The Information Set of Individual Investors*

Toomas Laarits
NYU Stern School of Business

Jeffrey Wurgler
NYU Stern School of Business and NBER

October 13, 2024

Abstract

We examine the browser histories of a representative sample of individual investors to observe their interest in different types of information. The median individual investor spends two minutes of observable research on a stock before trading it, primarily just before the trade; the mean investor spends twelve minutes. Individual investors have little specific interest in traditional risk statistics and spend the most time reviewing price charts, followed by analyst opinions. Across stocks, research interest is correlated with stock size, and salient news can draw the attention of many investors. Across investors, research styles vary, and investors that focus on short-term information are more likely to trade high-volatility stocks. The results inform many theories of individual investor behavior.

* We thank Valerie Baldinger, Myeongrok Doh, and Cody Wang for outstanding research assistance and Terry Odean, Raghu Rau, Paul Tetlock, and participants at NYU Stern and the NBER Conference on Big Data and High-Performance Computing for Financial Economics for helpful comments.

I. Introduction

The information an investor uses to make decisions is a basic concept of investment theory. Asset pricing models make a wide range of assumptions about these "information sets." For example, standard models like the CAPM assume that investors know variances and covariances. De Long, Shleifer, Summers, and Waldmann (1990) assume that some investors know fundamental values and others trade based on irrelevant information. Hong and Stein (1999) assume that some investors follow the news and others react to price patterns. Microstructure models make still other assumptions. But in any plausible model, equilibrium prices depend on the trades and beliefs of investors, which in turn depend on the information to which they are exposed.

Information set concepts are just as central to understanding empirical results. Thousands of studies have examined how prices reflect past price information, other publicly available information, or private information, for example. And a large literature explores how attention to specific types of information, such as technical patterns or company announcements, affect investor trading patterns and thus their portfolios.

It is in general not possible to know much about the information sought, possessed, or used by an investor. Empirical tests of efficient markets shed only indirect light on what information investors employ, and likewise structural assumptions are required to back out beliefs based on portfolio choices. The panoply of frameworks and empirical results is unsurprising given the inability to observe the information environment.

In this paper, we use browser history data on hundreds of individual investors to shed light on the information they seek and obtain. Browser history data, also known as clickstream data, show the series of URLs visited by the user. Clickstream data are collected by a variety of

companies and are analyzed by, for example, online retailers who use it to influence how consumers navigate their websites. Our data provider makes extensive efforts to balance the panel against the U.S. households using the Internet at the time of the sample, which is early 2007. Bogan (2008) reports that even by 2002, over 80% of Schwab clients' daily trades were made online, and presumably this fraction was even higher several years later.

The data illuminate the nature of the information an individual investor sought and had at hand when trading, and thus the nature of the information that the trade could have incorporated into market prices and how the information shaped the investor's own portfolio in the process. It provides, a number of new and first-order insights about individual investor research interests and practices. We summarize them here and expect that different readers will be interested in different findings. For individual investors in our sample:

- The median household spends about two minutes on research on the ticker that they trade. The bulk of this research is done just before the trade.
- More research time is spent on price charts and then analyst-related information than other categories of information. Most trades are preceded by a snapshot page which includes a range of statistics, but these may be presented by default by the broker and many investors don't display active interest beyond this summary page. There is almost no explicit interest in standard portfolio statistics such as beta or volatility.
- Individual investors have different research styles. Some spend more time on stock research than others; some review more stocks while others concentrate; and some focus on indicators of short-term prospects while others on long-term fundamentals.

- Investors' research styles connect to the characteristics of stocks that they trade. Investors who have a high level of research interest in general, and a specific interest in short-term information, tilt their research interest towards high volatility stocks.

The current paper—an overall description of the types of information that individual investors seek, what they care about most, and by implication what they essentially ignore—is an initial step in a broader agenda using browser data. There are natural follow-up questions. We mention two. First, how does a given investor get from what she *observes* to what she *trades*? If an investor observes a particular price pattern in a chart, for example, how does that help her decide between buying, selling, or ignoring a particular holding? To the extent that information feeds into trading, browser data can shed light on how individual investors' portfolios are formed, because at a mechanical level portfolio composition is the cumulative outcome of past trading.

Second, what is the "right" amount of research for individual investors? One view is that investors should spend almost no time analyzing stocks; if they insist on trading individual stocks, they should take a dartboard approach and trust in market efficiency to ensure average returns. An alternative view, apparently appealing to many individual investors, is that there are returns to information gathering that justify the effort, and depending on the relationship between effort and performance there is a nonzero optimum. This becomes the proverbial empirical question: How does the depth and type of *research* that investor conducts relate to the *performance* of her trades? Is it really true that "more research is needed"?

The paper proceeds as follows. Section II describes the nature of the data and the sample of investors (investor-households). Section III discusses how we tabulate stock research. Section IV reviews average total research time, most-used domains, and stocks of most interest. Section

V tabulates the information of most interest. Section VI considers the timing of research vis-à-vis trading. Section VII characterizes research styles. Section VIII concludes.

II. Browsing data and sample

We begin with an introduction to the data set, including a stylized example, and discuss its unique advantages and remaining limitations. We then describe the specific household ("individual investor") sample.

A. Data source

The clickstream data we use was introduced in a legal context by Bakos, Marotta-Wurgler, and Trossen (2011) and Marotta-Wurgler (2011, 2012) and in a law and finance context by Laarits, Marotta-Wurgler, and Wurgler (2024). It includes the browsing behavior of tens of thousands of U.S. households across January, February, March, and June 2007. The data were made available by a major online research company. To paraphrase Bakos et al. (2011), the panel of households, recruited using random-digit-dialing distributions, installed a plug-in that collected the timing and sequence of URLs (link addresses) pages visited. All confidential or personally identifiable data, such as account numbers, addresses, passwords, and the like, had been removed by the data provider. Our focus is on the subset of households that trade individual stocks.

Inclusion in the data set is voluntary, but the data provider makes extensive efforts to reduce selection biases. Subject to any remaining selection bias and the noise from sampling, there is no reason to think that our results would not give a reasonably representative picture of the broader sample of U.S. individual investors trading online at the time.[1]

[1] As of 2002, there remained a small fraction of individual investors still calling in orders to Schwab over the phone or visiting branch offices (Bogan (2008)); this fraction likely continued to diminish over the five years before our

A stylized example based on raw data illustrates the level of detail that can be contained in link addresses. Table 1 reports fourteen minutes of a browsing session for a particular investor (to be precise, a particular household). A browsing session may begin at any link. This particular investor is a motorsports fan but soon switches to CNBC.com, where she clicks a link providing current data on several U.S. and international stock market indices, the US 10-year Treasury yield, and the USD-Euro exchange rate.

After a minute at that page, the investor logs into a brokerage account.[2] Within her online broker's research pages, she seeks out the day's most active names, which on that day included ImClone (IMCL) due to good news for its cancer drug's prospects (based on our own investigation). She consults the quarterly earnings performance for IMCL. From the link, we can infer that earnings were presented for the prior three years and estimates were shown for the next two years. In a separate panel, the chart showed the daily volume for IMCL over the prior three years, as well as a 13-day moving average. Further analysis of IMCL took place on the highly popular Yahoo Finance website. The investor obtains a quote on symbol ("?s=") IMCL and then looks up analysts' estimates ("ae") for same.

Following a check-in with race results, the investor returns to her brokerage's page and takes a different tactic. This time, she looks for trading ideas through a stock screener, in particular stocks with expected EPS growth of at least fifty percent over the next fiscal year. This would have yielded many results, Google (GOOG) among them. A click or two later, the investor comes to a snapshot page, which presents many types of financial and price information on the given stock in an abbreviated form. After looking at a price chart, the investor enters a

sample begins. We speculate that by 2007, investors who still hadn't transitioned to online trading were generally less active, less sophisticated, and less interested in formal stock research.

[2] The data provider scrubbed links to eliminate account numbers, personal information, email addresses, etc.

market order to buy ("ordertype=1") Google shares. Then she logs into e-mail and proceeds with other activities.

While every domain has a different directory layout and requires tedious processing to extract all the available information from query strings, as described in the Data Appendix, this example illustrates how the data can provide rich detail about the "research" process of individual investors. As a first step in analyzing this data set, the current paper documents the information investors care most about, the stocks they focus on, and the heterogeneity in research styles that they employ. By implication, the data also identify the categories of information that are ignored despite being a mouse click away.

Of course, there are limits to what we can observe. We do not directly observe what investors know or learn from sources outside these four months of Internet browsing. For example, we don't observe any information the investors gather offline. However, the impact of this limitation is greatly diminished by the fact that, as of 2007, the Internet was not just the easiest but often the only practical way to gather most of the data items that we track. Metrics such as consensus earnings, beta, insider selling, institutional ownership, recent price trends, the day's most active stocks, forward P/E ratios were easily accessible through free, continuously updated, one-stop-shop data feeds just a click or two away from the Buy and Sell buttons. Consider the wide range of information sought and obtained instantly by the investor in Table 1. Such convenience makes the Internet the obvious venue for stock research, at least for investors with the minimal level of sophistication needed to trade stocks online.

We also cannot determine an investor's accumulated "stock" of knowledge as opposed to the flow of information captured in the clickstream. In a stock-specific trading context, this is often not a major issue because many of the statistics consulted by our investor in Table 1 would

be useless if stale. An investor interested in one-week lookback price charts cannot store such information for future trades. On the other hand, dividend yield information, once learned, is more durable. One way to investigate these issues to see how research "ramps up" prior to the trade; if an investor was relying heavily on unobserved information, we might not expect a sudden concentration of research on that stock prior to the trade (but that is what we observe in most cases).

These limitations are most significant when it comes to intangible details about product lines or corporate reputation, as well as word of mouth from friends and family. Sentiments like "Apple makes cool products, and a growing number of young people seem to be using them" are abstract but may be relevant in cases where investors are also customers or have come in contact with the product line. Unobserved, informal information gathering is less likely to be important for investments in defense contractors, oil drilling equipment manufacturers, and indeed most of the thousands of exchange-traded stocks.

Other limitations may cause browser data to exaggerate rather than understate the scope of investor information. Most significantly, we cannot determine how well investors actually understand the information they access. If they struggle to interpret an income statement or a detailed analyst report, the time spent perusing such data may give the impression that they value these items more than they actually do. Furthermore, without the use of eye-tracking software, we cannot observe which specific parts of pages containing a mix of information, such as "snapshots" pages that list multiple statistics, capture an investor's attention. Our attribution of total page-viewing time across research categories becomes noisy when we must distribute it proportionally among the data items that were presented. This noise may average out somewhat

if different investors focus on different items of a multidimensional page, as the evidence of diverse research approaches indicates that they do.

Finally, spending only a few seconds viewing a bit of information is informative but it does not necessarily imply a lack of understanding. For example, investors familiar with the concept of beta will need only a few seconds to find that information for a given stock and, because it is a reasonably stable characteristic, it will relevant as long as the investor can remember it. Changes in analyst opinion, on the other hand, may not only take more time to find and understand, but it may have a short-term (perceived) investment relevance. Therefore, it's important not to assume that the time spent reviewing different types of information directly reflects their importance in making investment decisions. However, it is clear that if an investor has no knowledge of or hasn't looked into a particular type of information at all, it cannot influence their investment choices. For these reasons, it is important to consider not just the average amount of time an investor spends on a particular type of information, but the fraction of investors that is interested at all.

B. Identifying individual investors

We focus on households where we observe at least one trade in a specific ticker over the four-month sample; we will generally refer to households as individual investors, although in a few cases there may be more than one trader per household. Although some households may do some "research" yet not trade even once in four months, we regard them as much less interesting to study since they cannot be very influential. We focus on traders of individual U.S. stocks and do not include households who trade only options, mutual funds, international stocks, bonds, or ETFs. Research on such instruments is interesting but distinct. For example, options trading is

more short-term and event-based, and fund-level investing does not involve much company-specific research. We leave this to future work.

To identify trading households, we start with a list of online brokerages operating in 2007. Six of them feature prominently in our data. In the links to their domains, we search for words that suggest trading, such as buy, sell, ticker, order, trade, and so on, and proceed iteratively, manually inspecting each of thousands of potential links. Table 1 shows an example of a link that makes plain the details of a trade. (See the Data Appendix for additional detail.)

