volatility is set to its daily mean, we see the familiar pattern in which institutions buy in the smallest size bin and in bins greater than around \$10,000 in size. However, on days on which volatility is two standard deviations above its daily mean, institutional buying becomes more aggressive in the larger size bins. This concentration of large institutional trades in small firms on days when volatility is high suggests that institutions may be particularly urgent about their transactions at such times. The second interesting feature in figure 7 is that at times of high volatility, buying activity appears in the intermediate size bins of \$7000-\$10,000 where none had existed before. This is broadly consistent with a world in which institutions attempt to disguise their activity from the market maker, increasing 'stealth-trading' activity in medium size bins in small stocks at times of high volatility.

In the largest quintile of stocks, institutional trading disappears in the very smallest bin in days with high volatility, much like the behaviour in periods of high returns. Furthermore, institutional trading in large stocks shifts towards the intermediate size bins in periods of high volatility, akin to the behaviour in small stocks.

#### 4.2.3. Volume and Average Quoted Depth

**Robustness of Dollar Size Bin Specification** The flexibility of the Nelson-Siegel functional form also allows us to check whether our specification can be improved by alternative definitions of trade size bins. We currently define our bins in terms of the dollar size of a trade. This dollar based bin classification is motivated by the insight that we can use the wealth constraint experienced by individuals to try to separate the trading behaviour of institutions from that of individuals. In other words, individual investors generally either cannot trade large dollar trade sizes because they simply don't have the money, or dislike making large dollar trades because such trades would result in extremely concentrated and/or leveraged positions relative to their wealth.

Another possible constraint we could use to separate individuals from institutions is the liquidity constraint, i.e. institutions generally do not like to trade illiquid securities for a variety of reasons (such as the desire to window dress their portfolios). This, especially for active institutional traders, indicates a preference for more liquid trade sizes in which it is easier to increase or decrease holdings.[4] This in turn suggests that we redefine our bins each quarter in terms of percentiles of total trading volume that fall within each bin. Yet another approach is to specify bins in terms of multiples of average quoted depth, as a measure of the 'normal' or 'most liquid' trade size in a stock.

A straightforward way to check whether the liquidity constraint can help us better identify institutional ownership is to interact our dollar size bins with measures of liquidity. When

[4] Thanks to Soeren Hvikdjaer for first bringing this issue to our attention.

total daily volume or daily average quoted depth, are high relative to their time series means, does trading shift across into different size bins? Does time variation in liquidity generate variation in the preferred trading habitat of institutional investors?

**Results** When $\nu$ is volume, $\gamma$ is uniformly negative across size quintiles, and approximately the same size across quintiles. For the largest quintile of stocks, the magnitude of the coefficient indicates that when total daily volume in a stock increases by 1 percent of total shares outstanding, institutional holdings in the stock reduce by 2.6 basis points over the day. Institutions, therefore, appear to sell on net when liquidity is high. Turning to the patterns across trade size bins, when volume is set to its daily mean, we see the familar pattern in which institutions appear to trade in the smallest bin, and in bins greater than \$20,000 in size. Figure 9 shows the coefficients for the smallest quintile of stocks. On days when volume is two standard deviations above its daily mean there is a dramatic change in the shape of the function, with institutional trading increasing in all bins between \$7,000 and \$90,000 in size. This increase in the concentration of medium-size institutional trades in small firms on days when volume is high could be because institutions see an opportunity for stealth trading at times of high liquidity. The results for the largest quintile of stocks suggest that institutional trading becomes slightly more aggressive in all bins greater than \$20,000 in size in response to increases in volume.

When $\nu$ is average quoted depth, again $\gamma$ is uniformly negative across size quintiles, and ranges from -0.15 for the smallest quintile of stocks to -0.95 for the second largest quintile. For the largest quintile of stocks, $\gamma$ is -0.34, and the magnitude of the coefficient indicates that when average quoted daily depth increases by one percent of total shares outstanding, institutional holdings in the stock reduce by approximately 34 basis points over the day. This confirms our results from the volume interaction: institutions appear to sell on net when liquidity is high. Figure 10 shows the bin-specific coefficients for the largest size quintile of firms, in which the depth interaction results are quite pronounced. In the largest firms, when depth is high, institutional trading aggressively shifts into the smallest bins, up to around \$10,000 in size, and then picks up again in bins greater than \$50,000. The intermediate bins, as before, appear to be favoured by individuals.

The results for the average quoted depth interaction, along with the results from the volume interaction, suggest that 'small' dollar trade sizes may only appear that way when liquidity is low. Trade sizes previously considered small may seem larger to institutional investors when there is an easing of their liquidity constraint. These results suggest that augmenting our static dollar bin sizes with a time varying liquidity component might yield improvements to our method for predicting institutional ownership.

## 5. Conclusion

This paper has presented a technique for predicting quarterly institutional ownership using the "tape", the publicly available record of all trades and quotes within the quarter. The technique can be used to track high-frequency institutional trading in a large cross-section of stocks. In future research we plan to use this approach to measure patterns of institutional behavior around earnings announcements, stock splits, and other corporate actions.

The results of this paper shed light on the trading behavior of institutions. Total classifiable buy volume predicts increasing institutional ownership and total sell volume predicts decreasing institutional ownership. These results are consistent with institutions tending to buy at the ask and sell at the bid, or to buy on upticks and sell on downticks, and suggest that institutions demand liquidity rather than provide it. The coefficient on total sell volume is larger in absolute value than the coefficient on total buy volume, suggesting that institutions are particularly likely to demand liquidity when they sell. All these patterns are more pronounced in large stocks than in small stocks.

Classifying transactions by their size adds considerable explanatory power to our regressions. Buy volume in sizes between \$2,000 and \$30,000 is associated with decreasing institutional ownership, while buy volume in larger sizes predicts increasing institutional ownership. Interestingly, extremely small buys below \$2,000 also predict increasing institutional ownership, suggesting that institutions use these trades to conceal their activity or to round small positions up or down. All these patterns are reversed for sell volume, and are remarkably consistent across firm sizes.

The method we develop in this paper represents a marked improvement to the pre-existing approach to measuring high frequency institutional ownership. We find that the $R^2$ statistics generated by our regression methodology are much higher than that generated by the simple cutoff classification schemes that were previously employed. This remains true when we compare the out-of-sample $R^2$ statistics derived from our analysis to those from the cutoff classification schemes.

We also investigate changes in the trading patterns of institutions in response to movements in daily returns, volume and volatility. We find that periods of high returns appear to reduce institutional trading in the very smallest trade sizes. High volatility is accompanied by increased institutional trading intensity in large trade sizes, suggesting that institutions may have high demands for liquidity at such times. Finally, periods of high volume and quoted depth appear to generate greater institutional trading activity in small and medium size trades, suggesting that institutions may use periods of high market liquidity to increase stealth-trading activity.

# 6. Appendices

## 6.1. Appendix 1: Buy-Sell Classification

TAQ does not classify transactions as either buys or sells. To classify the direction of each trade, we use a matching algorithm suggested by Lee and Ready (1991). This algorithm looks at the trade price relative to quotes to determine whether a transaction is a buy or sell. The method works by matching trades to pre-existing quotes, based on time stamps. More precisely, we inspect quotes lagged by at least five seconds to avoid problems of stale reporting of quotes. If the trade price lies between the quote midpoint and the upper (lower) quote, the trade is classified as a buy (sell). If the trade price lies at the midpoint of the quotes, we use a tick test, which classifies trades that occur on an uptick as buys, and those on a downtick as sells. If the trade price lies at the midpoint of the quotes and the transactions price has not moved since the previous trade (trade occurs on a "zerotick"), Lee and Ready suggest classifying the trade based on the last recorded move in the transactions price. If the last recorded trade was classified as a buy (sell), then the zerotick trade is classified as a buy (sell). From Lee and Ready, trade-to-quote matching can be accomplished in 75.7% of trades, while tick tests are required in 23.8% of cases. The remaining trades take place outside the quoted spread.

The analysis in Lee and Radhakrishna (2000) evaluates the effectiveness of the Lee and Ready matching algorithm, using the TORQ database, which has buy-sell classified, institutional-individual identified data for 144 stocks over a 3 month period. They find that after removing trades with potentially ambiguous classifications (such as trades that are batched or split up during execution), the buy/sell classification algorithm is 93 percent effective. In particular, they find that the accuracy is highest (at 98 percent) when trade-to-quote matching can be accomplished, lower (at 76 percent) for those trades that have to be classified using a tick test, and lowest (at 60 percent) for those trades classified using a zerotick test. We eliminate this last source of variability in our data by terming as unclassifiable those trades for which a zerotick test is required. We further identify as unclassifiable all trades that occur in the first half hour of trading (since these come from the opening auction) as well as any trade that is reported as cancelled, batched or split up in execution. This last category of trades is identified as unclassifiable since we use trade size as one important input into our prediction of institutional ownership. A trade that is reported as being batched or split up cannot be unambiguously classified in terms of its size. We aggregate all unclassifiable trades together, and use the bin of unclassifiable trades as an additional input into our prediction exercise.

### 6.2. Appendix 2: Spectrum Institutional Ownership Data

A 1978 amendment to the Securities and Exchange Act of 1934 required all institutions with greater than $100 million of securities under discretionary management to report their holdings to the SEC. Holdings are reported quarterly on the SEC's form 13F, where all common-stock positions greater than 10,000 shares or $200,000 must be disclosed. These reports are available in electronic form back to 1980 from CDA/Spectrum, a firm hired by the SEC to process the 13F filings. Our data include the quarterly reports from the first quarter of 1993 to the final quarter of 2001. Throughout this paper, we use the term institution to refer to an institution that files a 13F. On the 13F, each manager must report all securities over which they exercise sole or shared investment discretion. In cases where investment discretion is shared by more than one institution, care is taken to prevent double counting.

Our Spectrum data have been extensively cleaned by Kovtunenko and Sosner (2003). They first identify all inconsistent records, those for which the number of shares held by an institution in a particular stock at the end of quarter $t-1$ is not equal to the number of shares held at the end of quarter $t$ minus the reported net change in shares since the prior quarter. They assume that the holdings data are correct for such observations, rather than the reported change data.

They proceed to fill in missing records, using the general rule that if a stock has a return on CRSP but does not have reported Spectrum holdings in a given quarter, holdings are set to zero. For the missing records inconsistent with this assumption (those for which holdings at the end of quarter $t$ are above the reported net change from previous quarter holdings), they fill in the holdings for the end of quarter $t-1$ as split-adjusted holdings in period $t$ less the reported net change in holdings.

The Spectrum 13F holdings file contains three columns: date, CUSIP code, identifier for the institution, and number of shares held in that stock by that institution on that date. All dates are end-of-quarter (March 31, June 30, September 30, or December 31). For each CUSIP and date we simply sum up the shares held by all institutions in the sample to get total institutional holdings of the security at the end of that quarter.

