"Analyzing Clickstream Data: The Information Set of Online Investors"

A research proposal for the July 17, 2021 NBER Conference on

Big Data and High-Performance Computing for Financial Economics

by

Toomas Laarits and Jeffrey Wurgler (NYU Stern)

April 5, 2021

Background and Initial Research Question

The sequence of browser URLs of an Internet user is called the clickstream. Clickstream data are collected by a variety of companies and are carefully analyzed by, for example, online retailers who use it to understand how consumers find their websites and maximize the probability with which visitors become buyers.

The proposed research agenda involves learning from the clickstreams of online investors. How do they do research prior to a trade? What sources of information do they consult? How does this research gathering process vary from investor to investor and stock to stock? The detailed nature of our own user-level clickstream data set, briefly described below, allows us to study the research and trading process of hundreds of online investors in unprecedented detail. To the best of our knowledge, the use of this scope of clickstream data is novel in finance.

Our first project will introduce clickstream data to the financial economics community and address the fundamental question of the online investor's "information set." The concept of information set permeates finance research, but much research necessarily assumes that the marginal investor somehow efficiently incorporates all relevant information or infers that she fails to do so for a particular category of information based on low-power tests involving under- or overreaction. While it is clearly impossible to know all information that an individual investor uses to inform a trade, we aim to show how clickstream data directly documents a key subset of research and the information set based on browsing behavior prior to an observed trade.

Example Data

Our clickstream data set contains the browsing history of a set of tens of thousands of households for the first three months of 2007. In raw form, the data set includes 2 TB of raw URL-level information; it is Big Data indeed. Due to a confidentiality agreement, we are unable to disclose the data provider, but the provider removed all personally identifying information prior to our acquisition. Out of further deference to confidentiality concerns, we also choose not to disclose the

online broker(s) whose investors we analyze.[1]

In raw form, rows of data look like this:

Figure 1: Sample Rows of Raw Clickstream Data.

household	time	date	link
1	8:55:20	5-Jan-07	http://streamerapi.finance.yahoo.com/streamer/1.0?s=NSMG.PK,USD=X&tt=1168007924&o=^DJI,^IXIC&marketid=us_market
1	8:55:34	5-Jan-07	http://my.yahoo.com/
1	8:57:14	5-Jan-07	https://BROKER.com/login
1	8:58:01	5-Jan-07	https://BROKER.com/home
1	8:58:16	5-Jan-07	https://BROKER.com/accounts/accountscombo
1	8:58:29	5-Jan-07	https://BROKER.com/alerts/alertinbox
1	8:58:50	5-Jan-07	https://BROKER.com/stocks/ownership/insiders.asp?symbol=ACAS
1	8:59:07	5-Jan-07	https://BROKER.com/portfolioview?portfolio=XXXXXXXX&_formtargetortfoliolist&traxui=ac_combo
1	9:01:33	5-Jan-07	https://BROKER.com/invest/createentry?Symbol=ACAS&ordertype=2&numshares=100

The broker has been anonymized by us and the account number was anonymized by the data provider, and for brevity we omit a limited number of fields containing demographic information.

We infer that in six minutes on the morning of January 5, 2007, this investor ("household 1") looked up a pink sheet stock (NSMG) on Yahoo! Finance; proceeded to login to a brokerage account; reviewed his/her accounts and inbox; reviewed insider ownership data on a (then-)Nasdaq stock ACAS; reviewed his/her portfolio; and then sold 100 shares of ACAS (order type=2 corresponds to a stock sale). Considering the nature of raw links, a long process of regular expressions programming is needed to extract usable fields.

Hence, we learn that this investor was concerned about insider ownership before trading a particular stock. We also know the information that is typically available on this broker's site that was not consulted, at least not in this interval, and that another stock was initially investigated but not traded, at least not in this interval. An important dimension of the data is that we can also see research that falls outside of the broker's own research ecosystem, such as analyst reports, SEC filings, Bloomberg news articles, etc. Given that there are many investors, many stocks, a time series of three months, and many categories of information that can be consulted online and documented using our methods, we hope to be able to make new statements about the nature of information gathered by online investors.

Preliminary Analyses

We include here a few very preliminary tables that give further context. We limit the data to households that made at least one trade of a stock that can be matched to a CRSP permno, and in these tables we limit to one broker.[2] For these households we show some basic summary

[1] In this level of anonymity, our data set resembles the treatment of other rich and unique investor data sets, such as online brokerage trades analyzed in several highly influential papers by Odean and coauthors from 1998 and onward, and the investor survey and matched broker trading records studied by Lease, Lewellen, and Schlarbaum in a series of articles in the 1970s.