C. Sample of investors

Ultimately, our data involve 484 households which make 2,911 total U.S. stock trades over the four-month period. The raw data for these households includes around 8.5 million clicks over roughly 60,000 total hours of Internet use of any type. Individual households spend anywhere between one and four months in the data. In order to maintain comparability across households who may be in the sample for a different number of months, we report statistics on household-level monthly averages.

The top panel of Table 2 shows basic demographic information. Head-of-household income is topcoded at $100,000/year ($155,000/year in 2024 dollars). The median is between $75,000 and $100,000. The average age of the head of household is 50 years. Households are dispersed across the U.S. and have a representative distribution of other household characteristics including family size and life stage. Since the sample is roughly balanced with respect to household characteristics and trading is anonymous, these statistics are not so interesting, and we omit them for brevity.

Panel B reports general browsing statistics for our investor sample. To give a sense of the intensity of web use, we define a *browsing session* as a series of clicks with a break of no longer

than 15 minutes. The mean household in this sample carries out 103 such browsing sessions per month, with the median households coming in at 92 sessions. The average length of a trading session is about 26 minutes. During these browsing sessions, the average household visits three broker or other finance websites in a month. They spend over three and a half hours on brokerage sites, and another one and a half hours on other finance websites. For comparison, all other browsing time adds up to just over 37 hours per month, indicating that these households spend over 10% of their total online time on finance-related sites. Note too that the time spent at brokerage and other finance sites are right-skewed. The median household spends 1.4 hours per month at brokerage sites with the 95th percentile coming in at close to fifteen hours.

Panel C of Table 2 summarizes trading activity in individual stocks that we were able to identify from the clickstream data; we will now switch from "per household" to "per investor" terminology. The average investor in our sample trades in 1.33 browsing sessions per month; the median investor records only half a browsing session with a trade in the average month. Recall that investors are required to trade once in the four-month period to be included in our sample, hence the theoretical minimum value of per month browsing sessions with a trade is 1/4. A closely related measure is the number of stock trades. The average investor in the sample trades 2.2 times per month, while the median trader sees .67 trades per month. Trading activity is quite skewed: the 95th percentile household trades more than 9 times per month. The trades are roughly split between buys and sells. The mean trade size is $13,540 and the median is $2,020 among the 191 investors who use brokers where the URLs contain the number of shares traded (we determine the price from CRSP and TAQ data). The act of trading does not take a lot of time: the average trading portion of a session that contains some trading in individual stocks (i.e., a session that may include stock research, sports news, etc.) lasts a little under three minutes.

To restate, these averages are based on a total of 2,911 stock trades by 484 investor-households over the course of four months with individual households spending anywhere between one to four months in the data.

III. Tabulating stock research

A. Content categories

Asset pricing research often distinguishes between information about cash flows and information about discount rates (or expected returns). In reality, investment-related information is organized differently. Many sites provide a fairly common list of company-specific data items, and investors can then drill down into these items within the same domain or follow hyperlinks to other domains.[3] Given remaining heterogeneity in how websites present content, however, we need to determine an appropriate level of granularity for categorizing content. If we are too granular, we risk focusing on investors who, perhaps by chance, connect to brokerages that offer more detailed information, thus failing to capture a representative investor profile. Conversely, if we are too coarse, we miss out on the full richness of the data.

Balancing these considerations led us to identify the following major content categories: (1) risk statistics, (2) dividends, (3) earnings, (4) other fundamentals (typically valuation ratios or financial statements), (5) analysts' and other third-party analyses and opinions, (6) events (such as acquisitions or FTC actions), (7) price charts and price data, (8) specialized technical signals, (9) informed ownership (by insiders, funds, institutions, or short sellers), (10) company website visits, (11) industry data or competitor comparisons, and (12) market-level data (such as index fund tickers or market indices).

[3] See https://finance.yahoo.com/quote/AAPL for a current example.

We classify remaining links as NEC if the topic is of limited interest, or as indeterminate if the link constitutes stock research but lacks sufficient detail for precise categorization. Each category will be discussed in detail. Documenting how individual investors distribute their attention across these fundamental content categories is a unique opportunity provided by our data and a key contribution of this paper.

B. Content versus format

Note that our content categories are based on content, not format. Earnings information, for instance, can be presented as a bar chart, a news article, or a text press release. An analyst's opinion may be conveyed through text or summarized in a table. Past prices and returns represent content, while a chart or a table of prices represents different formats. Although links often suggest the presentation format, and the manner in which information is presented can significantly impact perception, as shown by extensive research on salience and framing effects, we focus on the content itself as the primary input for investment decisions. The analysis of formats is left for future research.

C. Content imputation

Investors frequently access detailed quote pages, often referred to as “snapshots” on various websites. These pages, available on every online broker and platforms like Yahoo Finance, serve as starting points for deeper research through other links and tabs. A snapshot page frequently marks the beginning, and sometimes the end, of an investor’s research on a particular ticker.

A snapshot page almost always contains current prices, daily percentage return, open/bid/ask, today’s volume, 52-week range, market cap, dividend or yield, a brief earnings statistic, and a price chart with a default lookback period. Beyond those statistics, practices vary

across brokers. For example, as of 2007, beta was occasionally included. Since almost no investors explicitly drill down into risk statistics, as we will see, they would encounter beta only incidentally, as one among many items on a snapshot page. Brokers may also differ in their use of lookback periods in price charts. Some snapshots include current headlines or detailed analyst forecasts, compare today's volume to a moving average, or report other fundamental data. Others are sparse or provide numerical data without meaningful context.

Given the widespread use of the snapshot format, we used the Internet Archive and other archival sources to identify the content categories represented on a given domain's snapshot page at the time of our sample. We established rules to allocate the time spent on a snapshot page across individual content types. Consequently, our tabulations include research time that can be assigned to dividends based on explicit details in the link and sometimes additional "imputed" research time, assigned with less confidence, to dividends based on their presence on snapshot pages. We distinguish between directed/explicit time and imputed time in our tabulations, with further details provided in the Data Appendix.

Other situations where imputations are required involve message boards and news articles for which we do not have headlines or coverage on the Internet Archive. For example, we may know that an investor may be scrolling through a stock-related message board, but we do not know the content itself. We view message boards as merely a form for presenting content rather than content itself. The Data Appendix explains how we use present-day Reddit r/stocks message board discussions to understand the typical topics of conversation on stock message boards, which we then impose on the message board time in our data.

News articles present a similar challenge. We may know a news article is about Intel but not whether it involves an earnings announcement or an analyst interview. In this case, we use

RavenPack's detailed content classifications to distribute time spent across content categories. The Data Appendix details this procedure as well. Again, due to the noise introduced by these imputation-based approximations, we distinguish between imputed time and explicit time in our tabulations below.

Table 3 summarizes our content-imputation techniques for these special situations. Snapshots vary from providing no analyst opinions or estimates at all to devoting up to one-third of the page to such content. Stock snapshots provided by some domains contain no dividends or earnings information at all, while others devote up to one-eighth of the page to these items. For any given website, the plurality and sometimes the majority of imputed time is attributed to price charts and price-related information, which are more visually prominent than numerical data items in small font. News reports often concern earnings and frequently contain price charts. Message board discussions most commonly focus on earnings.

IV. Total time spent on stock research, top sources, and top stocks

A. Total time spent on research

Table 4 presents a summary of the total online stock research activity conducted by 484 investors. We report statistics based on investor-level monthly averages, accounting for the fact that different households may be observed for varying durations, from one to four months.

The top row of Panel A indicates that, on average, investors engage in 29 browsing sessions per month that include some research activities. During these sessions, they view an average of 30 different stock tickers monthly. Each browsing session lasts an average of 3.5 minutes, amounting to close to two hours of research per month. Research efforts are notably right-skewed: the median investor examines 12 unique tickers monthly, whereas the 75th and

95th percentiles view 32 and 119 unique tickers, respectively; median monthly research time is 36 minutes, with the 75th percentile at 103 minutes and the 95th percentile at 520 minutes. Additionally, we report total research time per trade executed by the household, a normalization used extensively in subsequent tables. The average investor spends 138 minutes researching per trade, with the 25th percentile at 9 minutes and the 75th percentile at 107 minutes.

Panel B summarizes the non-research time investors spend on brokerage and other financial websites. On average, they spend over two hours per month on non-research activity such as logging on, browsing the broker's homepage, reviewing their portfolio, and so on.

B. Top domains used for stock research

Where do investors do most of their online stock research? Table 5 details the average number of minutes per trade across all research content categories that our trading households spend on different domains. We anonymize most of the specific domains to protect the intellectual property of the data provider, but we can establish that most research can be broken down into research on Yahoo Finance, brokerage domains, and other sites that present financial news or other financial statistics. We can also identify time spent on the SEC website and corporate websites of traded tickers.

We unmask Yahoo Finance, which has been highlighted by publicly available analysis of other samples to be the most popular finance research site from at least the date of our sample through the present.[4] It is also the single most important site by a substantial margin. Yahoo Finance reports financial data, news, analyst opinions, filings, and other information in a standardized format and does not require an account to use. Given its popularity, Yahoo Finance

[4] See https://ir.comscore.com/news-releases/news-release-details/yahoo-finance-ranks-top-financial-news-research-site-us-more-18 and https://www.semrush.com/website/top/united-states/finance/.

also appears at or near the top of any Google search involving a particular stock, which in turn perpetuates its dominance. The average of our 484 trading households spends 44 minutes per trade there on identifiable stock research; the 259 that actually use the site within our sample period spend a mean of 83 minutes and a median of six minutes per trade there.

The second set of columns shows that slightly over half of the time on Yahoo Finance can be matched to specific stock tickers. The third set of columns includes only research links matched to tickers traded by the household at least once, significantly reducing the time and number of households trading on information from the site. The gap reflects some combination of time spent researching stocks already owned but not traded within the sample (such stocks are unobservable by us) and time considering stocks that are never traded.

The next major sources of research are the brokerage sites themselves. Each investor in our sample trades through one of the indicated brokers. We exclude the time spent trading and and non-research activities at the same broker. Since investors typically use only their own broker for broker-domain research, the columns for households with nonzero time spent on that broker's website are the most useful, because they exclude non-account holders who couldn't access that broker's domain anyway. On average, investors that spend any time on research at their broker's website spend less than an hour per trade, with the median time in the single digits of minutes per trade. Many investors spend a minute or less per trade, with nearly every investor doing at least a few seconds of broker-site research over the four-month sample, although this may be passive in the sense that the broker may make it impossible for an investor to enter an order without being presented with at least a cursory quote page.

Proceeding down the list of top domains, 32 investors visited the SEC's website, typically following a link to an official statement provided by another domain. However, only 11 of these

investors obtained a document that we can attach to a company large enough to be in the CRSP sample. Note that we are able to count only the time spent on the SEC website, not the time spent reading any downloaded documents or documents opened in applications outside the browser. It may be the case that a few households do this, but the table maintains some perspective: Among 484 households and 2,911 trades, only one household reviewed information from the SEC website on a stock that they traded within the sample.

Another potential source of "research" is the websites of publicly traded U.S. corporations, specifically those matched to CRSP tickers. However, determining whether this browsing constitutes deliberate investment-focused research requires some constraints as well as detailed examination of the clickstream, as much internet activity on domains owned by public corporations (e.g., shopping on Amazon or reviewing bank balances at Bank of America) is not investment-relevant. If the clickstream itself is not sufficiently detailed about the investor's intent and/or when an investor accesses earnings numbers or financial statements on a corporation's website, in which case we categorize the time based on its content (e.g., earnings or fundamentals research) under the view that the domain for finding information is less significant than the content, we impose the requirement that a visit to a corporate website qualifies as "research" only when the same household clicks at least one link on another domain that implies investment-related research on the associated ticker. We additionally impose a running constraint that company website research cannot add up to more than two times the investor's overall research on that ticker.