## References

[1] Barclay, Michael and Jerold B. Warner, 1993, "Stealth Trading and Volatility: Which Trades Move Prices?" *Journal of Financial Economics*, 34, 281-305.

[2] Chakravarty, Sugato, 2001, "Stealth-Trading: Which Traders' Trades Move Stock Prices?" *Journal of Financial Economics,* 61, 2, 289–307

[3] Chen, Joseph, Harrison Hong and Jeremy Stein, 2002, "Breadth of Ownership and Stock Returns," *Journal of Financial Economics*, 66, 171-205.

[4] Chen, Hsiu-Lang, Narasimhan Jegadeesh, and Russ Wermers, 2000, "The value of active mutual fund management: an examination of the stockholdings and trades of fund managers," *Journal of Financial and Quantitative Analysis,* 35, 343-368.

[5] Cohen, Randy, Paul Gompers, and Tuomo O. Vuolteenaho, 2002, "Who underreacts to cashflow news? Evidence from trading between individuals and institutions," *Journal of Financial Economics,* 66, 2-3, pp.409-462.

[6] Daniel, Kent, D., Mark Grinblatt, Sheridan Titman, and Russ Wermers, 1997, "Measuring mutual fund performance with characteristic-based benchmarks," *Journal of Finance,* 52, 1035-1058.

[7] Daves, Phillip, James Wansley, and Rongrong Zhang, 2003, "Reported Trading Volume on the NYSE and the Nasdaq," working paper, University of Tennessee.

[8] Gompers, Paul A., and Andrew Metrick, 2001, "Institutional investors and equity prices," *Quarterly Journal of Economics* 116, 229-260.

[9] Greene, Jason, 1995, "The impact of limit order executions on trading costs in NYSE stocks: an empirical examination" Indiana University Working Paper

[10] Grinblatt, Mark and M. Keloharju, 2000a, “The investment behavior and performance of various investor-types: A study of Finland’s unique data set,” *Journal of Financial Economics* 55, 43-67.

[11] Grinblatt, Mark and M. Keloharju, 2000b, “What makes investors trade?,” *Journal of Finance*, 56, 589-616.

[12] Grinblatt, Mark, Sheridan Titman and Russ Wermers, 1995, “Momentum Investment Strategies, Portfolio Performance, and Herding: A Study of Mutual Fund Behavior,” *American Economic Review*, 85, 1088-1105.

[13] Kovtunenko, Boris and Nathan Sosner, 2003, “Sources of Institutional Performance,” unpublished working paper, Harvard University.

[14] Lakonishok, Josef, Andrei Shleifer, and Robert Vishny, 1992, “The Impact of Institutional Trading on Stock Prices,” *Journal of Financial Economics*, 32, 1, 23-43.

[15] Lee, Charles M. C. and Mark J. Ready, 1991, “Inferring trade direction from intraday data,” *Journal of Finance* 46, 733-746.

[16] Lee, Charles M. C. and Balkrishna Radhakrishna, 2000,. “Inferring Investor Behavior: Evidence from TORQ Data,” *Journal of Financial Markets,* 3(2): 83-111.

[17] Lo, Andrew W. and Jiang Wang, 2000, “Trading Volume: Definitions, Data Analysis, and Implications of Portfolio Theory,” *Review of Financial Studies,* 13, 257-300.

[18] Nelson, Charles R. and Andrew F. Siegel, 1987, “Parsimonious Modeling of Yield Curves,” *Journal of Business* 60, 473-489.

[19] Nofsinger, John R. and Richard W. Sias, 1999, “Herding and feedback trading by institutional and individual investors,” *Journal of Finance* 54, 2263-2295.

[20] Shalen, Catherine T., 1993, “Volume, Volatility, and the Dispersion of Beliefs,” *The Review of Financial Studies*, 6, 405-434.

[21] Shao, J. 1989 "The Efficiency and Consistency of Approximations to the Jackknife Variance Estimators", *Journal of the American Statistical Association,* 84, 114-119

[22] Shao, J. and Wu, C. F. J, 1989 "A General Theory for Jackknife Variance Estimation," *Annals of Statistics*, 17, 1176-1197

[23] Wermers, Russ, 1999, "Mutual fund herding and the impact on stock prices," *Journal of Finance* 54, 581-622.

[24] Wermers, Russ, 2000, "Mutual fund performance: An empirical decomposition into stock-picking talent, style, transaction costs, and expenses," *Journal of Finance* 55, 1655-1695.

## Table I: Summary Statistics for Firm Size Quintiles

Table I presents means, medians and standard deviations for the TAQ and Spectrum variables in our specifications. All data are independently winsorized at the 2.5 standard deviation level. The variables are in sequence, the total buyer initiated orders in TAQ classified by the Lee and Ready algorithm; the total seller initiated orders, similarly classified; the total unclassifiable volume (those transacted in the opening auction, reported as cancelled, or unclassifiable as a buy or a sell by the LR algorithm); the total volume (the sum of the previous three variables); the net order imbalance (total classifiable buys less total classifiable sells); and finally, the change in quarterly 13-F institutional ownership as reported in the Spectrum dataset as a fraction of CRSP shares outstanding. All TAQ variables are normalized by daily shares outstanding as reported in CRSP, and then summed up to the quarterly frequency. All summary statistics are presented as annualized percentages (standard deviations are annualized under the assumption that quarterly observations are iid). The columns report these summary statistics first for all firms, and then for firm size quintiles, where firms are sorted quarterly by market capitalization (size).

| | **All** | **Small** | **Q2** | **Q3** | **Q4** | **Large** |
|---|---|---|---|---|---|---|
| **Mean** | | | | | | |
| ***TAQ Total Buys*** | 31.83 | 20.88 | 27.42 | 33.56 | 39.20 | 38.04 |
| ***TAQ Total Sells*** | 30.99 | 23.13 | 28.58 | 33.00 | 36.45 | 33.75 |
| ***TAQ Unclassifiable*** | 15.39 | 9.66 | 13.21 | 15.94 | 18.85 | 19.29 |
| ***TAQ Total Volume*** | 78.31 | 53.82 | 69.28 | 82.62 | 94.61 | 91.14 |
| ***TAQ Net Imbalance*** | 0.96 | -2.13 | -1.08 | 0.63 | 2.87 | 4.49 |
| ***Spectrum Change*** | 0.60 | -1.31 | 0.29 | 1.73 | 1.49 | 0.77 |
| **Median** | | | | | | |
| ***TAQ Total Buys*** | 23.84 | 13.72 | 18.76 | 24.85 | 31.40 | 30.58 |
| ***TAQ Total Sells*** | 23.90 | 15.80 | 20.63 | 25.58 | 29.90 | 27.41 |
| ***TAQ Unclassifiable*** | 11.57 | 5.70 | 8.73 | 11.47 | 15.04 | 15.81 |
| ***TAQ Total Volume*** | 60.42 | 36.43 | 49.26 | 63.03 | 77.00 | 74.35 |
| ***TAQ Net Imbalance*** | 0.55 | -1.22 | -0.62 | 0.15 | 1.62 | 3.09 |
| ***Spectrum Change*** | 0.43 | -0.03 | 0.41 | 1.65 | 1.35 | 0.98 |
| **Standard Deviation** | | | | | | |
| ***TAQ Total Buys*** | 13.48 | 11.00 | 13.04 | 13.84 | 14.23 | 12.82 |
| ***TAQ Total Sells*** | 12.40 | 11.32 | 12.55 | 12.81 | 12.83 | 11.30 |
| ***TAQ Unclassifiable*** | 6.79 | 5.69 | 6.69 | 7.00 | 7.03 | 6.21 |
| ***TAQ Total Volume*** | 31.68 | 27.09 | 31.27 | 32.62 | 33.06 | 29.46 |
| ***TAQ Net Imbalance*** | 5.07 | 4.87 | 5.07 | 5.22 | 5.15 | 4.23 |
| ***Spectrum Change*** | 8.94 | 7.48 | 9.40 | 9.87 | 9.59 | 8.03 |

## Table II: Summary Statistics for Bins and Firm Size Quintiles

Table II presents medians (top panel) and standard deviations (bottom panel) for total TAQ buys + sells classified by the Lee and Ready algorithm (normalized by firm shares outstanding). All data are winsorized at the 2.5 standard deviation level. All summary statistics are reported in annualized percentage terms (standard deviations are annualized under the assumption that quarterly observations are iid).

| **Median** | **Small** | **Q2** | **Q3** | **Q4** | **Large** |
|---|---|---|---|---|---|
| **Buys + Sells** | | | | | |
| **0-2000** | 2.126 | 0.680 | 0.163 | 0.000 | 0.000 |
| **2000-3000** | 1.727 | 0.728 | 0.409 | 0.166 | 0.000 |
| **3000-5000** | 3.438 | 1.675 | 1.004 | 0.700 | 0.319 |
| **5000-7000** | 2.597 | 1.738 | 1.045 | 0.691 | 0.405 |
| **7000-9000** | 1.943 | 1.623 | 1.051 | 0.713 | 0.385 |
| **9000-10000** | 0.793 | 0.710 | 0.512 | 0.330 | 0.139 |
| **10000-20000** | 4.902 | 6.038 | 5.025 | 3.555 | 1.954 |
| **20000-30000** | 2.056 | 3.351 | 3.526 | 3.070 | 1.831 |
| **30000-50000** | 1.784 | 3.934 | 4.584 | 4.692 | 3.215 |
| **50000-70000** | 0.544 | 2.378 | 3.085 | 3.353 | 2.630 |
| **70000-90000** | 0.000 | 1.628 | 2.352 | 2.737 | 2.194 |
| **90000-100000** | 0.000 | 0.586 | 0.998 | 1.180 | 0.988 |
| **100000-200000** | 0.000 | 3.562 | 6.626 | 8.567 | 7.630 |
| **200000-300000** | 0.000 | 1.329 | 3.411 | 5.059 | 5.064 |
| **300000-500000** | 0.000 | 1.139 | 3.623 | 5.941 | 6.554 |
| **500000-700000** | 0.000 | 0.000 | 1.853 | 3.452 | 4.116 |
| **700000-900000** | 0.000 | 0.000 | 1.067 | 2.237 | 2.814 |
| **900000-1000000** | 0.000 | 0.000 | 0.000 | 0.792 | 1.087 |
| **>1000000** | 0.000 | 0.000 | 4.018 | 9.543 | 14.178 |

| **Standard Deviation** | **Small** | **Q2** | **Q3** | **Q4** | **Large** |
|---|---|---|---|---|---|
| **Buys + Sells** | | | | | |
| **0-2000** | 1.905 | 1.002 | 0.529 | 0.317 | 0.134 |
| **2000-3000** | 1.095 | 0.744 | 0.459 | 0.324 | 0.197 |
| **3000-5000** | 1.861 | 1.444 | 0.910 | 0.610 | 0.382 |
| **5000-7000** | 1.575 | 1.348 | 0.910 | 0.610 | 0.404 |
| **7000-9000** | 1.336 | 1.228 | 0.892 | 0.606 | 0.376 |
| **9000-10000** | 0.658 | 0.620 | 0.477 | 0.332 | 0.224 |
| **10000-20000** | 3.554 | 3.610 | 3.009 | 2.293 | 1.421 |
| **20000-30000** | 2.013 | 2.243 | 2.080 | 1.835 | 1.245 |
| **30000-50000** | 2.216 | 2.720 | 2.679 | 2.514 | 1.903 |
| **50000-70000** | 1.317 | 1.757 | 1.799 | 1.753 | 1.470 |
| **70000-90000** | 0.940 | 1.329 | 1.399 | 1.372 | 1.155 |
| **90000-100000** | 0.427 | 0.606 | 0.642 | 0.635 | 0.542 |
| **100000-200000** | 1.981 | 3.193 | 3.730 | 3.840 | 3.291 |
| **200000-300000** | 1.047 | 1.736 | 2.214 | 2.395 | 2.132 |
| **300000-500000** | 1.188 | 1.942 | 2.600 | 2.925 | 2.699 |
| **500000-700000** | 0.705 | 1.178 | 1.615 | 1.832 | 1.786 |
| **700000-900000** | 0.467 | 0.861 | 1.168 | 1.329 | 1.296 |
| **900000-1000000** | 0.134 | 0.385 | 0.530 | 0.582 | 0.570 |
| **> 1000000** | 1.949 | 3.481 | 4.655 | 5.791 | 5.865 |