[2] A natural question to address in the paper, but not here, is the extent to which these households are representative of online traders or traders in general.

statistics about their Internet usage, time spent on the broker's website, research types pursued, and most-searched tickers.

Specifically, Table 1 summarizes the data on the household-month level. In Panel A we summarize time spent online, time spent at the broker's website, and the number of browsing sessions, where one session is defined as a continous clickstream with no break longer than fifteen minutes. Panel B shows summary statistics regarding stock research and trading. The average household spends 1.35 hours per month researching and executes 4 trades. Both statistics are right skewed—the median household spends a quarter hour researching and executes two trades. Panel C shows a limited set of demographic data.

In Table 2 we show some major categories of research we have been able to extract using regular expression programming. The Table shows the total amount of time spent on a given type of research, as well as the number of households carrying out the type of research, and the total number of CRSP permno's captured. News and snapshots of a stock are the most popular types of research, but we also see a substantial amount of time spent on learning about stock fundamentals, and on charting.

Table 3 displays time spent in the same research categories, and adds the mean and median of time spent by households, conditional on the household spending non-zero time in the research category. Additionally, the table normalizes the mean and median amount spent in a given type of research category normalized with the number of trades by the same household.

Finally, in Table 4 we list the most researched tickers among the traders. For each ticker we include the total amount researched, the number of unique households researching the ticker, and the number of unique URLs pertaining to the ticker visited on the broker's website. 40 of the 170 households in this sample carry out research on Microsoft, while Altria is the stock with the longest cumulative amount of research time spent.

Directions Forward

The analysis above is a sketch of our first project. We would benefit from an NBER audience to discuss this paper and potential follow-up research we are pursuing. One follow-up research topic is how "research styles" differ across investors and how they match up to "trading styles." Another is how trading performance varies with the depth and nature of pre-trade research. Is more research by investors better, worse, or just a waste of time? Yet another topic is the extent to which there are cross-broker differences in investor research behavior and how the information that brokers choose to present shape their investors' trading behavior. An extreme recent example here is Robinhood, which presents only a limited amount of information. Our data complements that research using more typical brokers and a more representative set of investors.

Thank you for considering our proposal.

Table 1: Household-Month Level Summary Statistics This table summarizes the data on the household-month level. The sample is restricted to households that trade at least once with this broker. All time variables measured in hours. Dollar amounts measured in thousands. Panel A shows household-month level averages of time spent online, and time spent on the broker's website. A browsing session is defined as time spent online with no gap in activity longer than 15 minutes. Panel B. summarizes research and trading data. Panel C summarizes demographic data: household income and age of the head of household.

Panel A.

	mean	sd	p1	p10	p50	p90	p99	count
Hours online	42.90	28.97	2.49	11.85	36.11	78.60	137.73	170
Hours at broker website	4.36	7.16	0.10	0.29	1.50	12.44	36.97	170
Hours elsewhere	38.66	27.89	2.34	10.29	32.28	72.57	137.21	170
# Browsing sessions	104.59	56.69	13.00	33.83	95.54	172.50	303.00	170
Browsing session length	0.43	0.24	0.08	0.21	0.38	0.71	1.35	170

Panel B.

	mean	sd	p1	p10	p50	p90	p99	count
# Tickers researched	65.56	147.56	1.00	1.00	14.33	177.42	857.50	170
Hours researched	1.35	3.15	0.00	0.01	0.26	4.42	18.03	159
# Trades	3.97	5.94	1.00	1.00	2.00	9.67	33.00	170
# Tickers traded	2.67	3.19	1.00	1.00	1.50	6.00	18.00	170
# Trading sessions	2.54	2.89	1.00	1.00	1.50	5.63	16.00	170
Trading session length	1.07	1.71	0.09	0.23	0.69	1.94	5.41	170
Dollar amount	9.60	29.71	0.00	0.24	2.63	13.88	169.57	120
Share buys	0.49	0.41	0.00	0.00	0.50	1.00	1.00	170

Panel C.

	mean	sd	p1	p10	p50	p90	p99	count
Income	76.21	24.95	12.50	37.50	75.00	100.00	100.00	155
Age	47.62	14.39	22.00	30.00	47.00	65.00	86.00	161

Table 2: Total Research Time by Research Type. This table summarizes total hours spent on research in each of the listed categories. # Households refers to the number of households carrying out a given type of research. # Permno counts the number of unique Permno's that are researched in each of the categories.