Under these rules, 252 households in the sample consulted websites of publicly traded companies of apparent research interest to them. According to the rightmost column, 92 households that clicked on a company domain, conducted other research on the same stock, and

traded the stock in our sample, spent a mean of two minutes per month on corporate sites and a median of a few seconds.

C. Most-researched stocks

Another basic exercise is to examine the cross-section of research interest across stocks. In Figure 1, we present a scatterplot of stocks that are among both the top 100 stocks by market cap and the top 100 by the number of households that engaged in investment-related research during the sample period. At the very top is Apple, Inc. (AAPL), with more than *one-quarter* of investors researching Apple in just the four sample months. Again, this tabulation is not simply counting households that peruse Apple's corporate website for computer specifications or replacement cables, because as discussed above we do not count any time spent at a company website as "research" unless the same household also studied the company from another investment-relevant perspective. In general, there is a strong correlation between breadth of interest and market cap. For example, Microsoft (MSFT) ranks second in breadth of research interest and third in market cap. The top five companies by market cap all fall within the top ten in terms of breadth of interest.

Conversely, some stocks receive less research attention than their capitalization would suggest. For example, oil stocks tend to attract less research attention relative to their size. This might be because potential investors understand that the major driver of oil company profits is oil prices and view firm-specific research as less useful.

D. Research around events

Examining research activity around specific events provides insight into both the nature of our sample and actual investor research behavior. In Figure 2, we analyze interesting periods for three companies listed at the top of Table 6, comparing the breadth of research interest in our

data to search intensity measured by Google Trends. Importantly, there is no mechanical relationship between the two; Google search activity, like all search activity, is excluded from our research measures.

Microsoft is the second-most researched stock in our sample. Panel A looks at a month of activity that includes a quarterly earnings announcement. We plot research behavior in the clickstream data, measured as the number of investors engaged in investment-relevant research, alongside Google Trends search volume for the ticker symbol "MSFT" as a proxy for investment-focused search activity. There is a strong agreement between the two methodologies. The decline of interest on weekends is evident in both sources, and weekday variations roughly correlate as well. In both data sets, breadth of interest peaks around the quarterly earnings announcement.

This graph makes several key points. First, it confirms that our sample is representative of general investment-related interest: Google Trends is based on a very large random sample of almost all search activity, and the clickstream data sample, though smaller by orders of magnitude, qualitatively matches this activity. Second, the clickstream sample of a few hundred investors is large enough to detect daily fluctuations in research interest, at least concerning significant events among large companies. Third, our sample validates that Google Trends (on ticker symbols) can accurately measure time variation in investor interest and attention, which is not obvious due to its opaque normalizations.[5] This finding is of broad relevance because Google Trends data are available over the history of Internet usage and across search terms. Finally, there appears to be at least some incremental interest in earnings announcements of major companies; this is somewhat useful because one cannot otherwise distinguish whether the millions of

[5] For the same reason, it is not possible to directly use Google Trends data to measure absolute levels of investment-related interest, i.e., the number or fraction of Googlers interested in a particular term.

individual investors, or simply the thousands of institutional and professional investors, would be driving Google searches for, e.g., MSFT.

Apple is the most-researched company in our sample. Panel B shows a month of activity that includes both the announcement of the iPhone and a regular quarterly earnings announcement. We again observe a local maximum in clickstream data interest around the earnings announcement and another spike around the product announcement. Here, Google Trends data have the advantage that separate search terms can be used to isolate iPhone product interest from other investment-related interest. Clickstream data, on the other hand, have the advantage that one can identify and isolate households who do other investment-related research on Apple from the mass of households who are simply interested in a novel consumer product.

Our third example, in Panel C, involves Sirius, which Table 6 shows is an outlier with much greater research interest than its capitalization would suggest. This traces to a merger announcement during the sample period. Extrapolating the number of investors in our sample who "research" Sirius to the population implies a notable concentration of investor attention to this small stock around this specific event. The popular talk show host Howard Stern moved to Sirius in January 2006, potentially priming investor attention to events involving the company. Also, once again, investor interest as indicated in our data lines up with Google Trends.

V. Research interest in different types of information

We can now proceed to a central part of the analysis, which is a breakdown of the types of information our individual investors are exposed to and/or explicitly seek. The results are reported in Tables 7 and 8. In keeping with prior, domain-level results we report statistics on investor time spent in specific content categories, normalized by the number of trades the

investor carries out. We report means and medians across all 484 households, means and medians conditioning on the household doing some research in the category, as well as the number of investors doing research in the category. We further break out the results for all research by the household, and separately for tickers that the same household trades during the sample period.

As described in Section III C., a good amount of online research takes place on "snapshot" pages which are a combination of information belonging to different categories. In order to give a full sense of the composition of the investors' information diet, in Table 8 we replace content categories that are combinations of underlying content with imputed breakdowns. We organize the rows in Tables 7 and 8 in thematic groups and describe the results using the same structure.

A. Risk statistics

The standard academic advice for individual investors is simply to allocate a fraction of portfolio wealth into a diversified set of risky assets (or index) and the remainder in low-risk Treasuries. Each investor in our sample has eschewed that advice to some extent, since he or she is trading individual stocks. For such investors, academics offer the advice to at least be sure to distinguish alpha from beta. From an academic perspective, it is quite interesting to see the extent to which investors consult standard risk statistics such as beta and volatility.

Investors do not display much explicit interest in standard risk statistics. As shown in the first row of Table 7, only seven out of 484 investors displayed an explicit interest in beta or volatility; three of these drilled down into risk information on a ticker that they traded. The households would get some additional exposure to such risk metrics from snapshot pages, which sometimes include them. The most popular finance site in the sample, Yahoo Finance, as well as

two of the major brokerages, did not present any standard risk statistics on their snapshot page, so even the imputed time on standard risk statistics remains less than two minutes per trade dedicated to such content, as shown in the first row of Table 8.[6]

B. Hard fundamentals: Dividends, earnings, and other fundamental data

A second broad category of information that households may seek out is "hard" fundamentals: dividends, earnings, valuation ratios, and the accounting statements and regulatory filings that underlie these figures. We find more interest in this type of research. 199 households seek out some sort of earnings information, with around two minutes of research for the mean household, and 66 of the households study earnings of stocks they trade within the sample. Earnings information is prevalent on snapshot pages; consequently, almost all households have some imputed exposure to earnings-related information. Dividends attract at least some direct attention by 79 households, and those with that particular interest tend to spend relatively more time on research: the mean dividend-researching household spends close to 4 minutes per trade, and the median spends 2/3 of a minute, numbers comparable to those for earnings research.

A considerable number of households—255—seek out non-earnings, non-dividend fundamental information. This category includes valuation ratios such as price to sales, annual reports, and regulatory filings. In total, we see over four minutes of hard fundamentals research per trade.

C. Analysts

Analyst reports and estimates are a relatively popular category. 346 households pursue this type of research, making them second only to price charts, snapshots, and news in the number of households in this notion of popularity, and 131 households seek analyst information

[6] Yahoo Finance did not include beta to their snapshot page until 2012, according to the Internet Archive.

on the stocks they trade. The mean household spends over 12 minutes per trade on analyst reports, and conditional on doing any such research they spend approximately 18 minutes. Time spent on analyst reports is quite skewed, with the median household with that interest only carrying out 3 minutes per trade.

Taking into account the analyst-related content on snapshot and other imputed sites increases the importance of this category even further: with imputed content categories reported in Table 8 we see over 18 minutes per trade of analyst-related content. Investors tend to see analyst content on the stocks they trade: 293 households see any sort of analyst information on the stocks they trade.

D. Prices, price charts, and technical signals

By far the most prevalent types of pages the investors see are price charts and snapshots. Price chart refers to any sort of page that either just shows a chart of prices, or reports the price information in tabular form. Snapshot refers to a combination page of price information, as well as a host of other categories such as analysts or earnings, as described in Section III C.

The average household spends over 38 minutes per trade on price chart pages and another 37 minutes on snapshot pages. Nearly all of the households in the sample do some of this research: 448 and 451 out of 484 in the snapshot and price chart categories, respectively. Over 320 households in both categories do this kind of research on any of the stocks that they trade.

Most snapshot or price chart pages contain a price chart showing the evolution of the nominal share price over the prior month; only one broker and one finance website deviate from this standard, having one-year default lookbacks. Theory gives no guidance as to how far back investors should be looking into past prices. Intuitively, price charts provide some notion of risk, which is useful since we have already seen that investors are not seeking out detailed risk

statistics; but, for full context, a stock's price should be plotted against a market index, and that is generally not included in the default. Perhaps another plausible intuition is that investors with shorter investment horizons would be looking at shorter lookbacks.

Figure 3 shows the frequency of various lookback horizons that individual investors use. The top panel includes the page's default; over 80% of price chart consultations/observations do not deviate from the default lookback of one month. The bottom panel shows the frequency distribution of lookback periods when investors do opt out of the website's default. When doing so, investors tend to look further back to gather more historical information. They are most likely to opt for lookback windows between one month and one year, with only a few zooming into more recent data or zooming out beyond one year. This pattern seems consistent with the fact that most individual investors have an average holding period longer than one month, but the main takeaway is that most investors observe past prices only to the default lookback. It is intriguing to think about how the one-month default lookback could relate to the stylized facts of reversal within one month (Jegadeesh (1990)) and momentum at longer horizons (Jegadeesh and Titman (1993)).

A category closely associated with short lookback-window charts is technical analysis. We find that 147 households look at charts that include additional elements of technical analysis, such as moving averages, oscillators, Bollinger bands, or simpler return-based indicators such as 52-week highs. Overall time spent on technical analysis adds up to about 8 minutes per trade for the households who carry out any such research.

E. Company website visits

A potentially important source of stock market information is knowledge about the company's products. To capture this aspect of investor familiarity with the stock, we track time

the households spend on the URLs associated with tickers that they separately research on any of the brokerage or finance sites. This means that an investor who has a Bank of America account, shops on Amazon.com, or routinely uses search websites such as Google or Yahoo does not get "research credit" for these tickers unless there is some other evidence of investment-related interest into those tickers. Specifically, we impose the constraint that the running total of company website research time can be no more than twice the running total of all other research on the ticker.

We find that investors in general have reasonable exposure to the websites of companies they research. 289 households visit the website of some company that they also research in other respects. However, as a reflection of our 2x constraint, the overall time spent at company websites is quite modest, only 1 minute per trade. Easier to interpret is the number of households that have visited the website of a stock they trade: 96, or close to 20% of our investors.

F. Industry- and market-level data, ownership

The final thematic group pertains to industry- and market-level data. We classify as industry any sort of research on the firm's sector or competitors. We classify as market pages that give an overview of market-wide developments, such as index returns and market-wide news, we additionally define as "market" news any type of snapshot or price chart information sought on stock indices. Markets as a category attracts the attention of 215 households, with a rather skewed time distribution of 12 minutes on average, but only .7 minutes for the median household with nonzero time. Because of its definition, we never see "market" research on any traded stock. Industry attracts the attention of a similar number of households but is quite prevalent at snapshot pages and so 261 households are exposed to this type of information in the imputed content categories.

Ownership research includes information about short-selling, institutional or mutual fund ownership, or insider ownership. This groups together market participants that an average individual investor might suspect—usually, with good reason—to have superior information. 120 out of 484 investors in our sample consult such information explicitly, and 46 of them did so in a stock that they traded in the sample. The median investor who ever consults this information does so for under a minute per month; it does not take long to read the percentage of institutional ownership or short interest outstanding. Some type of ownership information is included on most major brokers' snapshot pages, although there was none on the Yahoo Finance site snapshot at the time. Since most trades confront the snapshot page, the imputed time on ownership data is substantially higher, but overall, this category remains among the least important in terms of research time share.

G. Indeterminate and NEC

Having classified the information into the above categories we are left with about 10 minutes per trade of research that is "indeterminate", meaning we are not able to ascribe any content category based on the URL we observe, and 5 minutes per trade that is "NEC" (Not Elsewhere Classified), where we can deduce the type of information but it's sufficiently rare that it doesn't warrant its own category. (We tabulated research time on risk statistics, despite its rarity, because of theoretical interest in that category.)