## Table III: Regression Specifications on Total Buys, Sells and Net Flows

Table III presents estimates of several specifications, in which the dependent variable is the change in Spectrum institutional ownership as a fraction of shares outstanding. The first panel below presents the independent variables in rows: an intercept, the lagged level of Spectrum institutional ownership as a percentage of the shares outstanding of the firm, the lagged change in Spectrum ownership computed similarly, the total unclassifiable volume in TAQ, total buyer initiated trades and total seller initiated trades. The second panel uses the same first three independent variables, but uses total net flows (total buys less total sells) as the fourth independent variable. Different specifications use different combinations of these independent variables. Specifications D-F are the same as specifications A-C, except that they incorporate quarter-specific time dummy variables. t-statistics computed using the delete-1 jackknife method are reported in parentheses below the coefficients.

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| ***Intercept*** | 0.007 | 0.007 | 0.011 | 0.011 | | | | |
| | (23.905) | (23.841) | (34.623) | (34.161) | | | | |
| ***Lagged Spectrum Level*** | | | -0.014 | -0.012 | | | -0.014 | -0.012 |
| | | | -(17.715) | -(16.110) | | | -(17.583) | -(16.025) |
| ***Lagged Spectrum Change*** | | | | -0.065 | | | | -0.064 |
| | | | | -(10.304) | | | | -(10.125) |
| ***TAQ Unclassifiable*** | | 0.034 | 0.033 | 0.034 | | 0.004 | 0.001 | -0.001 |
| | | (2.305) | (2.234) | (2.322) | | (0.221) | (0.067) | -(0.039) |
| ***TAQ Total Buys*** | 0.362 | 0.354 | 0.375 | 0.382 | 0.361 | 0.360 | 0.381 | 0.389 |
| | (33.867) | (31.221) | (33.000) | (33.555) | (33.675) | (30.987) | (32.750) | (33.395) |
| ***TAQ Total Sells*** | -0.453 | -0.462 | -0.460 | -0.470 | -0.447 | -0.448 | -0.446 | -0.456 |
| | -(38.314) | -(37.819) | -(37.766) | -(38.532) | -(37.804) | -(36.317) | -(36.244) | -(36.967) |
| ***R-Squared*** | 0.046 | 0.046 | 0.052 | 0.056 | 0.045 | 0.045 | 0.050 | 0.054 |
| ***N*** | 63403 | 63403 | 63403 | 63403 | 63403 | 63403 | 63403 | 63403 |
| ***N(Firms)*** | 3334 | 3334 | 3334 | 3334 | 3334 | 3334 | 3334 | 3334 |
| ***Time Dummies?*** | No | No | No | No | Yes | Yes | Yes | Yes |

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| ***Intercept*** | 0.000 | 0.005 | 0.011 | 0.010 | | | | |
| | (0.791) | (19.474) | (33.319) | (32.792) | | | | |
| ***Lagged Spectrum Level*** | | | -0.016 | -0.014 | | | -0.015 | -0.013 |
| | | | -(20.760) | -(19.329) | | | -(19.713) | -(18.217) |
| ***Lagged Spectrum Change*** | | | | -0.063 | | | | -0.064 |
| | | | | -(10.003) | | | | -(10.042) |
| ***TAQ Unclassifiable*** | | -0.133 | -0.094 | -0.098 | | -0.134 | -0.097 | -0.101 |
| | | -(16.875) | -(11.302) | -(11.626) | | -(17.069) | -(11.664) | -(12.018) |
| ***TAQ Net Flows*** | 0.349 | 0.388 | 0.404 | 0.412 | 0.351 | 0.391 | 0.405 | 0.414 |
| | (32.465) | (35.230) | (36.814) | (37.456) | (32.573) | (35.409) | (36.868) | (37.587) |
| ***R-Squared*** | 0.032 | 0.042 | 0.049 | 0.052 | 0.032 | 0.042 | 0.049 | 0.052 |
| ***N*** | 63403 | 63403 | 63403 | 63403 | 63403 | 63403 | 63403 | 63403 |
| ***N(Firms)*** | 3334 | 3334 | 3334 | 3334 | 3334 | 3334 | 3334 | 3334 |
| ***Time Dummies?*** | No | No | No | No | Yes | Yes | Yes | Yes |

## Table IV: Size Quintile Specific Regressions of Spectrum Change on Total TAQ Flows

This table presents estimates of specification F from Table III, estimated separately for stocks sorted into market capitalization quintiles. The dependent variable in all specifications is the change in Spectrum institutional ownership as a fraction of shares outstanding. The first panel below presents the independent variables in rows: the lagged level of Spectrum institutional ownership as a percentage of the shares outstanding of the firm, the total unclassifiable volume in TAQ, total buyer initiated trades and total seller initiated trades. The second panel uses the same first three independent variables, but uses total net flows (total buys less total sells) as the fourth independent variable. All specifications incorporate quarter-specific time dummy variables. t-statistics computed using the delete-1 jackknife method are reported in parentheses below the coefficients.

| | **Small** | **Q2** | **Q3** | **Q4** | **Large** |
|---|---|---|---|---|---|
| ***Lagged Spectrum Level*** | -0.041 | -0.026 | -0.020 | -0.020 | -0.026 |
| | -(14.874) | -(12.160) | -(9.860) | -(8.959) | -(10.259) |
| ***Lagged Spectrum Change*** | -0.049 | -0.030 | -0.032 | -0.078 | -0.165 |
| | -(3.141) | -(2.285) | -(2.559) | -(5.722) | -(10.514) |
| ***TAQ Unclassifiable*** | -0.094 | 0.030 | 0.065 | 0.036 | 0.026 |
| | -(2.766) | (0.759) | (1.667) | (0.967) | (0.739) |
| ***TAQ Total Buys*** | 0.172 | 0.218 | 0.356 | 0.472 | 0.612 |
| | (6.517) | (7.989) | (13.823) | (17.979) | (21.808) |
| ***TAQ Total Sells*** | -0.234 | -0.308 | -0.471 | -0.550 | -0.712 |
| | -(9.348) | -(11.743) | -(17.442) | -(18.533) | -(22.246) |
| ***R-Squared*** | 0.082 | 0.047 | 0.060 | 0.067 | 0.105 |
| ***N*** | 12516 | 12609 | 12621 | 12732 | 12925 |
| ***N(Firms)*** | 1131 | 1357 | 1319 | 1161 | 735 |
| ***Time Dummies?*** | Yes | Yes | Yes | Yes | Yes |

| | **Small** | **Q2** | **Q3** | **Q4** | **Large** |
|---|---|---|---|---|---|
| ***Lagged Spectrum Level*** | -0.042 | -0.027 | -0.023 | -0.022 | -0.028 |
| | -(15.113) | -(12.997) | -(11.556) | -(10.469) | -(11.941) |
| ***Lagged Spectrum Change*** | -0.048 | -0.029 | -0.031 | -0.077 | -0.162 |
| | -(3.075) | -(2.213) | -(2.457) | -(5.628) | -(10.355) |
| ***TAQ Unclassifiable*** | -0.176 | -0.110 | -0.119 | -0.083 | -0.122 |
| | -(9.973) | -(5.734) | -(6.075) | -(4.469) | -(6.891) |
| ***TAQ Net Flows*** | 0.207 | 0.261 | 0.396 | 0.493 | 0.633 |
| | (8.854) | (10.768) | (16.425) | (19.428) | (22.911) |
| ***R-Squared*** | 0.081 | 0.045 | 0.057 | 0.066 | 0.102 |
| ***N*** | 12516 | 12609 | 12621 | 12732 | 12925 |
| ***N(Firms)*** | 1131 | 1357 | 1319 | 1161 | 735 |
| ***Time Dummies?*** | Yes | Yes | Yes | Yes | Yes |

## Table V: Estimates of Spectrum-TAQ Quarterly Predictive Regression

This table presents estimates from an equation relating quarterly Spectrum institutional ownership to TAQ for different size quintiles of stocks. Here, the dependent variable is the change in quarterly 13-F institutional ownership from Spectrum (as a fraction of firm shares outstanding). In order, the dependent variables are the lagged level of the Spectrum institutional ownership fraction, the total unclassifiable volume in TAQ, and Net Flows, which are the number of shares bought less shares sold traded within dollar cutoff bins from TAQ (normalized by CRSP daily shares outstanding, and then summed up to the quarterly frequency). All specifications incorporate quarter-specific time dummy variables. t-statistics computed using the delete-1 jackknife method are reported below coefficients in parentheses.