Research Category	Hours	# Households	# Permno's
NEC	6.8	57	14
charts-NEC	58.8	96	519
charts-dividends	3.6	10	40
charts-earnings	3.4	6	23
charts-volume	10.8	43	171
fundamentals-NEC	19.8	56	290
fundamentals-earnings	4.7	44	203
fundamentals-insiders	1.9	32	74
news	216.9	97	2,010
prices	20.1	48	184
snapshot	272.9	154	1,534
technical	0.6	3	104
watchlist	1.7	43	113

Table 3: Total Research Time by Research Type, Normalized by Number of Trades. This table summarizes total time spent on research in each of the listed categories. # HH refers to the number of households carrying out a given type of research. The two columns labeled "Mean" measure the mean amount of minutes spent in the category by a household, and the mean amount of time spent in the category, normalized by number of trades. The two columns labeled "Median" measure the median amount of minutes spent in the category by a household, and median amount of time spent in the category, normalized by number of trades. The numbers in the last four columns are calculated conditional on the household doing any research in the given category.

			Mean		Median	
Research Category	Hours	# HH	Minutes	Minutes/Trades	Minutes	Minutes/Trades
NEC	6.78	57	7.13	1.92	1.60	0.45
charts-NEC	58.80	96	36.75	4.14	7.42	1.32
charts-dividends	3.58	10	21.47	5.27	1.63	0.69
charts-earnings	3.43	6	34.27	3.98	1.45	0.53
charts-volume	10.82	43	15.09	3.40	1.78	0.56
fundamentals-NEC	19.80	56	21.22	2.02	4.19	0.50
fundamentals-earnings	4.66	44	6.36	1.89	0.85	0.29
fundamentals-insiders	1.93	32	3.62	1.65	0.56	0.22
news	216.91	97	134.17	9.18	24.97	2.62
prices	20.06	48	25.08	1.99	4.80	0.39
snapshot	272.93	154	106.34	21.51	22.22	5.22
technical	0.64	3	12.77	7.85	0.68	0.29
watchlist	1.70	43	2.38	0.45	0.34	0.05

Table 4: Top Researched Tickers. Tickers researched by more than 15 individual households. # HH refers to the total number of households researching a given stock. Total hours refers to the cumulative time spent researching the stock. # Pages refers to total number of research pages visited.

Ticker	Current Name	Permno	# HH	Total hours	# Pages
MSFT	MICROSOFT CORP	10107	40	2.7	196
AAPL	APPLE INC	14593	33	2.1	219
CSCO	CISCO SYSTEMS INC	76076	30	1.3	176
PFE	PFIZER INC	21936	29	2.1	147
INTC	INTEL CORP	59328	27	3.8	262
GE	GENERAL ELECTRIC CO	12060	27	1.7	149
SIRI	SIRIUS X M HOLDINGS INC	80924	25	1.0	112
XOM	EXXON MOBIL CORP	11850	23	1.3	66
YHOO	YAHOO INC	83435	22	1.5	64
C	CITIGROUP INC	70519	22	0.6	74
T	A T T INC	66093	21	3.0	254
GOOG	ALPHABET INC	90319	21	1.8	170
MO	ALTRIA GROUP INC	13901	21	5.5	303
GS	GOLDMAN SACHS GROUP INC	86868	20	2.4	132
WMT	WALMART INC	55976	19	1.4	92
DELL	DELL INC	11081	19	1.0	74
AMGN	AMGEN INC	14008	19	1.1	119
VZ	VERIZON COMMUNICATIONS INC	65875	19	1.0	68
GM	GENERAL MOTORS CORP	12079	19	0.8	73
JPM	JPMORGAN CHASE CO	47896	19	0.9	38
LVLT	LEVEL 3 COMMUNICATIONS INC	85991	19	1.6	190
HPQ	H P INC	27828	18	0.9	62
JNJ	JOHNSON JOHNSON	22111	18	0.6	58
EBAY	EBAY INC	86356	17	1.5	171
DIS	DISNEY WALT CO	26403	17	1.0	88
F	FORD MOTOR CO DEL	25785	17	0.9	67
HD	HOME DEPOT INC	66181	17	0.5	37
NYX	N Y S E EURONEXT	91143	16	1.3	149
GPS	GAP INC	59010	16	0.7	43
BA	BOEING CO	19561	16	0.9	80