H. Summing up

In all, the average (mean) household carries out 138 minutes of research per trade, with the median household seeing 36 minutes of research per trade. Of the 138 minutes, just 11.5 minutes is matched to a ticker that the household trades within the sample period. The content

breakdown across traded and untraded stocks is summarized in Panel B of Figure 4. The overall content categories, both original and imputed, are reported in Panel A of Figure 4.

VI. Timing: Research in the run-up to a trade

How does research accelerate before a trade? Theory gives no general guidance here, but some investors clearly do respond to news to an extent, and some types of information that investors access quickly become stale, e.g., short-lookback price charts or the Sirius merger announcement in Figure 2. Other investment-relevant information, such as dividend policy, is more persistent and needs to be consulted less often. The temporal pattern of overall research relative to actual trading is another aspect of individual investor behavior that our data can characterize.

Although this is conceptually straightforward exercise, there are complexities in implementation. The simplest case is an investor that executes a single trade in a single ticker at t=0 and does no other trading throughout the other months of the sample. If so, it is straightforward to sum up research time related to the traded ticker, as well as untraded tickers, and associate all of that research to the trade. But in the more general case in which an investor makes multiple trades in the same ticker, or in multiple tickers, contemporaneously or across days, one must be careful to understand and disentangle potential double-counting issues.

In Figure 5, we summarize all ticker-matched research in the calendar week prior to a trade. Panel A starts by simply conditioning on a buy or sell trade at t=0 (of which we have N=1,431 buys and N=1,480 sells) and looks back one calendar week to calculate the cumulative amount of research that is matched to any ticker, as a function of time remaining to the next trade, regardless of whether it is regarding the ticker traded at t=0. Each vertical bar marks a 24-

hour period, and the periodic flat spots reflect afterhours pauses in research activity. We see an exponential increase in overall research time prior to a trade being executed; that is, the stock of research knowledge is rapidly augmented in the hours just prior to a trade. Roughly speaking, about half of the preceding week's ticker-matched research occurs in the 24-hour period prior to a buy, with a concentration in the minutes preceding the trade. Research prior to sells is somewhat more spread out. The mean cumulative research is about 20 minutes for buys and 15 minutes for sells.

The median behavior in Panel B looks different. The median household carries out no research at all until five days before a trade, and a considerable fraction of ticker-matched research that does take place occurs in the three or four minutes prior to the trade. Comparing the median and the mean behavior per trade implies that research time is skewed, and a considerable fraction of trades are preceded by at best a few minutes of ticker-specific research. Some sales may be driven by liquidity needs, but on average the cumulative pre-trade research time and pattern is similar to buys.

Next, we narrow in on the portion of this research that is attached to the specific ticker that was traded at t=0. The mean is in Panel C and median is Panel D. Given that market prices are impacted by trading, these pictures shed light on the amount and timing of stock-specific research that is injected into the market for a typical trade by an individual investor. We can again see research before a buy ramping up in the hour or two preceding a trade, with a spike in the minutes just before. In Panel D, we see that the median trade is preceded by only 2.5 minutes of stock-specific research. One can reasonably ask whether so little ticker-specific research could contribute much to a successful aggregation of new information into a stock's price—or, from

the investor's perspective, reflect a thoughtful approach to picking stocks and assembling a portfolio.

VII. Heterogeneity in research styles

The results so far document household interest in various types of information. We now seek to establish whether there are any cross-investor patterns to the types of research they carry out. To this end, we carry out principal components analysis on investor-level research intensity in different categories.

For this analysis we restrict the data to research matched to tickers, and we create indicator variables for any household-category that sees more than 2 minutes of research per month. We augment these category-level indicator variables with a number of other measures that capture salient aspects of an investors research style. This is a way of correlating research interests with other aspects of trading behavior and understanding more general research-investing styles.

First, we add dummy variables for those households that are in the top quintile in terms of trades they carry out, or in terms of tickers they have researched. We label these variables "High Trade" and "High Permno", respectively. We also include two Herfindahl index measures: one for the concentration among stocks researched ("HHI Research"), and one for the concentration among stocks traded ("HHI Trade"). Finally, we make use of the data we have on charts and classify those investors who spend more than 30% of their chart time with lookbacks less than one year as short lookback window investors ("Short Lookback"), while those spending more than 30% of their chart time on look-back windows of ten years or longer are "Long Lookback".

The first three principal components and the associated eigenvalues are reported in Table 9. The first three PCs account for 24%, 12%, and 8% of the variation in the data and have plausible economic interpretations. PC1 is a *level* factor. It loads roughly equally on all the content categories and has a negative weight on the two Herfindahl measures, indicating that a high PC1 loading reflects a household that researches across different categories of information and tends to cast a wide net in terms of tickers searched.

PC2 loads negatively on "High Trade" and "High Permno" on the one extreme and loads positively on "HHI Trade" and analysts and technical on the other extreme. We therefore conceptualize PC2 as a *focus-versus-breadth* factor: a negative loading captures investors who research and trade a number of stocks, and a positive loading captures investors who trade in a concentrated set of stocks, and research analyst, technical, and industry information in particular.

While less statistically important than the other principal components, PC3 is the most intriguing in terms of the style of research interests it captures. It loads positively on research on dividends, earnings, other fundamentals, as well as "HHI Research", indicating a concentrated view on a small number of stocks. It loads negatively on snapshot, pricechart, analysts, as well as short lookback windows. We conceptualize PC3 as capturing a fundamental researcher with a focus on a few stocks and an interest in cash-flows on the one extreme, and an investor with an emphatic short-term focus on the other extreme. One might call this a *fast-versus-slow-information* factor.

These investor research styles roughly correspond to the types of stocks they tend to research. We illustrate this finding by generating an investor-level tilt in normalized stock characteristics, such as return volatility, skewness, beta, age, size, dividend yield, nominal price.

The first principal component of the investor-level characteristic tilts loads strongly positively on return volatility and loads negatively on firm size.

The first three columns of Table 10 report investor-level regressions of PC1, PC2 and PC3 on the same households' characteristic tilt. The second set of three columns repeats the same analysis, but uses the household's return volatility tilt directly. In both specifications there is a positive correlation with PC1 and a negative correlation with PC3. The positive correlation with PC1 (level) indicates that investors who have a high level of research interest in general tend to tilt towards high volatility stocks. The negative correlation with PC3 (fast-versus-slow-information) loading indicates that investors who tend to focus on fundamental research and shy away short-term information tend to tilt away from the high volatility stocks.

VIII. Conclusion and Implications

We document stylized facts about the research interests and information set of individual investors. These shed light on the information and beliefs that individual investor trades could plausibly inject into market prices and on what information shapes investor beliefs and portfolios. The main findings include:

- The median trade is preceded by a few minutes of stock-specific research, concentrated just before the trade.
- As a group, individual investors spend more time on price charts, followed by analyst-related information, than other categories of information. Most trades are preceded by the presentation of a snapshot of a dozen or so statistics, but these may be forced by the broker, and investors often don't display an interest in going beyond this summary.

- Individual investors have heterogeneous research styles. Some spend more time on stock research than others; some review more stocks while others concentrate; and some focus on indicators of near-term prospects while others on long-term fundamentals. This behavior is correlated with stock characteristics in intuitive ways.

Is such behavior optimal? This paper does not answer this question. But, consider the contemporary trading environment. A more quantitative fund manager might spend years and millions of dollars refining algorithms—and then execute trades, given fresh data, within hours, minutes, or seconds. A traditional active fund manager might employ a number of industry specialists and spend days to weeks contemplating a large new position. It is at best optimistic to think that the typical individual investor could outperform such competitors with their own few minutes of research. The optimal amount of stock-specific research for the typical individual investor might be essentially zero, as advocates of indexing would maintain. We are exploring the extent to which an individual investor's research—which is often perfunctory but occasionally in-depth—pays off in terms of better performance.

References

Bakos, Yannis, Florencia Marotta-Wurgler, and David R. Trossen, 2014, Does anyone read the fine print? Consumer attention to standard-form contracts, *Journal of Legal Studies* 43, 1-35.

Barber, Brad M., and Terrance Odean, 2001, The Internet and the investor, *Journal of Economic Perspectives* 15, 41-54.

Barber, Brad M., and Terrance Odean, 2007, All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors, *Review of Financial Studies* 21, 785-818.

Boehmer, Ekkehart, Charles M. Jones, Xiaoyan Zhang, and Xinran Zhang, 2021, Tracking retail investor activity, *Journal of Finance* 76, 2249-2305.

Bogan, Vicki, 2008. Stock market participation and the internet, *Journal of Financial and Quantitative Analysis*, 43(1), 191-211.

Choi, James J., David Laibson, and Andrew Metrick, 2002, *Journal of Financial Economics* 64, 397-421.

Da, Zhi, Joseph Engelberg, and Pengjie Gao, 2011, In search of attention, *Journal of Finance* 66, 1461-1499.

Da, Zhi, Joseph Engelberg, and Pengjie Gao, 2014, The sum of all FEARS: Investor sentiment and asset prices, *Review of Financial Studies* 28, 1-32.

Drake, Michael S., Darren T. Roulstone, and Jacob R. Thornock, 2012, Investor information demand: Evidence from Google searches around earnings announcements, *Journal of Accounting Research* 50, 1001-1040.

Jegadeesh, Narasimhan, 1990, Evidence of predictable behavior of security returns, *Journal of Finance* 45, 881-898.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65-91.

Laarits, Toomas, Florencia Marotta-Wurgler, and Jeffrey Wurgler, 2024, How much do individual investors care about SEC-mandated disclosures? Evidence from IPO and mutual fund prospectuses, New York University work in progress.

Lawrence, Alastair, James Ryans, Estelle Sun, and Nikolay Laptev, 2018, Earnings announcement promotions: A Yahoo Finance experiment, *Journal of Accounting and Economics* 66, 399-414.

Lease, Ronald C., Wilbur G. Lewellen, and Gary G. Schlarbaum, 1974, The individual investor: Attributes and Attitudes, *Journal of Finance* 29, 413-433.

Lease, Ronald C., Wilbur G. Lewellen, and Gary G. Schlarbaum, 1976, Market segmentation: Evidence on the individual investor, *Financial Analysts Journal* 32, 53-60.

Marotta-Wurgler, Florencia, 2012, Will increased disclosure help? Evaluating the recommendations of the ALI's 'Principles of the Law of Software Contracts,' *University of Chicago Law Review* 78, 165-186.

Marotta-Wurgler, Florencia, 2012, Does contract disclosure matter? *Journal of Institutional and Theoretical Economics* 168, p. 94-123.

Merton, Robert C., 1987, A simple model of capital market equilibrium with incomplete information, *Journal of Finance* 42, 483-510.

Miller, Edward M., 1977, Risk, uncertainty, and divergence of opinion, *Journal of Finance* 32, 1151-1168.

Reinganum, Marc R., 1981, Misspecification of capital asset pricing: Empirical anomalies based on earnings yields and market values, *Journal of Financial Economics* 9, 19-46.

Sicherman, Nachum, George Loewenstein, Duane J. Seppi, and Stephen P. Utkus, 2015, Financial attention, *Review of Financial Studies* 29, 863-897.

Thayer, Jane, 2011, Determinants of investors' information acquisition: Credibility and confirmation, *The Accounting Review* 86, 1-22.

Table 1. Stylized example of browsing data. A constructed example of a household's fourteen-minute browsing session in raw data format. The session indicates visits to non-finance websites, finance or financial news websites, and broker websites. The finance and broker websites indicate the nature of the content that is displayed, tickers involved, and some details of trades.