| | **Small** | **Q2** | **Q3** | **Q4** | **Large** |
|---|---|---|---|---|---|
| ***Lagged Spectrum Level*** | -0.040 | -0.027 | -0.024 | -0.021 | -0.026 |
| | -(14.469) | -(12.799) | -(11.885) | -(9.839) | -(10.879) |
| ***Lagged Spectrum Change*** | -0.053 | -0.044 | -0.055 | -0.097 | -0.184 |
| | -(3.417) | -(3.384) | -(4.615) | -(7.148) | -(11.894) |
| ***Total Unclassifiable*** | -0.156 | -0.086 | -0.077 | -0.085 | -0.099 |
| | -(8.789) | -(4.471) | -(3.829) | -(4.455) | -(5.296) |
| ***Net Flows*** | | | | | |
| *0-2000* | 0.832 | 3.639 | 4.103 | 1.893 | -1.254 |
| | (4.770) | (5.245) | (2.802) | (0.769) | -(0.233) |
| *2000-3000* | -0.378 | -0.193 | -3.966 | -3.724 | -4.485 |
| | -(1.552) | -(0.215) | -(2.480) | -(1.749) | -(1.452) |
| *3000-5000* | -0.459 | -2.361 | -1.810 | -2.508 | -1.966 |
| | -(3.155) | -(4.372) | -(1.694) | -(1.646) | -(0.873) |
| *5000-7000* | -0.349 | -1.576 | -3.285 | 1.025 | -4.804 |
| | -(2.297) | -(3.152) | -(3.214) | (0.601) | -(2.240) |
| *7000-9000* | -0.348 | -0.605 | -2.165 | -2.292 | -1.796 |
| | -(2.029) | -(1.252) | -(2.246) | -(1.411) | -(0.801) |
| *9000-10000* | -0.074 | -1.285 | -3.530 | -5.361 | 1.534 |
| | -(0.285) | -(1.747) | -(2.576) | -(2.312) | (0.455) |
| *10000-20000* | 0.052 | -1.084 | -1.169 | -2.084 | -3.027 |
| | (0.603) | -(4.827) | -(3.079) | -(2.891) | -(2.614) |
| *20000-30000* | 0.060 | 0.011 | -0.375 | -0.991 | 1.585 |
| | (0.521) | (0.046) | -(0.894) | -(1.451) | (1.181) |
| *30000-50000* | 0.263 | 0.283 | 0.286 | 0.662 | 0.107 |
| | (2.488) | (1.536) | (0.962) | (1.459) | (0.120) |
| *50000-70000* | 0.328 | 0.510 | 0.452 | 1.038 | 1.497 |
| | (2.590) | (2.358) | (1.357) | (2.110) | (1.992) |
| *70000-90000* | 0.606 | 0.777 | 0.525 | 1.021 | 1.603 |
| | (3.654) | (3.287) | (1.515) | (2.090) | (1.848) |
| *90000-100000* | 0.992 | 1.200 | 0.812 | 1.944 | -0.090 |
| | (3.583) | (3.272) | (1.612) | (2.641) | -(0.079) |
| *100000-200000* | 0.615 | 0.628 | 0.790 | 0.721 | 0.668 |
| | (6.467) | (5.409) | (5.364) | (3.492) | (1.846) |
| *200000-300000* | 0.312 | 0.937 | 1.049 | 0.935 | -0.096 |
| | (2.360) | (6.321) | (6.175) | (3.994) | -(0.262) |
| *300000-500000* | 0.198 | 0.471 | 0.915 | 0.958 | 0.338 |
| | (1.679) | (3.959) | (6.887) | (5.545) | (1.299) |
| *500000-700000* | 0.119 | 0.310 | 0.715 | 0.764 | 1.066 |
| | (0.699) | (2.182) | (4.638) | (3.811) | (3.776) |
| *700000-900000* | 0.463 | 0.353 | 1.118 | 1.041 | 2.319 |
| | (1.695) | (2.121) | (6.450) | (4.871) | (7.611) |
| *900000-1000000* | 2.004 | 0.442 | 0.473 | 1.188 | 3.720 |
| | (2.270) | (1.355) | (1.674) | (3.556) | (7.644) |

| | | | | | |
|---|---|---|---|---|---|
| *> 1000000* | 0.184 | 0.215 | 0.269 | 0.287 | 0.378 |
| | (2.730) | (4.989) | (7.165) | (7.115) | (8.187) |
| $R^2$ | 0.097 | 0.084 | 0.121 | 0.109 | 0.139 |
| ***N*** | 12516 | 12609 | 12621 | 12732 | 12925 |
| ***N(Firms)*** | 1131 | 1357 | 1319 | 1161 | 735 |
| ***Time Dummies?*** | Yes | Yes | Yes | Yes | Yes |

## Table VI: Evaluating the Lee-Radhakrishna Method Using Spectrum and TAQ

This table presents $R^2$ statistics for various specifications of the Lee-Radhakrishna regression of the change in quarterly 13-F institutional ownership as reported in the Spectrum dataset (as a fraction of CRSP shares outstanding *O*) $\Delta(S_{i,t}/O_{i,t})$ on the lagged level and change in institutional ownership, and net quarterly flows from TAQ (normalized by CRSP daily shares outstanding, and then summed up to the quarterly frequency ). Here $f_{i,t}(c)$, represents net flows *greater* than \$$c$, and $f_{i,t}(-c)$ represents net flows *less* than \$$c$. We estimate variants of: $\Delta(S_{i,t}/O_{i,t}) = \alpha + \phi(S_{i,t-1}/O_{i,t-1}) + \rho\left(\Delta(S_{i,t-1}/O_{i,t-1})\right) + \beta_{c1} f_{i,t}(-c1) + \beta_{c2} f_{i,t}(c2) + \varepsilon_{i,t}$.

The specifications in rows labeled $\alpha = \hat{\alpha}_t$ include quarter-specific time dummies. The specification is estimated separately for different size quintiles of stocks (in columns). Row headings indicate estimates of the $R^2$ statistic under different coefficient restrictions, for different values of the dollar cutoff levels $c1$ and $c2$. The second to last row shows the out of sample $R^2$ of the CRV method evaluated over the second half of the sample using coefficients estimated over the first half of the sample. The final row shows the one period ahead out of sample $R^2$ of the CRV method estimated from a rolling regression updated each period.

| $R^2$ | **Small** | **Q2** | **Q3** | **Q4** | **Large** |
|---|---|---|---|---|---|
| ***c1 =2,000; c2=5,000*** | | | | | |
| $\alpha = 0, \beta_{c1} = -1, \beta_{c2} = 1$ | -0.082 | -0.094 | -0.050 | -0.009 | 0.044 |
| $\alpha = \hat{\alpha}_t, \beta_{c1} = -1, \beta_{c2} = 1$ | -0.080 | -0.086 | -0.033 | 0.010 | 0.068 |
| $\alpha = 0, \beta_{c1} = \hat{\beta}_{c1}, \beta_{c2} = \hat{\beta}_{c2}$ | 0.059 | 0.027 | 0.042 | 0.055 | 0.081 |
| $\alpha = \hat{\alpha}_t, \beta_{c1} = \hat{\beta}_{c1}, \beta_{c2} = \hat{\beta}_{c2}$ | 0.067 | 0.042 | 0.062 | 0.070 | 0.100 |
| ***C1=3,000;c2=10,000*** | | | | | |
| $\alpha = 0, \beta_{c1} = -1, \beta_{c2} = 1$ | -0.052 | -0.072 | -0.029 | 0.003 | 0.052 |
| $\alpha = \hat{\alpha}_t, \beta_{c1} = -1, \beta_{c2} = 1$ | -0.049 | -0.062 | -0.011 | 0.021 | 0.075 |
| $\alpha = 0, \beta_{c1} = \hat{\beta}_{c1}, \beta_{c2} = \hat{\beta}_{c2}$ | 0.063 | 0.032 | 0.057 | 0.067 | 0.093 |
| $\alpha = \hat{\alpha}_t, \beta_{c1} = \hat{\beta}_{c1}, \beta_{c2} = \hat{\beta}_{c2}$ | 0.071 | 0.048 | 0.078 | 0.081 | 0.111 |
| ***c1=3,000;c2=20,000*** | | | | | |
| $\alpha = 0, \beta_{c1} = -1, \beta_{c2} = 1$ | -0.025 | -0.048 | -0.008 | 0.016 | 0.060 |
| $\alpha = \hat{\alpha}_t, \beta_{c1} = -1, \beta_{c2} = 1$ | -0.021 | -0.037 | 0.010 | 0.033 | 0.083 |
| $\alpha = 0, \beta_{c1} = \hat{\beta}_{c1}, \beta_{c2} = \hat{\beta}_{c2}$ | 0.065 | 0.037 | 0.063 | 0.071 | 0.096 |
| $\alpha = \hat{\alpha}_t, \beta_{c1} = \hat{\beta}_{c1}, \beta_{c2} = \hat{\beta}_{c2}$ | 0.073 | 0.053 | 0.084 | 0.085 | 0.114 |
| ***c1=3,000;c2=50,000*** | | | | | |
| $\alpha = 0, \beta_{c1} = -1, \beta_{c2} = 1$ | -0.001 | -0.026 | 0.014 | 0.032 | 0.072 |
| $\alpha = \hat{\alpha}_t, \beta_{c1} = -1, \beta_{c2} = 1$ | 0.005 | -0.014 | 0.032 | 0.046 | 0.091 |
| $\alpha = 0, \beta_{c1} = \hat{\beta}_{c1}, \beta_{c2} = \hat{\beta}_{c2}$ | 0.065 | 0.040 | 0.068 | 0.074 | 0.100 |
| $\alpha = \hat{\alpha}_t, \beta_{c1} = \hat{\beta}_{c1}, \beta_{c2} = \hat{\beta}_{c2}$ | 0.072 | 0.056 | 0.088 | 0.087 | 0.116 |
| ***c1=5,000;c2=100,000*** | | | | | |
| $\alpha = 0, \beta_{c1} = -1, \beta_{c2} = 1$ | -0.005 | -0.017 | 0.029 | 0.042 | 0.082 |
| $\alpha = \hat{\alpha}_t, \beta_{c1} = -1, \beta_{c2} = 1$ | 0.004 | -0.002 | 0.048 | 0.054 | 0.098 |
| $\alpha = 0, \beta_{c1} = \hat{\beta}_{c1}, \beta_{c2} = \hat{\beta}_{c2}$ | 0.059 | 0.038 | 0.073 | 0.078 | 0.105 |
| $\alpha = \hat{\alpha}_t, \beta_{c1} = \hat{\beta}_{c1}, \beta_{c2} = \hat{\beta}_{c2}$ | 0.067 | 0.056 | 0.095 | 0.091 | 0.119 |
| ***CRV Out of Sample*** | | | | | |
| *Q1 1997: Q4 2000* | 0.073 | 0.061 | 0.072 | 0.080 | 0.116 |
| *Rolling One Period Ahead* | 0.077 | 0.072 | 0.092 | 0.089 | 0.117 |

## Table VII: Estimates of Nelson-Siegel Function Coefficients

This table presents nonlinear least squares estimates of the Nelson-Siegel (1987) function that relates the change in quarterly 13-F institutional ownership from Spectrum (as a fraction of firm shares outstanding) to exogenous variables and TAQ flows. The exogenous variables in order are: the dependent variables are the lagged level of the Spectrum institutional ownership fraction, lagged change in the Spectrum institutional ownership fraction and the total unclassifiable volume in TAQ. The coefficients on flows in various bins (indexed by *Z*, the midpoint of the range of dollar trade sizes captured in the bin) can be recovered from the coefficients below, using the function:

$$\beta(Z) = b_0 + (b_1 + b_2)[1 - e^{-Z/\tau}]\frac{\tau}{Z} - b_2 e^{-Z/\tau}$$