Row	Timestamp	Link
1	3/23/2007 2:31:11 PM	http://formula1.com/
2	3/23/2007 2:36:20 PM	http://cnbc.com/
3	3/23/2007 2:36:28 PM	http://quote.cnbc.com/quote.htm?symbols=.DJIA|.NCOMP|.SPX|GB;FTSE|DE;DAXX|FR;CAC|JP;N225|CN;SHI|US.10|$$EURUSD&requestMethod=quick&realtime=1
4	3/23/2007 2:37:30 PM	https://[broker-1].com/user/login
5	3/23/2007 2:37:37 PM	https://[broker-1].com/eresearch/markets_sectors/analysis/mostactives.jhtml
6	3/23/2007 2:38:01 PM	https://[broker-1].com/stocks/earnings/earnings.asp?symbol=IMCL
7	3/23/2007 2:38:22 PM	https://[broker-1].com/stocks/earnings/update_earningsChart.asp?showQTRChange=1&display=mountain&overlay=&indicator=volume,1,13,EMA,13&duration=1095&daysforward=545&year=2006
8	3/23/2007 2:38:51 PM	https://[broker-1].com/charts/update_chart.asp?duration=365&frequency=1day&scaling=linear&display=candlestickcolor
9	3/23/2007 2:39:14 PM	http://finance.yahoo.com/
10	3/23/2007 2:39:20 PM	http://finance.yahoo.com/q?s=IMCL
11	3/23/2007 2:39:25 PM	http://biz.yahoo.com/bw/070530/000000000.html?.v=1
12	3/23/2007 2:39:29 PM	http://finance.yahoo.com/q?s=IMCL
13	3/23/2007 2:39:32 PM	http://finance.yahoo.com/q/ao?s=IMCL
14	3/23/2007 2:40:52 PM	http://formula1.com/
15	3/23/2007 2:43:03 PM	https://[broker-1].com/tradingideas/screener/stock_screener.asp
16	3/23/2007 2:43:10 PM	https://[broker-1].com/tradingideas/screener/stock_screener_results.asp?c=IBES.EPSGrowthNFY:GEQ:50:LSS:100001
17	3/23/2007 2:43:33 PM	https://[broker-1].com/common/symbol_info/symbolInfo.asp?symbol=GOOG
18	3/23/2007 2:43:39 PM	https://[broker-1].com/stocks/snapshot/snapshot.asp?symbol=GOOG
19	3/23/2007 2:44:11 PM	https://[broker-1].com/stocks/charts/charts.asp?symbol=GOOG
20	3/23/2007 2:45:02 PM	https://[broker-1].com/invest/socreateentry?ordertype="sell"&symbol=GOOG&shares=100

21	3/23/2007 2:45:19 PM	https://[broker-1].com/user/logout
22	3/23/2007 2:45:31 PM	http://us.f518.mail.yahoo.com/ym/login?login

Table 2. Summary statistics. Household-level statistics.

		Percentiles					
	Mean	5th	25th	Median	75th	95th	N
Panel A.							
Income	71.66	12.50	75.00	75.00	100.00	100.00	445
Age	50.08	29.00	39.00	50.00	59.00	74.00	467
Panel B.							
N Browsing Sessions	102.79	25.00	65.75	92.17	137.00	205.67	484
Minutes/Session	25.84	10.09	16.28	22.84	32.03	51.77	484
N Unique Broker/Finance Dom.	3.12	0.50	1.50	2.75	4.50	7.00	484
Hours/Month Broker Site	3.57	0.14	0.43	1.36	3.53	14.66	484
Hours/Month Other Finance Site	1.59	0.00	0.02	0.24	1.26	6.42	484
Hours/Month Other Site	37.33	6.47	18.37	30.93	49.98	86.72	484
Panel C.							
N Sessions with a Trade	1.33	0.25	0.33	0.50	1.00	6.00	484
N Stock Trades	2.20	0.25	0.33	0.67	1.71	9.50	484
N Stock Buys	1.10	0.00	0.00	0.25	0.67	5.00	484
N Stock Sells	1.11	0.00	0.00	0.33	1.00	4.50	484
N Unique Tickers Traded	1.47	0.25	0.33	0.50	1.25	5.50	484
USD Traded (in thousands)	13.54	0.05	0.33	2.02	8.40	52.03	191
Minutes/Session Trading	2.72	0.10	0.56	1.40	3.26	10.00	484
Minutes/Month Trading	6.46	0.11	0.68	1.97	5.65	25.69	484

Table 3. Imputation of mixed-content links to specific content categories. Certain types of pages in the raw data lead to content that involves multiple categories. For example, several finance and brokerage domains contain snapshot pages, and between 0% and 11% of the content on those pages involve traditional risk statistics. The imputed content distribution for time on message boards is based on the distribution of content on Reddit r/stocks message board discussions. The imputed content distribution for financial news of unknown type is based on the distribution of content across RavenPack's detailed finance article content classifications. The last two rows do not sum to 100% because of remaining NEC or Indeterminate content. See Data Appendix for further details.

	Proportional Imputation to Content Category									
	Risk Stats	Earnings	Dividends	Other fundls.	Analysts	Industry	Events	Owner-ship	Price Charts	Technical
Snapshot (min-max)	0-0.11	0-0.13	0-0.13	0-0.14	0-0.33	0-0.04	0-0.15	0-0.13	0.38-0.79	0-0.20
Message Board - Indeterminate	0	0.27	0.02	0.07	0.02	0	0.14	0.01	0.26	0.02
News - Indeterminate	0.06	0.31	0.03	0.03	0.07	0.11	0.01	0.02	0.12	0.10

Table 4. Aggregate research activity. N=484 households. All measures are constructed on the household-month level.

		Percentiles				
Browsing Activity (Per HH-Month)	Mean	5th	25th	Median	75th	95th
All Research						
N Sessions Including Research	28.94	0.67	4.25	14.29	37.83	104.00
N Unique Tickers Researched	29.97	1.00	4.00	12.00	31.67	119.00
N Unique Broker/Finance Dom.	3.12	0.50	1.50	2.75	4.50	7.00
Minutes/Session Research	3.49	0.27	1.02	2.23	4.38	11.18
Minutes/Month Research	114.13	1.00	7.25	36.20	103.33	520.34
Minutes/Trade Research	138.37	1.61	9.17	36.06	107.41	638.23
All Other Financial						
N Sessions	23.05	1.00	5.25	14.13	35.00	72.00
Minutes/Session	5.38	0.89	2.14	4.12	6.60	14.84
Minutes/Month	151.32	4.41	17.96	57.00	150.13	620.91

Table 5. Research time across domains. Household-level total minutes normalized by the number of trades by the household. Broker websites are those for which we observe trading activity and are anonymized. "Finance" sites contain financial news or other statistics; these are anonymized except for Yahoo Finance and the SEC website. Company websites involve visits to the domain of a stock researched on other sites at least once in the sample.

	All Research				Research Matched to Non-Traded Ticker				Research Matched to Traded Ticker			
Domain	Mean All HH	N t>0 HH	Mean t>0 HH	Median t>0 HH	Mean All HH	N t>0 HH	Mean t>0 HH	Median t>0 HH	Mean All HH	N t>0 HH	Mean t>0 HH	Median t>0 HH
finance-yahoo	44.46	259	83.09	6.88	19.48	211	44.69	4.01	4.51	133	16.41	1.13
broker-1	12.34	132	45.24	6.15	5.28	93	27.5	4.03	1.38	104	6.44	1.48
broker-2	10.89	87	60.6	10.42	3.18	75	20.54	3.21	0.45	67	3.26	1.15
broker-3	10.73	126	41.22	13.46	4.33	119	17.62	5.98	1.34	101	6.4	2.19
broker-4	9.47	81	56.59	5.52	1.67	47	17.18	5.57	0.2	34	2.8	1.37
broker-5	8.7	171	24.63	5.77	2.93	129	10.99	2.42	0.79	90	4.25	0.72
finance-B	7.76	173	21.7	1.4	3.24	100	15.7	0.54	0.22	39	2.68	0.27
finance-D	7.04	166	20.52	1.42	4.49	106	20.5	1.85	0.67	66	4.91	0.82
finance-G	3.38	37	44.24	0.63	3	35	41.48	0.36	0.27	10	13.05	1.21
broker-6	3.33	17	94.89	1.47	2.6	15	83.8	0.97	0.13	9	6.73	0.32
finance-E	3.26	6	262.88	201.16	2.71	6	218.49	161.38	0.07	5	7.21	5.46
finance-F	3.11	2	751.48	751.48	3.06	2	740.8	740.8	0.03	2	7.71	7.71
finance-H	2.36	134	8.54	1.33	0.46	42	5.25	0.56	0.02	18	0.5	0.44
finance-I	2.15	57	18.22	1.98	0.92	49	9.13	1.88	0.86	30	13.8	1.19
finance-J	1.22	36	16.39	1.16	1.03	32	15.56	0.93	0.1	11	4.41	0.82
finance-K	1.2	20	29.02	8.38	1.1	17	31.32	8.68	0.09	14	3.15	1.88
Company Website	1.18	290	1.96	0.13	0.76	266	1.38	0.04	0.42	97	2.08	0.13
finance-L	1.09	31	16.97	3.32	0.26	20	6.23	1.31	0.02	10	0.84	0.91
finance-M	1.04	47	10.68	0.83	0.5	37	6.49	0.33	0.04	14	1.51	0.2
finance-N	1	16	30.24	0.98	0.12	11	5.41	0.84	0.01	1	2.47	2.47
finance-O	0.63	51	6.01	0.88	0.52	51	4.97	0.78	0.01	17	0.36	0.17
finance-P	0.56	47	5.75	1.1	0.11	19	2.86	1.15	0.17	10	8.41	0.41
finance-Q	0.34	12	13.6	2.78	0.01	7	0.77	0.26	0	5	0.14	0.06
finance-T	0.31	38	3.92	0.64	0.18	24	3.65	0.92	0.01	7	0.95	0.17

finance-C	0.3	69	2.07	0.58	0.02	15	0.73	0.27	0	4	0.33	0.21
finance-R	0.27	45	2.94	0.16	0.25	33	3.62	0.18	0.02	20	0.42	0.05
finance-sec	0.21	32	3.19	0.87	0.16	19	4.08	0.87	0	1	0.84	0.84
broker-7	0.08	2	19.57	19.57	0	1	2.12	2.12				
broker-8	0	4	0.34	0.4	0	4	0.21	0.13	0	2	0.27	0.27
Total	138.4				62.4				11.8			

Table 6. Most popular stocks to research.

			Research Excluding Website			Research Including Website		
Name	Ticker	Market Cap (in billions)	N t>0 HH	Mean t>0 HH	Median t>0 HH	N t>0 HH	Mean t>0 HH	Median t>0 HH
APPLE INC	AAPL	74	128	2.4	0.63	128	2.74	0.69
MICROSOFT CORP	MSFT	302	108	4.5	0.76	108	5.01	0.9
CISCO SYSTEMS INC	CSCO	161	101	3.47	0.95	101	3.48	0.95
GENERAL ELECTRIC CO	GE	372	97	2.34	0.37	97	2.44	0.37
CITIGROUP INC	C	271	92	1.96	0.51	92	2.11	0.51
INTEL CORP	INTC	121	87	2.27	0.57	87	2.29	0.57
FORD MOTOR CO DEL	F	15	87	2.71	0.67	87	2.76	0.67
SIRIUS SATELLITE RADIO INC	SIRI	5	86	3.14	0.68	86	3.39	0.73
A T & T INC	T	236	83	3.69	0.77	83	4.01	0.77
GOOGLE INC	GOOG	114	78	3.59	0.52	78	3.71	0.53
ALTRIA GROUP INC	MO	183	73	4.48	1.25	73	5.03	1.35
DELL INC	DELL	54	70	1.62	0.34	70	1.98	0.37
EXXON MOBIL CORP	XOM	432	69	2.63	0.47	69	2.66	0.47
WAL MART STORES INC	WMT	199	65	2.03	0.34	65	2.33	0.4
GENERAL MOTORS CORP	GM	19	64	2.63	0.28	64	2.64	0.28
PFIZER INC	PFE	189	63	2.33	0.48	63	2.33	0.48
YAHOO INC	YHOO	39	62	4.46	0.38	62	5.31	0.38
HOME DEPOT INC	HD	83	61	3.67	0.43	61	3.8	0.58
GOLDMAN SACHS GROUP INC	GS	90	57	3.41	0.69	57	3.45	0.69
CONOCOPHILLIPS	COP	109	56	1.27	0.33	56	1.31	0.33

Table 7. Minutes of research per trade (original content categories). Total number of minutes of research divided by total number of trades.