All specifications incorporate quarter-specific time dummy variables. All t-statistics are computed following estimation of the nonlinear parameter $\tau$, using the delete-1 jackknife method. These are reported below coefficients in parentheses.

| | **Small** | **Q2** | **Q3** | **Q4** | **Large** |
|---|---|---|---|---|---|
| ***Lagged Spectrum Level*** | -0.040 | -0.026 | -0.022 | -0.017 | -0.025 |
| | -(14.445) | -(12.412) | -(10.878) | -(8.321) | -(10.277) |
| ***Lagged Spectrum Change*** | -0.050 | -0.039 | -0.049 | -0.095 | -0.180 |
| | -(3.205) | -(3.040) | -(4.138) | -(6.966) | -(11.543) |
| ***Total Unclassifiable*** | -0.152 | -0.062 | -0.038 | -0.048 | -0.077 |
| | -(8.547) | -(3.250) | -(1.959) | -(2.542) | -(4.163) |
| ***b0*** | 0.349 | 0.448 | 0.594 | 0.626 | 0.685 |
| | (10.704) | (16.022) | (24.270) | (24.391) | (19.362) |
| ***b1*** | 4.475 | 18.593 | 109.043 | 84.100 | -10.954 |
| | (6.880) | (8.695) | (8.266) | (3.984) | -(5.471) |
| ***b2*** | -8.652 | -31.156 | -145.007 | -118.802 | 8.951 |
| | -(8.296) | -(10.617) | -(9.799) | -(5.198) | (2.869) |
| ***Tau*** | 1000 | 998 | 501 | 501 | 5010 |
| $R^2$ | 0.091 | 0.073 | 0.107 | 0.096 | 0.126 |
| ***N*** | 12516 | 12609 | 12621 | 12732 | 12925 |
| ***N(Firms)*** | 1131 | 1357 | 1319 | 1161 | 735 |
| ***Time Dummies?*** | Yes | Yes | Yes | Yes | Yes |

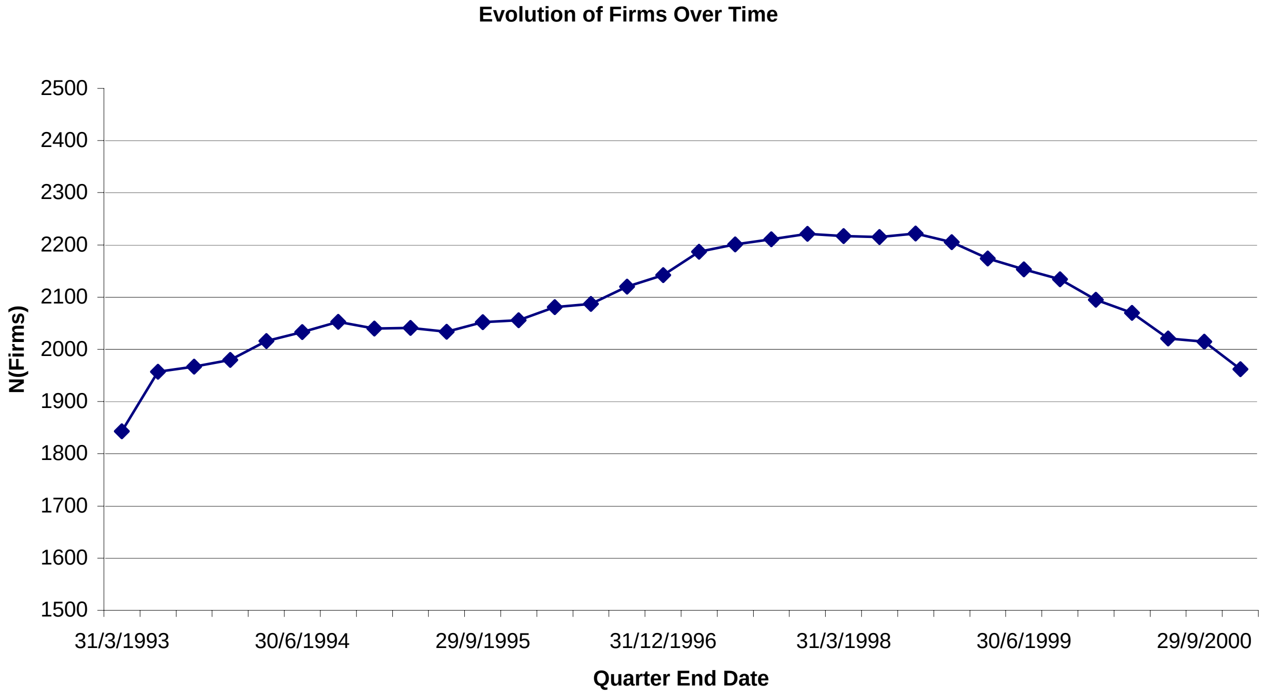
## Table VIII: Adding Interaction Terms to the Nelson-Siegel Specification

The first row of Table VII presents $R^2$ statistics for estimates of the baseline Nelson-Siegel specification relating the lagged level of quarterly institutional ownership, total unclassifiable volume and net flows in bins of different sizes (measured daily and aggregated up to the quarterly frequency) to the quarterly change in institutional ownership as measured in Spectrum. The rows below present in order, the $R^2$ from a Nelson-Siegel specification estimated using the same right hand side variables as above as well the variable itself (quarterly aggregated daily returns, absolute returns, volume and average quoted depth); the percentage change from the baseline $R^2$ upon addition of the variable; the $R^2$ from adding quarterly aggregated daily interactions between the flows in various bins and returns, absolute returns, volume or average quoted depth (as indicated by the headings); the marginal percentage increase in the $R^2$ from adding the interaction variables (over and above the variable itself) to the N-S specification; and finally the total percentage increase in $R^2$ from adding both the variable and the interactions. Columns present the results for the different size quintiles of stocks.

| | Small | Q2 | Q3 | Q4 | Large |
|---|---|---|---|---|---|
| **Baseline Nelson-Siegel $R^2$** | 0.091 | 0.073 | 0.107 | 0.096 | 0.126 |
| **Returns** | | | | | |
| N-S and Returns | 0.103 | 0.094 | 0.132 | 0.117 | 0.132 |
| Percentage Increase in $R^2$ | 13.91% | 29.45% | 23.58% | 22.16% | 5.18% |
| N-S, Interactions and Returns | 0.113 | 0.099 | 0.137 | 0.127 | 0.138 |
| Marginal Percentage Increase in $R^2$ | 9.08% | 5.36% | 3.75% | 8.66% | 4.10% |
| **Total Increase in $R^2$** | 24.25% | 36.39% | 28.22% | 32.74% | 9.48% |
| **Absolute Returns** | | | | | |
| N-S and Absolute Returns | 0.107 | 0.077 | 0.107 | 0.100 | 0.126 |
| Percentage Increase in $R^2$ | 17.60% | 5.59% | 0.12% | 3.71% | 0.45% |
| N-S, Interactions and Absolute Returns | 0.114 | 0.095 | 0.138 | 0.125 | 0.146 |
| Marginal Percentage Increase in $R^2$ | 6.80% | 23.19% | 29.34% | 25.61% | 15.30% |
| **Total Increase in $R^2$** | 25.60% | 30.08% | 29.50% | 30.28% | 15.81% |
| **Volume** | | | | | |
| N-S and Volume | 0.091 | 0.074 | 0.108 | 0.096 | 0.127 |
| Percentage Increase in $R^2$ | 0.80% | 1.17% | 1.00% | 0.07% | 0.60% |
| N-S, Interactions and Volume | 0.096 | 0.078 | 0.112 | 0.102 | 0.134 |
| Marginal Percentage Increase in $R^2$ | 5.28% | 6.70% | 4.28% | 6.72% | 6.20% |
| **Total Increase in $R^2$** | 6.12% | 7.94% | 5.32% | 6.79% | 6.84% |
| **Average Quoted Depth** | | | | | |
| N-S and Depth | 0.096 | 0.074 | 0.110 | 0.107 | 0.127 |
| Percentage Increase in $R^2$ | 5.52% | 1.72% | 3.03% | 11.88% | 0.54% |
| N-S, Interactions and Depth | 0.098 | 0.078 | 0.116 | 0.125 | 0.142 |
| Marginal Percentage Increase in $R^2$ | 2.29% | 5.35% | 5.22% | 16.76% | 11.92% |
| **Total Increase in $R^2$** | 7.93% | 7.16% | 8.41% | 30.63% | 12.52% |

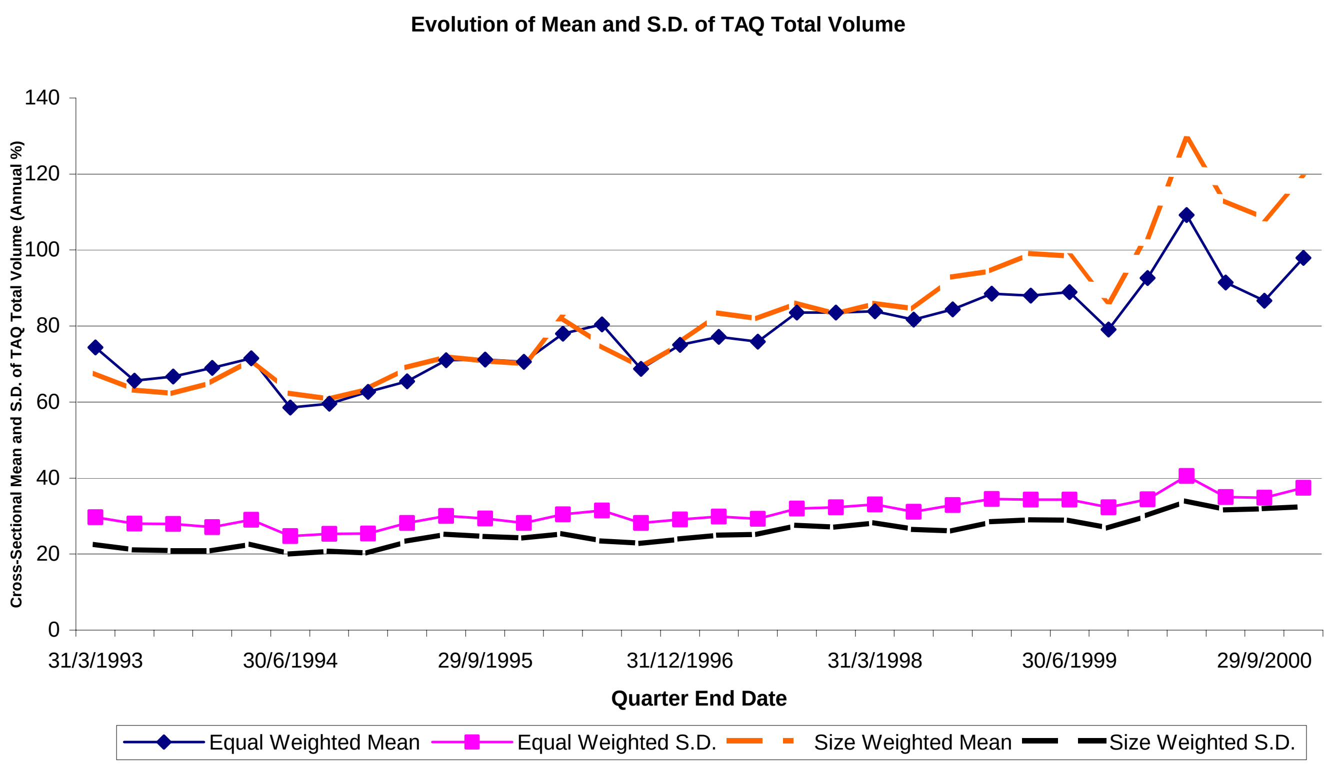
## Figure 1

This figure plots the evolution of the number of firms in our sample across time measured in quarters. The sample consists only of firms issuing common stock on the NYSE or AMEX exchanges. The data begin in the first quarter of 1993, and end in the final quarter of 2000.