	All Research					Research Matched to Non-Traded Ticker					Research Matched to Traded Ticker				
Research Category	Mean All HH	Median All HH	N t>0 HH	Mean t>0 HH	Median t>0 HH	Mean All HH	Median All HH	N t>0 HH	Mean t>0 HH	Median t>0 HH	Mean All HH	Median All HH	N t>0 HH	Mean t>0 HH	Median t>0 HH
Risk Statistics	0.0	0.0	7	0.3	0.1	0.0	0.0	2	0.1	0.1	0.0	0.0	3	0.2	0.3
Earnings	1.3	0.0	178	3.5	0.4	0.8	0.0	127	3.0	0.3	0.1	0.0	67	0.6	0.2
Dividends	0.2	0.0	15	6.9	0.1	0.2	0.0	11	9.2	0.3	0.0	0.0	7	0.3	0.1
Other Fundamentals	1.3	0.0	229	2.7	0.5	0.6	0.0	175	1.6	0.3	0.1	0.0	92	0.6	0.1
Analysts	12.3	0.9	341	17.5	2.6	6.1	0.0	227	13.0	1.0	0.7	0.0	128	2.5	0.4
Events	0.0	0.0	24	0.8	0.3	0.0	0.0	7	0.2	0.1					
Ownership	0.4	0.0	120	1.8	0.4	0.3	0.0	82	1.8	0.2	0.0	0.0	46	0.4	0.1
Pricechart	39.1	5.8	450	42.1	6.7	16.1	1.8	404	19.3	3.7	3.6	0.4	336	5.2	1.2
Technical	2.3	0.0	215	5.2	0.9	0.9	0.0	145	3.0	0.5	0.1	0.0	64	1.1	0.3
Company Website	1.2	0.0	290	2.0	0.1	0.8	0.0	266	1.4	0.0	0.4	0.0	97	2.1	0.1
Snapshot	37.0	6.5	452	39.6	8.3	23.1	2.1	409	27.4	3.6	4.4	0.3	325	6.6	0.9
News-Indeterminate	14.2	2.2	410	16.7	3.6	3.8	0.2	316	5.8	1.0	0.6	0.0	193	1.6	0.3
Message Board	6.3	0.0	85	36.1	1.4	0.1	0.0	30	2.2	0.9	1.0	0.0	15	32.8	4.5
NEC	12.2	0.9	391	15.1	2.0	7.1	0.2	319	10.8	1.0	0.1	0.0	124	0.5	0.1
Indeterminate	10.4	1.0	370	13.7	2.2	2.5	0.1	292	4.2	0.7	0.5	0.0	210	1.2	0.2
Mean Total across HH	138.4					62.4					11.8				
Median Total across HH	35.7					11.4					2.0				

Table 8. Minutes of research per trade (imputed content categories). Total number of minutes of research divided by total number of trades.

	All Research					Research Matched to Non-Traded Ticker					Research Matched to Traded Ticker				
Research Category	Mean All HH	Median All HH	N t>0 HH	Mean t>0 HH	Median t>0 HH	Mean All HH	Median All HH	N t>0 HH	Mean t>0 HH	Median t>0 HH	Mean All HH	Median All HH	N t>0 HH	Mean t>0 HH	Median t>0 HH
Risk Statistics	1.8	0.2	359	2.4	0.4	0.6	0.0	298	1.0	0.1	0.2	0.0	210	0.4	0.1
Earnings	9.4	1.6	456	9.9	1.9	3.1	0.4	408	3.7	0.6	0.8	0.0	321	1.2	0.1
Dividends	4.2	0.7	456	4.4	0.8	2.4	0.2	407	2.8	0.4	0.4	0.0	320	0.6	0.1
Other Fundamentals	5.2	0.9	454	5.5	1.1	2.8	0.3	411	3.3	0.5	0.5	0.0	326	0.8	0.1
Statements	0.1	0.0	126	0.4	0.0	0.1	0.0	85	0.4	0.0	0.0	0.0	39	0.1	0.0
Analysts	18.3	2.0	437	20.3	2.8	10.2	0.4	378	13.1	1.1	1.6	0.0	294	2.6	0.2
Industry	0.9	0.0	183	2.3	0.3	0.1	0.0	136	0.5	0.1	0.1	0.0	94	0.7	0.1
Events	4.1	0.6	430	4.6	0.9	2.0	0.2	376	2.6	0.3	0.4	0.0	279	0.7	0.1
Ownership	1.8	0.2	427	2.1	0.3	0.9	0.0	367	1.2	0.1	0.1	0.0	261	0.3	0.1
Pricechart	63.8	14.6	482	64.1	14.6	30.6	5.3	445	33.2	6.6	5.9	1.1	408	6.9	1.5
Technical	5.8	0.8	455	6.1	0.9	2.7	0.2	411	3.2	0.3	0.5	0.0	325	0.7	0.1
Company Website	1.2	0.0	290	2.0	0.1	0.8	0.0	266	1.4	0.0	0.4	0.0	97	2.1	0.1
Markets	0.9	0.0	85	4.9	0.2	0.0	0.0	30	0.3	0.1	0.1	0.0	15	4.4	0.6
NEC	10.6	1.6	434	11.8	2.1	3.5	0.3	371	4.5	0.6	0.3	0.0	218	0.6	0.1
Indeterminate	10.5	1.0	373	13.7	2.2	2.6	0.1	301	4.2	0.7	0.5	0.0	220	1.2	0.2
Mean Total across HH	138.4					62.4					11.8				
Median Total across HH	35.7					11.4					2.0				

Table 9. Principal components of household-level research focus.

Research Category or *HH Style Characteristic*	First PC	Second PC	Third PC
High Trade	0.34	-0.42	0.06
Snapshot	0.31	0.13	-0.33
News-Indeterminate	0.30	0.18	-0.05
High Permno	0.30	-0.49	0.03
Pricechart	0.30	0.18	-0.32
Earnings	0.26	0.21	0.40
Other Fundamentals	0.25	0.17	0.38
Technical	0.24	0.26	0.30
Analysts	0.23	0.27	-0.24
Short Lookback	0.22	0.22	-0.01
Ownership	0.15	0.21	0.00
Dividends	0.12	-0.05	0.49
Long Lookback	0.06	-0.17	0.01
HHI Research	-0.28	0.03	0.31
HHI Trade	-0.33	0.40	0.02
Eigenvalue	3.72	1.82	1.24
Proportion Explained	24.8	12.15	8.27

Table 10. Household-level research principal components and characteristic tilts.

	PC 1	PC 2	PC 3	PC 1	PC 2	PC 3
Characteristic PC1	0.0742*	0.0123	-0.0543**			
	(2.32)	(0.55)	(-3.00)			
Return SD tilt				0.394**	0.0755	-0.287***
				(2.81)	(0.69)	(-3.45)
Constant	-3.00e-09	3.68e-09	2.99e-09	0.0341	0.00653	-0.0249
	(-0.00)	(0.00)	(0.00)	(0.39)	(0.10)	(-0.51)
Observations	484	484	484	484	484	484
R^2	0.008	0.000	0.013	0.010	0.001	0.016

Figure 1. Market cap versus number of households conducting research. Sample limited to the top 100 stocks by market cap, a number of which do not fall within the top 100 most-broadly researched.

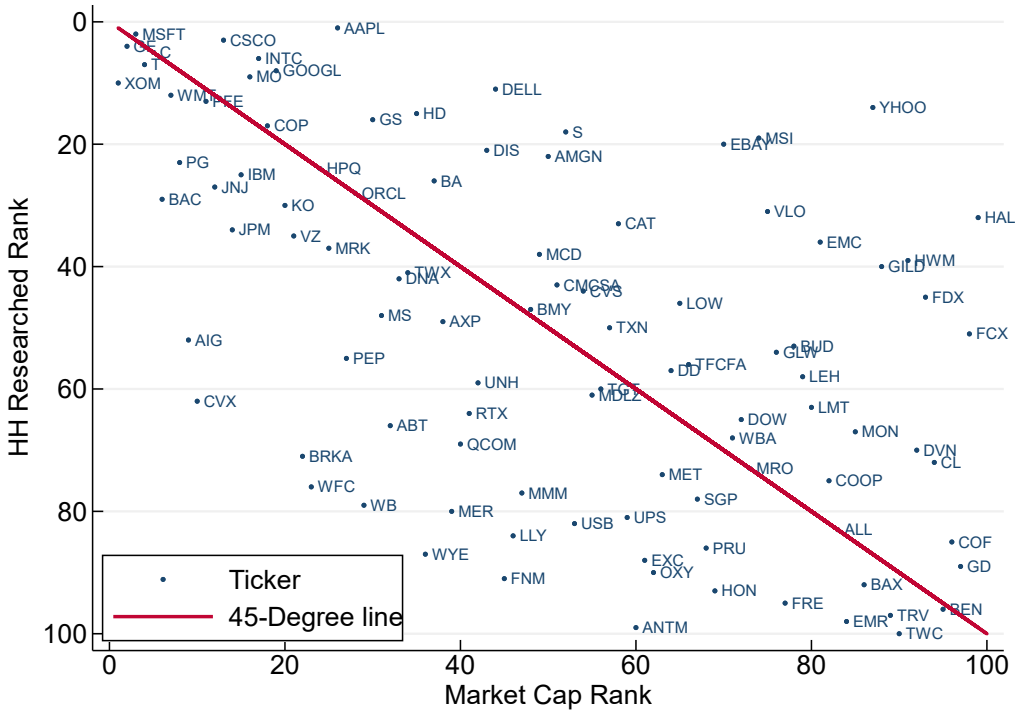


Figure 2. Research around events: Clickstrem data vs. Google Trends. Panel A shows the number of investors in the clickstream data with research activity on Microsoft stock in the month of its January 25, 2007 quarterly earnings announcement and compares this to the Google Trends index for interest in "MSFT". Panel B shows these two series around Apple's January 9, 2007 introduction of the iPhone and Apple's January 17, 2007 quarterly earnings announcement, and Panel C shows these series around the February 19, 2007 announcement of a merger between Sirius and XM. Weekends are shaded.

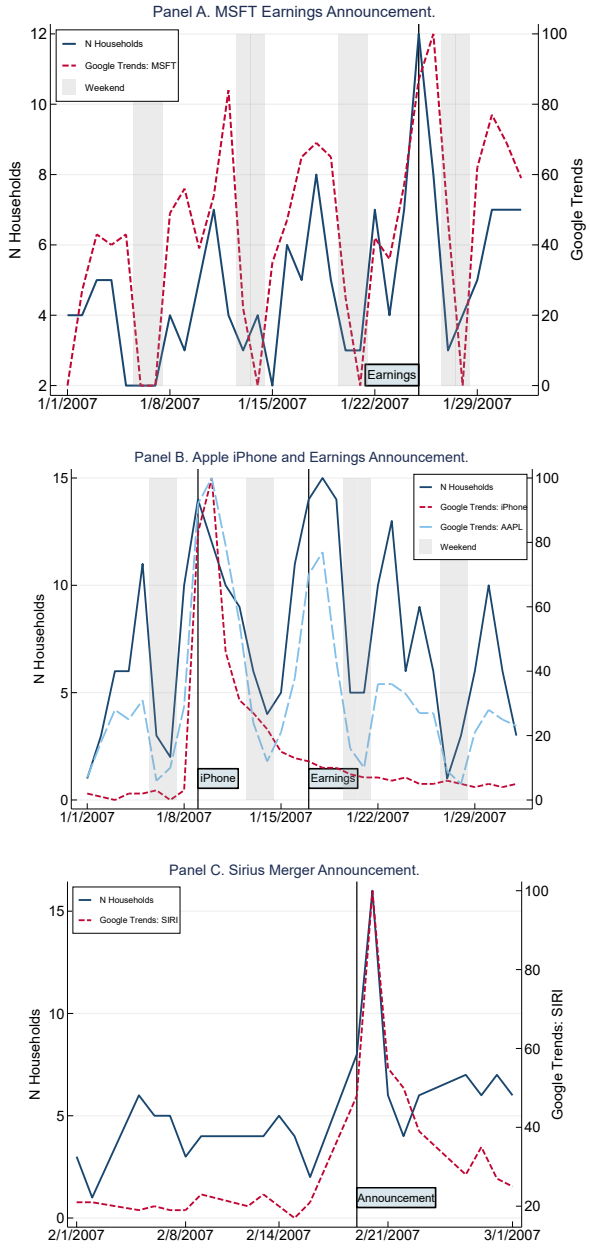


Figure 3. Length of chart lookback windows.

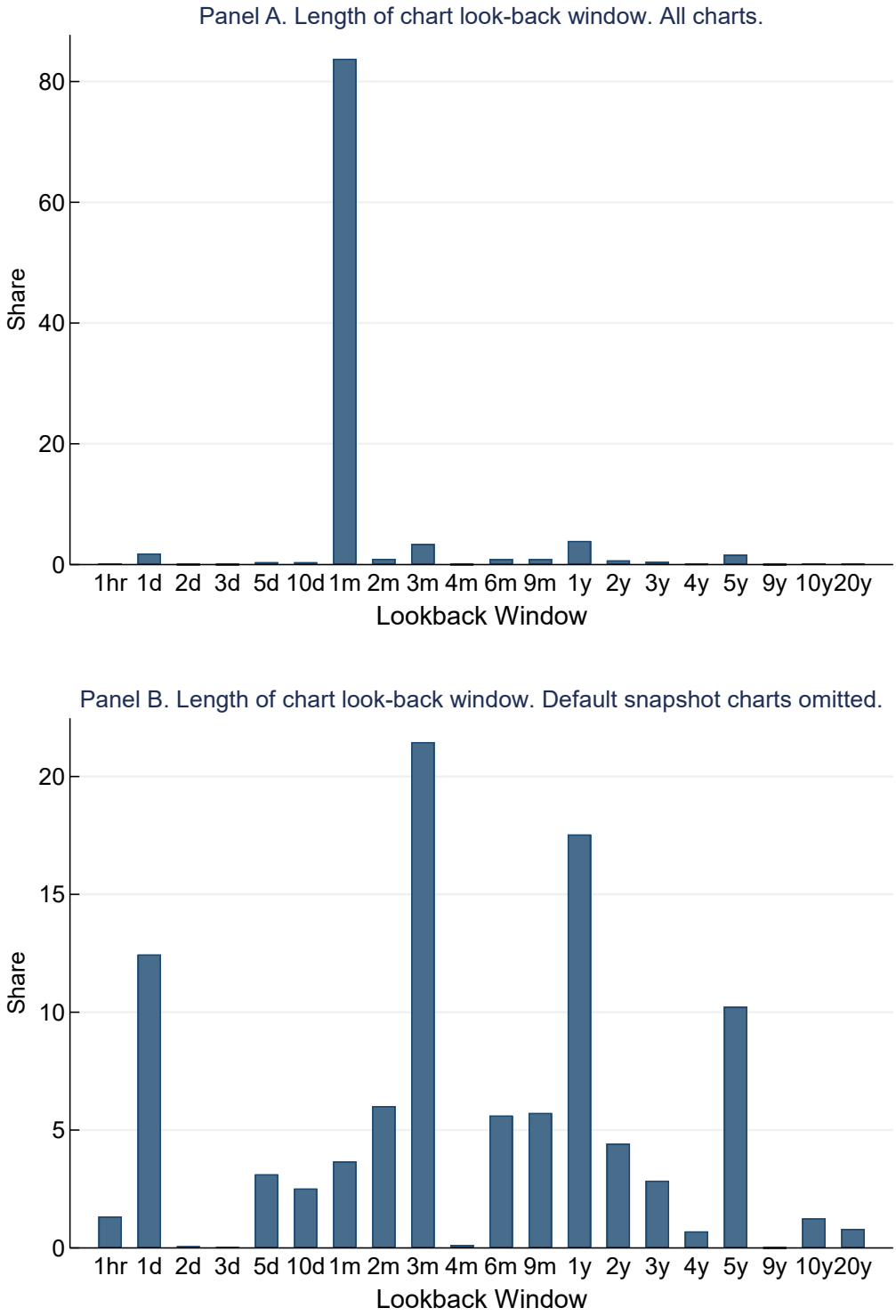


Figure 4. Total research time by category. Panel B restricts to research matched to ticker only.

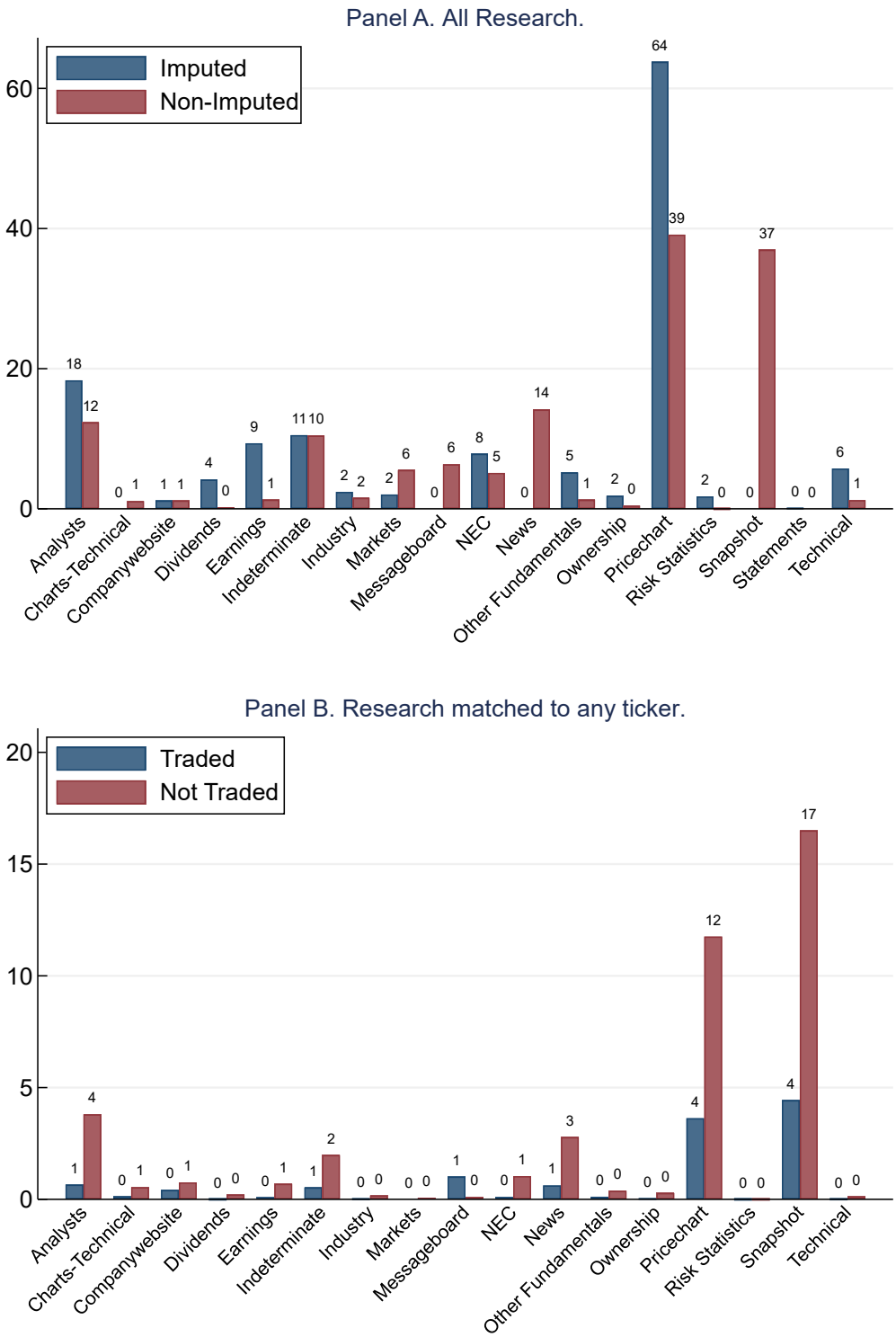
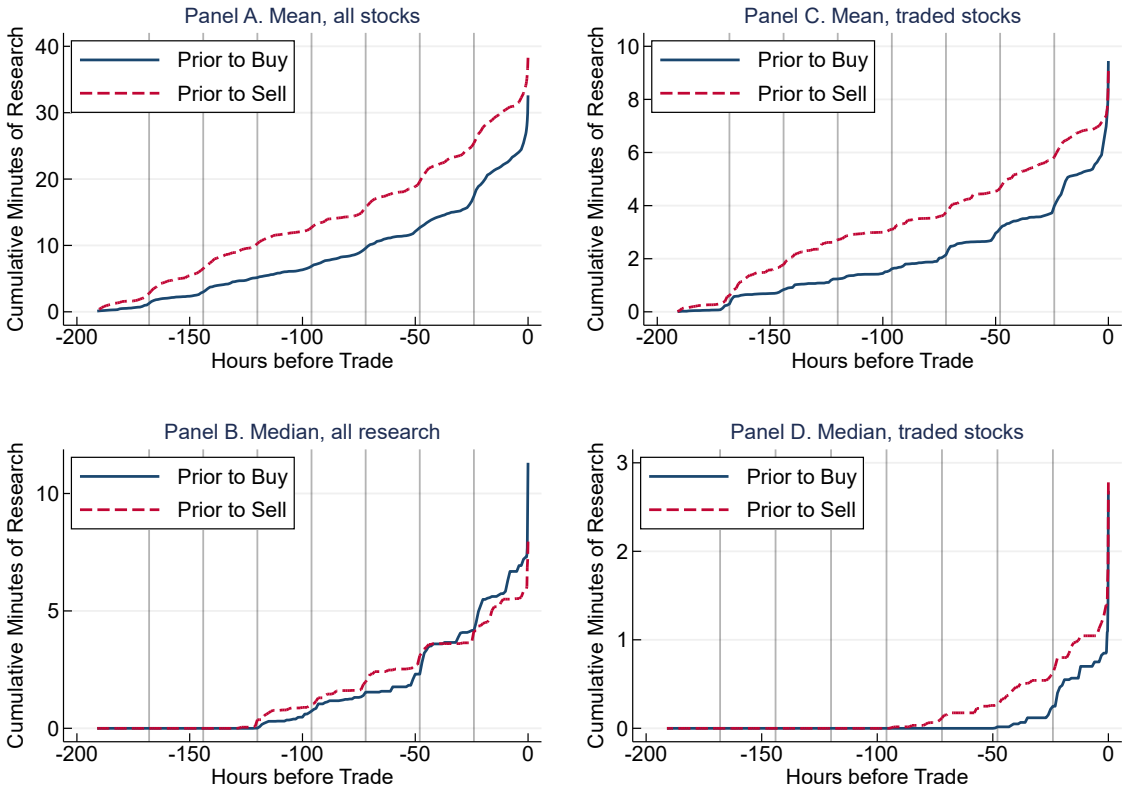


Figure 5. Research in the runup to trades.



a

Data Appendix

This Data Appendix describes a number of data processing steps. The starting point is clickstream data for a nationally representative sample of households for January, February, March, and June of 2007. Households move in and out of the sample and are present between one and four months. For each household we observe the timestamped sequence of all URLs ("links") that they visited, active time spent at each page, and a rough categorization of the type of content served at the specific page. The raw click data comprises roughly two billion URLs.

I. Constructing trading household clickstream data

A. Identifying trading households

We begin with a list of 11 online brokerages operating in 2007 and the associated web domain names. Using the domain names, we read through the entire dataset and create a subset of clicks on broker websites. Within this set of clicks we flag clicks that contain terms potentially indicative of stock trades: we look for terms such as "trade", "order", "buy", "sell", "shares" and so on. For instance, the URL in row 20 of Table 1:

```
https://[broker1].com/invest/socreateentry?ordertype=sell&s
ymbol=GOOG&shares=100
```

would be flagged for the inclusion of at least two such terms. Whenever we find a link that might represent a trade, we use regular expression ("regex") matching to find all links on that broker's site that follow the same syntax. We additionally use regular expression matching to read out the stock ticker, the trade action (buy or sell), and, if available, the quantity traded and the trade price. For the above example we would use the regex pattern:

```
socreateentry\?ordertype=(\w+)\&symbol=(\w+)\&shares=(\d+)
```

to read out the action, stock ticker, and quantity traded, in that order. From the above entry we would read out "Sell", the ticker GOOG, and the quantity of 100 shares. We match the traded tickers to the CRSP header file and keep trades that that correspond to common stock and ADRs.