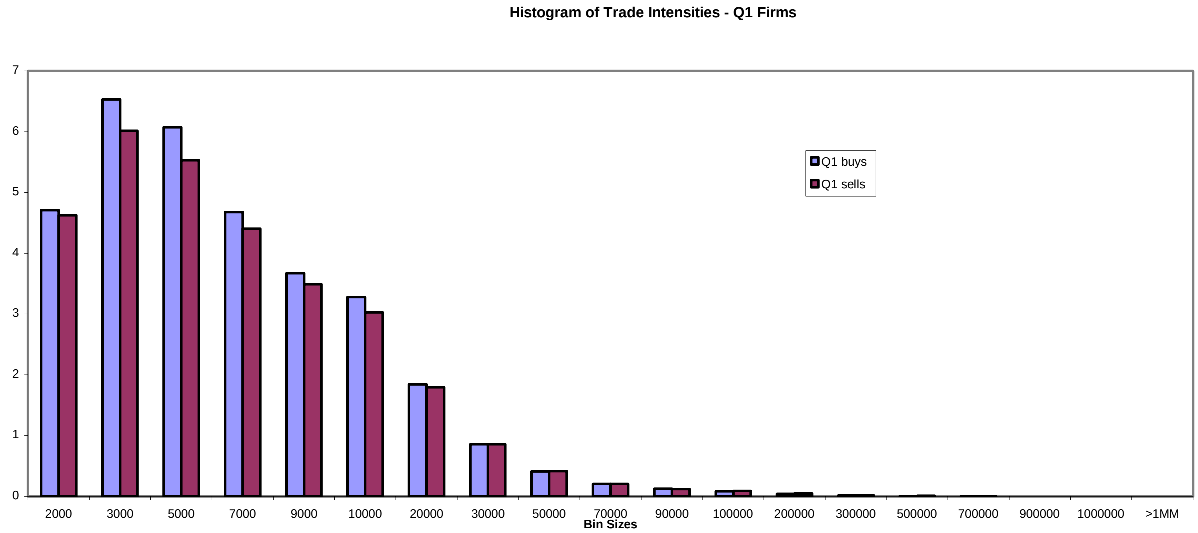

## Figure 2

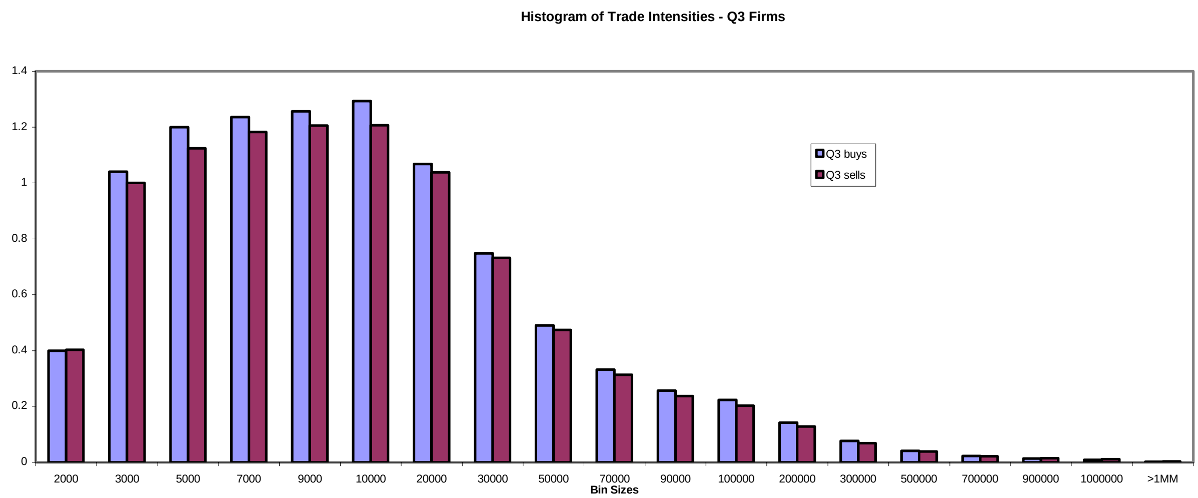
This figure plots the equal and market capitalization weighted means and standard deviations across all firms each quarter of the total volume of shares traded as a percentage of shares outstanding for each firm. The volume measure is obtained by summing all trades reported for each stock-quarter in the Transactions and Quotes (TAQ) database of the NYSE. Total shares outstanding and market capitalization for each firm is obtained from CRSP.

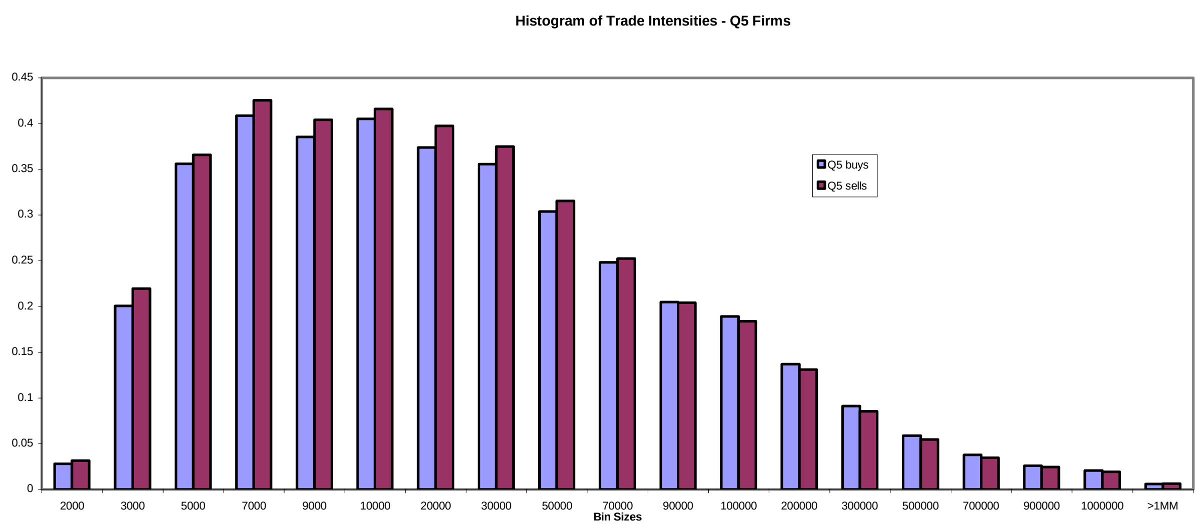


# Figure 3

Figure 3 plots histograms of trade intensities (total volume as a percentage of shares outstanding in each bin divided by relative bin width), for dollar trade size bins that aggregate TAQ trades classified into buys and sells. A bin size of $5 million is assigned to the largest bin. The three panels show, in sequence, histograms for small, median and large firms sorted quarterly into quintiles based on relative market capitalization (size).

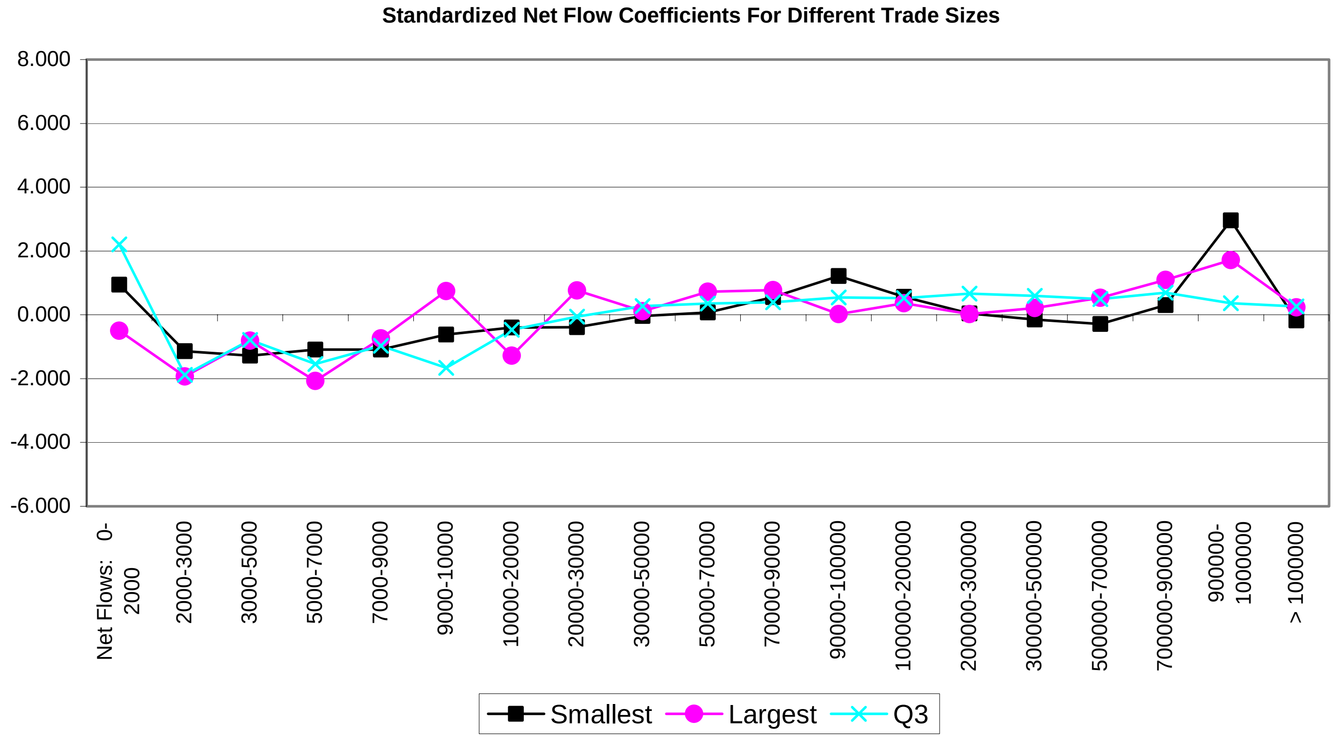


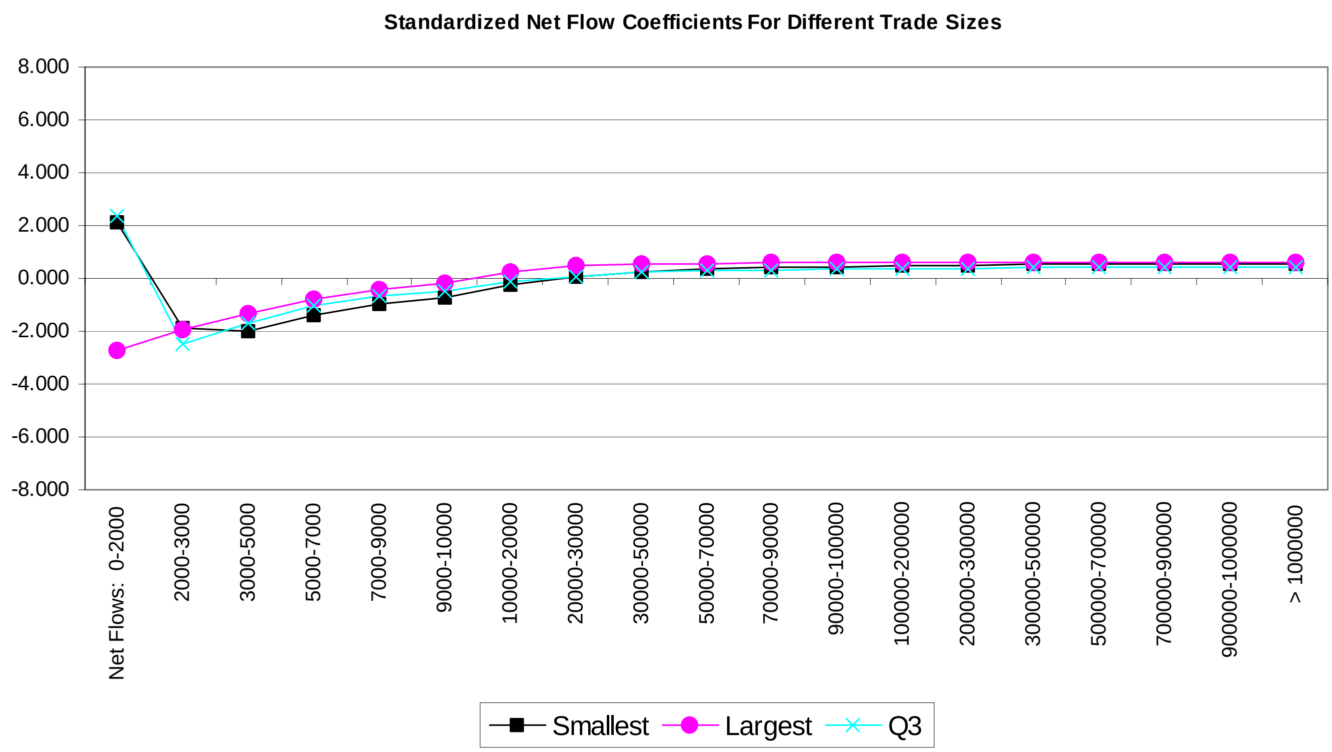
## Figure 4

This figure plots the net flow coefficients for each trade size bin, for the Q1, Q3 and Q5 firms in our sample. The coefficients are standardized by removing the within quintile cross-sectional mean of bin coefficients, and dividing by the cross-sectional standard deviation of bin coefficients.




## Figure 5

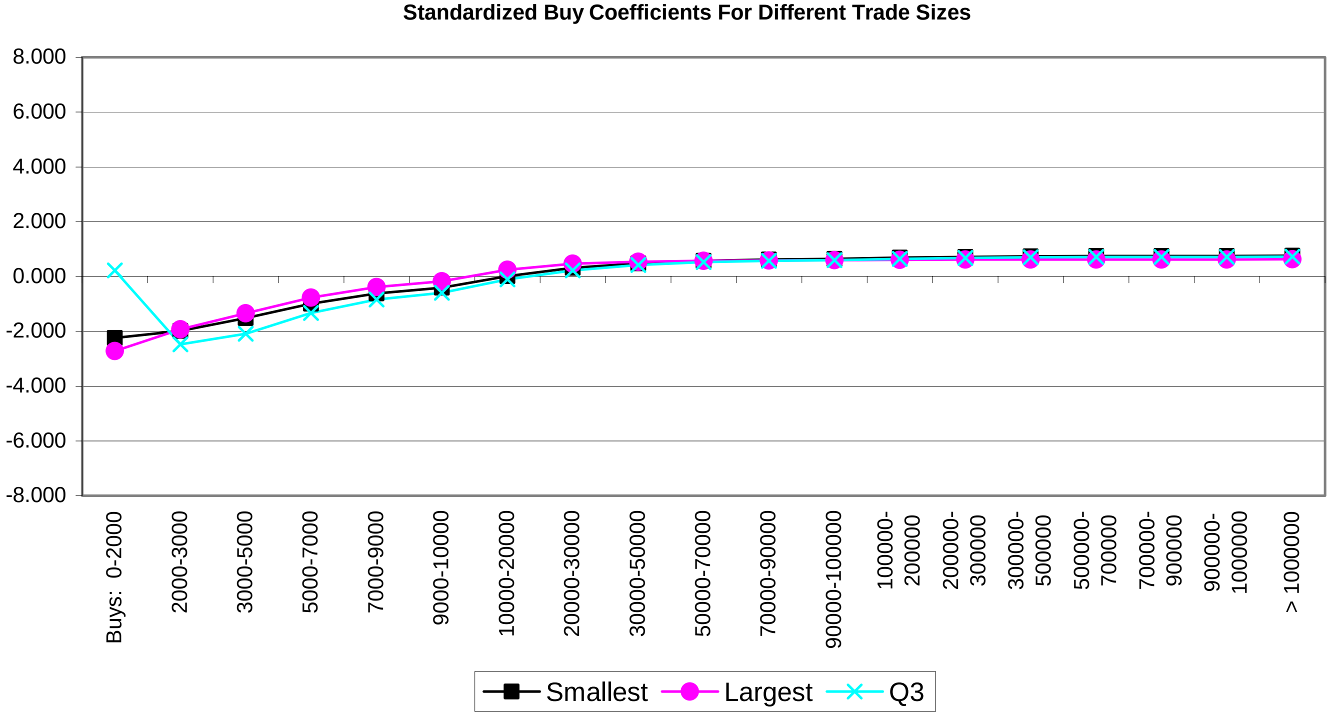
This figure plots the net flow coefficients estimated using the method of Nelson and Siegel [1987] for each trade size bin, for the Q1, Q3 and Q5 firms in our sample. The coefficients are standardized by removing the within quintile cross-sectional mean of bin coefficients, and dividing by the cross-sectional standard deviation of bin coefficients.



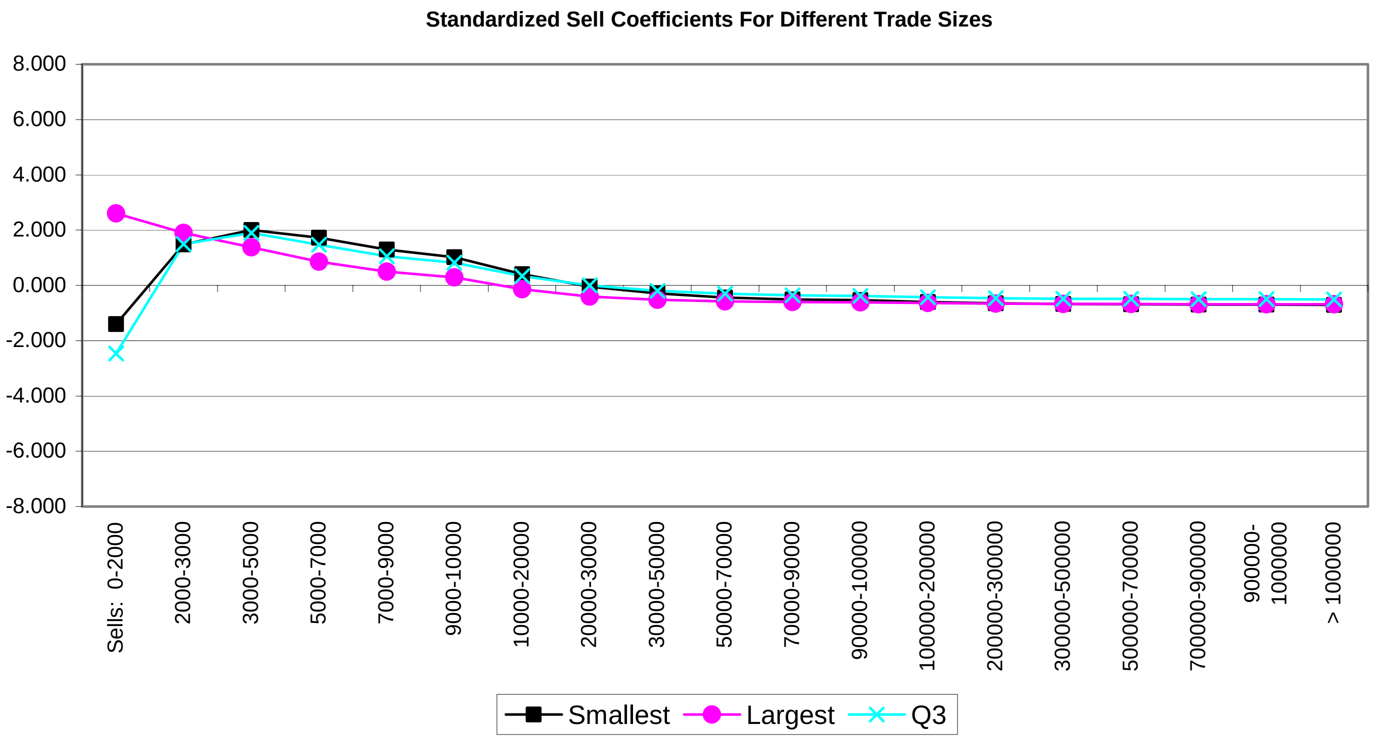

## Figure 6

This figure plots the buy and sell coefficients estimated using the method of Nelson and Siegel [1987] for each trade size bin, for the Q1, Q3 and Q5 firms in our sample. The coefficients are standardized by removing the within quintile cross-sectional mean of bin coefficients, and dividing by the cross-sectional standard deviation of bin coefficients. The top panel shows the buy coefficients, and the bottom panel the sell coefficients.