Having identified potential trades, we take a close look at the click sequences to confirm that they represent a stock trade. In many cases, a number of clicks occur in rapid succession pertaining to the same underling ticker. When we see a sequence of potential trade clicks relating to the same ticker and with the same trade action (buy or sell) with a gap of no longer than 30 minutes we collapse those clicks into one trade, represented by the last click in the sequence.

With this procedure we end up with 2,911 total common stock and ADR trades by 484 distinct trading households.

B. Processing traders' browsing data

Once stock trading households are identified, we set about documenting the information they acquire before and after trading. We read through the raw data once more and keep all the clicks associated with these 484 households. We end up with over 8.5 million individual clicks corresponding to some 60,300 combined hours of browsing.

At this point we carry out two pre-processing steps. First, we check the URLs for special characters and characters that are reported in Unicode values and transform them into corresponding characters. This step makes it easier to process the URLs via regular expression matching. Second, we organize the URLs in groups corresponding to the host site. For instance, all clicks with the stems biz.yahoo.com, finance.yahoo.com, and quote.yahoo.com are assigned to the same group—Yahoo Finance—which is subsequently processed in one shared routine.

The websites we process in this manner are brokerage sites and non-broker finance sites such as Yahoo Finance, CNBC, Marketwatch, and the SEC.

II. Reading out tickers and research categories

A. Processing ticker-matched research

Our first goal is to identify clicks associated with stock tickers. We proceed in an iterative manner: we look at the URLs until we find one that represents research associated with a ticker. For instance, at Yahoo Finance we might come across a URL like in row 10 of Table 1:

`http:/finance.yahoo.com/q/ao?s=IMCL`

and, with the help of the Wayback Machine at

`https://web.archive.org/web/20070407055358/http:/finance.yahoo.com/q/ao?s=IMCL`

we learn that on the abbreviation "ao" refers to "analyst opinions". We then use a simple regex pattern to classify all such URLs as pertaining to analyst information and associate them with the corresponding ticker.

In general, in the case of non-broker finance websites such as Yahoo Finance we are often able to find a version of the page on the Wayback Machine, allowing us to directly verify the type of content. Brokerage sites, however, typically require logging in to view the research material, and hence are not captured by the Wayback Machine. However, we were in many cases able to find promotional material or video guides that illustrated the content on various pages. In some cases, however, we have to make the determination based on the URL alone. For instance, a URL might contain the term "analystresearch" and a ticker, in which case it's clear that the investor was looking up analyst reports. In instances where we can deduce nothing about the type of content, we classify the content as "indeterminate".

In the case of broker websites, we typically first allow the structure of the URLs to dictate the original classification of a given link. For instance, we might come across a URL like in row 17 of Table 1:

```
https://[broker1].com/stocks/snapshot/snapshot.asp?symbol=G
OOG
```

and a regex pattern that reads out the terms separated by forward slashes identifies a slew of different research sites: stock news, fundamentals, ownership and so on. We continue processing URLs in this way until we can no longer find URLs pertaining to research that contain ticker information.

B. Special considerations

A number of clicks require further attention to fully extract the available information. The first such special case involves URLs that reference multiple tickers, such as row 3 in Table 1. In such cases we create duplicates of the underlying link and associate one copy of the link to each of the associated tickers. We then split the time spent on the original link evenly between each of the newly created copies.

Another special case involves URLs corresponding to charts. These often refer to multiple tickers as well as multiple types of information. For instance, a chart might plot the prices and volume of two stocks. In that case, we first proceed as above and split the original URL into copies corresponding to each of the associated tickers. These resulting URLs, in turn, are again split up to correspond to the different types of information. As before, we evenly split the time spent at the original site between all the resulting copies of the URL. Chart URLs often contain further information about the information the investor is seeking out, such as the lookback window and frequency of data plotted. We again use regular expression matching to extract such information. For instance, consider the URL in row 7 of Table 1:

```
https://[broker1].com/stocks/earnings/update_earningsChart.
asp?showQTRChange=1&display=mountain&overlay=&indicator=vol
ume,1,13,EMA,13&duration=1095&daysforward=545&year=2006
```

This link reveals that the investor was charting earnings over a period of 3 years (1095 days) with a period of 1.5 years (545 days) predicted into the future.

C. Cleaning tickers

Having processed the URLs and extracted ticker information we separately process the ticker information that we read out using regular expression matching. We first read out any terms appended to the ticker, such as country indicators for stocks trading abroad, ".PK" for pink sheets, and so on. We then match the resulting tickers to the CRSP header file. We inspect the unmatched tickers and match by hand, whenever possible. Reasons for unmatched tickers can be caused by stocks with multiple share classes (such as GOOG and GOOGL), or user input errors when searching for a particular entity. On occasion the ticker field instead represents the investor's input of a company name, say "WALLMART". Whenever obvious, we assign such user-inputted fields to the associated ticker, though we do not count the time from clicks involving such misspelled tickers in the research time totals.

We separately create indicators for identifiers that pertain to indexes, mutual funds, and to currencies. Tickers corresponding to indexes and currencies are often indicated by a special character on brokerage sites. For instance, in row 3 of Table 1 indexes are indicated by a leading period, while exchange rates are indicated with a double dollar sign.

D. Processing other activity at brokerage sites

Having exhausted clicks that contain ticker information, we turn to classifying other time spent at broker and finance sites. We first seek to classify research time without explicit tickers in

the URL. We follow a process very similar to the one described above and use regular expression patterns to read out any information pertaining to the content of the URL.

Additionally, we classify other activity on broker sites that may reference tickers but does not constitute research. The two main categories that might contain ticker-related information are "portfolio" and "trading": clicks involving reviewing current investment positions and clicks involving inputting and confirming trades. We also classify other time spent at brokerage sites: time spent on logging in, time on the main homepage, and time on various types of non-investing activities such as educational tools, bill payments, banking, and so on.

E. Processing other research activity

As a final type of investor research activity, we keep track of time spent on websites of companies that are in the set of stocks otherwise researched by the investor. We use the COMPUSTAT header file to link CRSP Permno identifiers to company website URLs and manually augment the list of company websites the companies in the top 100 in terms of research activity.

The idea behind keeping track of this time is to provide a sense of an investor's familiarity with the products of a given company. For instance, the time an investor browses the newly released first generation iPhones on Apple.com would count as research time, if they ever research AAPL on any of the finance sites. We impose an upper bound on such research time by constraining that at each point it cannot exceed twice the total of other research on the same stock. This limit ensures that the time spent at company websites does not skew research time totals.

F. Assigning unmatched research time

As discussed above, not all research clicks reveal the underlying stock or stocks. For instance, consider the sequence of URLs in rows 10, 11 and 12 of Table 1:

```
http://finance.yahoo.com/q?s=IMCL
http://biz.yahoo.com/bw/070530/000000000.html?.v=1
http://finance.yahoo.com/q?s=IMCL
```

Here the investor is looking at the "snapshot" page for ImClone Corporation (IMCL), then reads a Business Wire press release where the ticker is not included in the URL, and then returns to the IMCL snapshot page. While we don't know the content of the press release—and that particular page is not available via the Wayback Machine—we conjecture that the press release likely pertained to IMCL as it was preceded and succeeded by research on that ticker.

Based on the prevalence of such instances, we seek to assign unmatched research clicks to tickers if the nearest ticker-matched research before and after the click in question pertains to the same underlying ticker. We apply this procedure only if the bracketing ticker-matched clicks contain exactly one ticker.

G. Revising processed data

An important final part of our procedure is an iterative process of reviewing the processed data. We do this in three main ways.

Firstly, in order to ensure that we have captured all the URLs with ticker information we search the links for popular tickers that have relatively unique letter sequences. For instance, the most researched ticker in our data is AAPL, and for each of the sites we confirm that the search of the string "AAPL" does not yield in any URLs that contain ticker information but are not yet processed with a corresponding regular expression.

Secondly, we search the links with keywords relating to the main research categories such as “snapshot”, “analyst”, and “earnings” to ensure that all the relevant links are classified appropriately.

Thirdly, we collapse research time by site and category and compare the percentages across different brokers and non-broker sites. If a category is conspicuously absent in a particular website, we go back and look for evidence of that type of research.

III. Recombining and harmonizing content categories

A. Original research categories

The above procedure results in close to 3,000 hours of total research time, with roughly 1,350 hours matched to any sort of identifier, be it a ticker, a company name, or an index. This data is made up of 30 distinct websites, with the most important, Yahoo Finance, contributing close to 30% percent of total research hours.

As described in Step 1 above, in processing the research clicks at the various websites we typically take the approach of classifying based on the categorizations used on the website itself. Naturally, the categories of data that we obtain in this way are not perfectly aligned across the different websites, though there are numerous broad commonalities across the websites. For instance, all of the brokerage and finance sites see more visits to a standard page best classified as a “snapshot”: a ticker-specific summary page of different types of information such as price quotes, a chart of recent prices, brief data on earnings and dividends, and potentially links to news and analyst views. Another type of page present on all sites is a charting page with prices, volume, and technical indicators such as moving averages, candlesticks, and so on.

Beyond these two dominant categories, there are two types of pages that are prevalent across all broker and non-broker finance sites: analyst reports and fundamentals such as earnings and dividends. Additionally, across different sites we may see pages on risk metrics, technical research, industry analysis, ownership, market events, or other categories.

For the less frequently clicked categories that do not amount to more than a handful distinct clicks, we code the research as "NEC" (for not elsewhere classified), indicating that we know the type of information the investor sought out but that there is not enough investor interest to warrant inclusion in the tables. For pages where we can't tell the content we code it as "indeterminate". There are two types of pages where the content is often not obvious from the link and not captured by the Wayback Machine and hence classified as "indeterminate": news and messageboards.

B. Constructed categories

In addition to the categories suggested by the finance website, we define the following three research categories.

Companywebsite refers to time spent at a URL associated with a company that the investor has researched in any of brokerage or finance sites and we impose a running constraint that companywebsite research can at no point add up to more than 2x of the research on the associated ticker.

Markets refers to time spent on snapshot pages where the underlying security is a stock market index, or on other pages pertaining to market-wide developments.

Businessconditions refers to time spent on pages relating to aggregate business conditions. We construct this category by looking for keywords associated with inflation,

unemployment, GDP, the FOMC. We additionally add time spent at the BLS and BEA websites to this category.

C. Reassigning research time to primary categories

The approach described so far seeks to best classify each URL an investor visits. Some of these pages are amalgamations of many types of underlying primary information, while others are quite specific in their nature.

In order to attempt a full accounting of the types of information presented to investors, we construct an alternative classification where we break down all research time into primary categories. To do this we need to estimate the breakdown of underlying content categories on three main types of pages that contain a mixture of information, or where we typically don't observe the type of content: snapshots, news, and message boards.

To break down the information content of snapshots, we seek out contemporaneous screenshots of the page, or, in the case of non-broker sites, we look them up on the Wayback Machine. We then measure the relative amount of line items corresponding to each primary category. We ignore redundant information (e.g., when dividends per share are presented in addition to yields), information presented without meaningful context (e.g., daily volume presented without some notion of average volume), and other information that seems unlikely to affect individual investor behavior.

For news clicks, we typically do not observe the news article. However, for two brokers we observe in the URL the article title and source (Reuters, WSJ, etc.), as well as any associated tickers. We extract the title, tickers, and the date, and match each click to an array of articles in the RavenPack data using the ticker and a date range of three days. We then rank this set of potential matches by measuring the distance between the two titles and keeping the closest,

subject to a minimum distance requirement. The resulting dataset of matched articles allows us to estimate the breakdown of news article content using the detailed content classifications provided by RavenPack.

We follow a similar procedure to estimate the content breakdown of message board discussions. We scrape the present-day Reddit r/stocks message board and for each discussion thread match the title, "flair" (user-supplied keywords), and original post content to our research content categories. We then back out content category breakdowns by ascribing equal time to each post, dividing time equally across all matched categories.