**Buys**

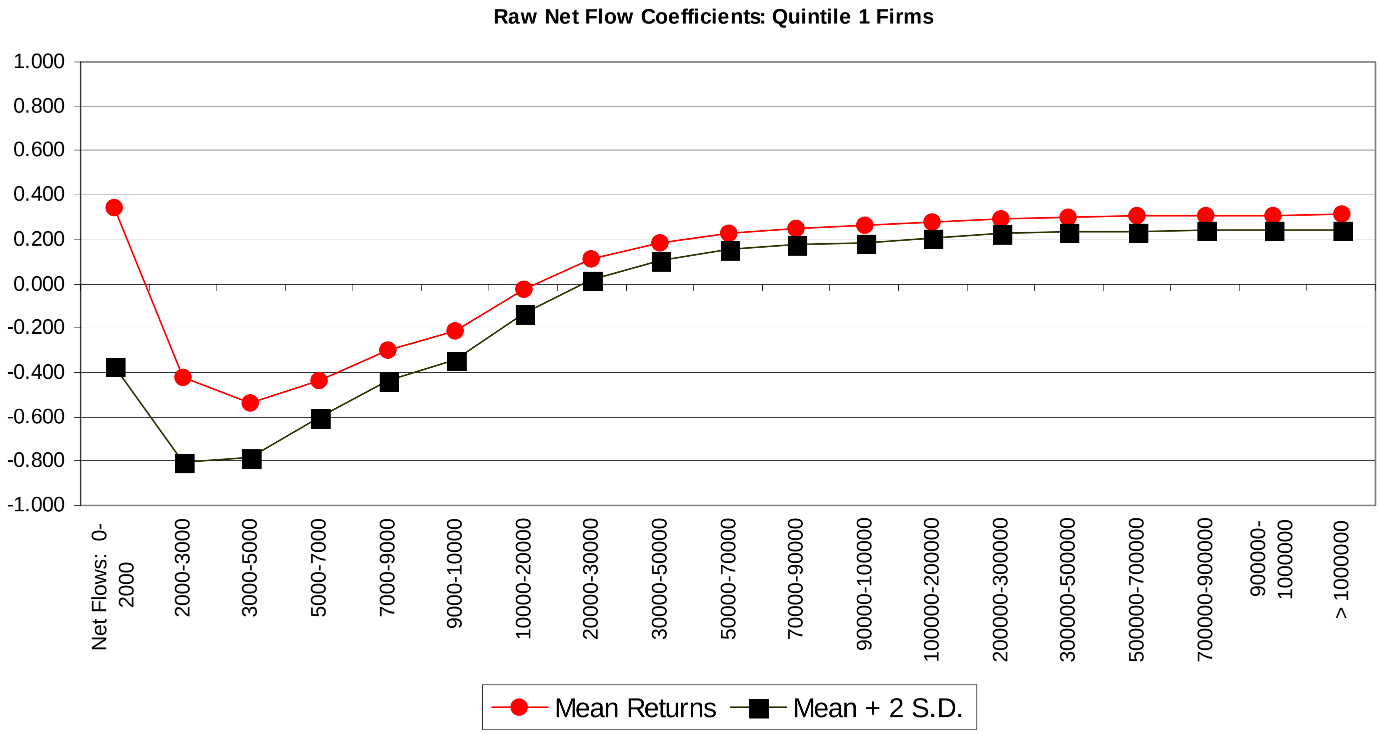



**Sells**



## Figure 7

This figure plots the net flow coefficients estimated using the method of Nelson and Siegel [1987] for each trade size bin, for the Q1 firms in our sample, setting the value of the return interaction to its daily mean and two standard deviations above its daily mean.




## Figure 8

This figure plots the net flow coefficients estimated using the method of Nelson and Siegel [1987] for each trade size bin, for the Q1 firms in our sample, setting the value of the absolute return interaction to its daily mean and two standard deviations above its daily mean.

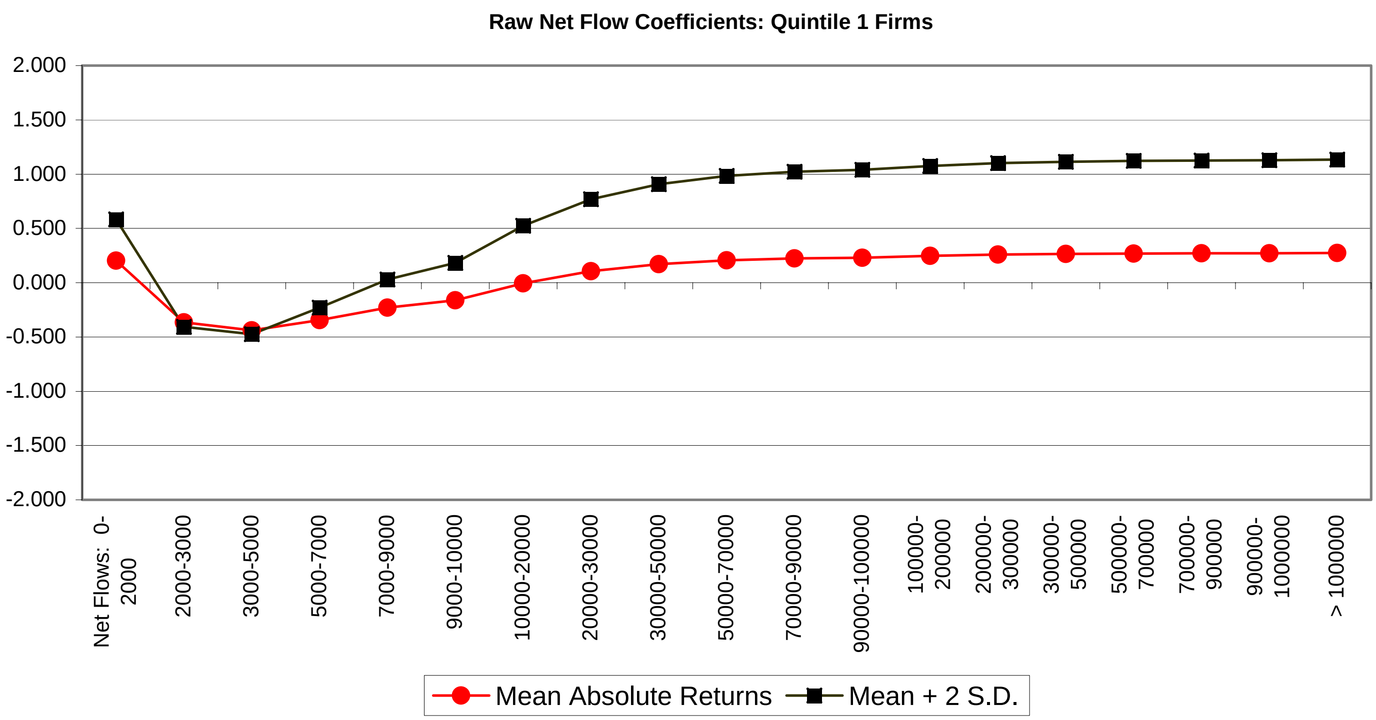


**Figure 9**

This figure plots the net flow coefficients estimated using the method of Nelson and Siegel [1987] for each trade size bin, for the Q1 firms in our sample, setting the value of volume to its daily mean and two standard deviations above its daily mean.

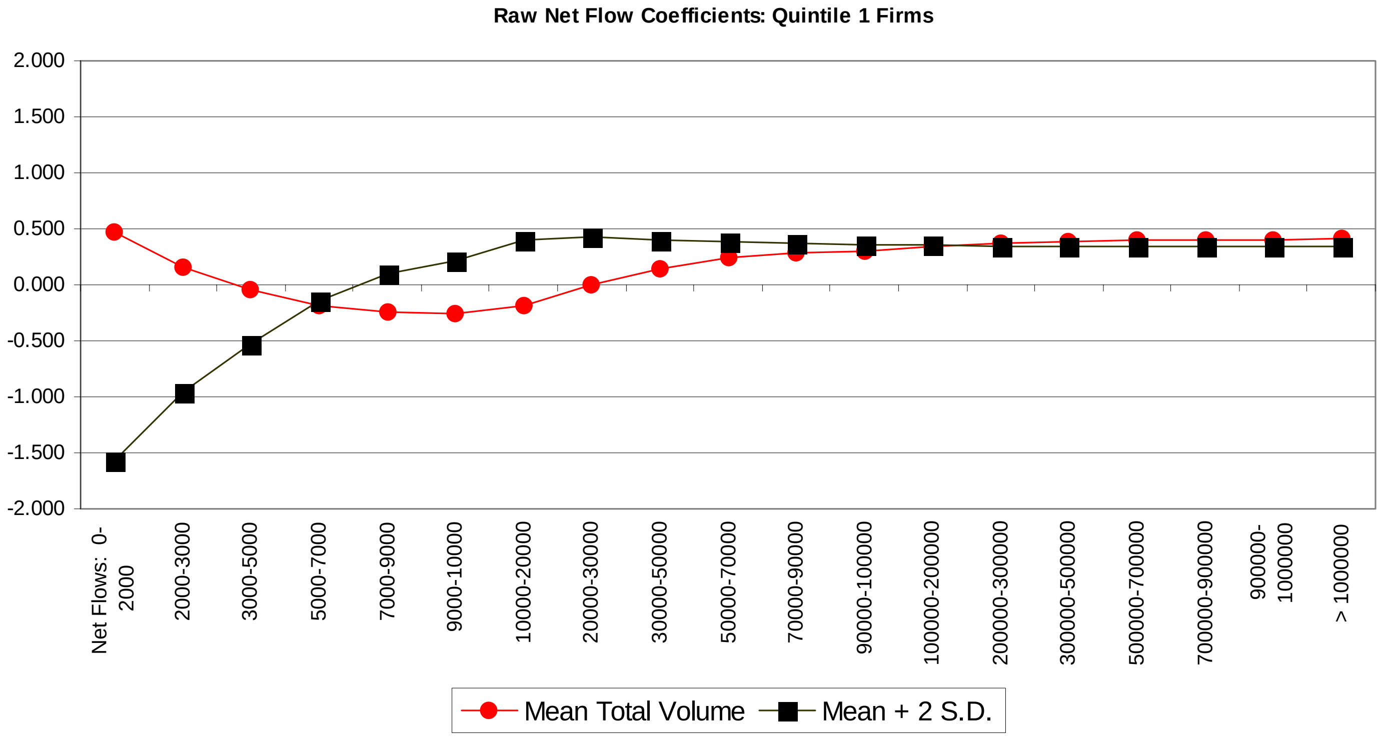



**Figure 10**

This figure plots the net flow coefficients estimated using the method of Nelson and Siegel [1987] for each trade size bin, for the Q5 firms in our sample, setting the value of average quoted depth to its daily mean and two standard deviations above its daily mean.