

The Lost World

Systematic Truncation of Current Financial Data and Its Impact on Research and Policy Making

Principle Investigator: Mao Ye

University of Illinois, Urbana-Champaign

Introduction

Recently, the size of trading data has increased exponentially due to computer-based high frequency trading and market fragmentation. The size of the data is already beyond the computing power of finance research using trade data. Here is a quote from the financial research page of Nanex:

“On Friday, Aug 5, 2011, we processed 1 trillion bytes of data for all U.S. equities, options, futures, and indexes. This is insane. A year ago, when we processed half of that, we thought it was madness. A year before that, when it was 250 billion bytes, we thought the same.”¹

The size of the data is currently a constraint for academic research. Some interesting questions cannot be addressed because of inadequate computing power. Some other questions can be addressed using more aggregated level of data, but important information is still missing. I also find that spurious inference can be drawn if a research is conducted using several existing datasets due to their aggregation and truncation. The computing power is also a constraint for policy making. On May 6, 2010, the Dow Jones Industrial Average experienced the biggest one day decline in the history (998.5 points, or about nine percent). Some stocks (e.g., Accenture, 3M) traded at pennies. Yet it takes five months for the SEC to figure out the reason due to the limitations of data processing ability:

“The reconstruction of even a few hours of trading during an extremely active trading day in markets as broad and complex as ours— involving thousands of products, millions of trades and hundreds of millions of data points—is an enormous undertaking. Although trading now occurs in microseconds, the framework and processes for creating, formatting, and collecting data across various types of market participants, products and trading venues is neither standardized nor fully automated.”²

¹ HFT is Killing the Emini: Enough Already! Research report from www.nanex.net

² U.S. Commodity Futures Trading Commission and U.S. Securities and Exchange Commission: Preliminary Findings Regarding the Market Events of May 6, 2010

Data

We have two unique datasets for our study: the first one is NASDAQ Totalview ITCH data, the second one is NASDAQ high frequency trading data. We also use the trade and quote data (TAQ), a widely used dataset in academic research. Because Totalview ITCH data provides more detailed and comprehensive information on trading, we can compare the Totalview ITCH and the TAQ to see the information missing in commonly used TAQ data.

The ITCH Totalview data includes the order entry, cancellation, and update as well as execution information for all the NASDAQ trades. The file is very large in size. We currently store all the files from 2010-2011 on PSC’s Blacklight supercomputer, and the file size is 7.5 TB. In fact, we have the data from 2000 till now. The file before 2010 is much smaller, but we estimate that we need probably 10-15 TB in total to store the complete series of our data from 2000-2011, and 10-15 more TB to store the file we generate during our analysis. We access the TAQ data from the Wharton School of Business’ research server. NASDAQ high frequency data is about 15 GB in size.

date	R	L	A	F	E	C	X	D	U
Panel A. Count of Each Message Type									
041511	7,829	169,358	107,990,296	4,399,554	5,693,004	80,777	1,089,360	107,792,080	22,933,350
041811	7,831	169,752	150,705,664	5,113,240	6,786,291	98,079	1,724,021	150,270,512	33,152,070
041911	7,843	169,409	118,482,480	4,387,052	5,847,933	75,845	1,165,250	118,113,288	23,020,108
042011	7,844	169,498	117,521,856	4,592,530	6,384,672	90,074	1,066,731	116,953,048	24,183,720
042111	7,869	171,355	99,653,168	4,431,085	5,558,716	75,509	884,554	99,603,136	19,334,540
042511	7,863	200,281	86,789,072	3,936,105	4,695,225	59,367	843,143	86,944,632	16,768,325
042611	7,863	170,373	103,546,776	9,097,849	6,031,116	75,790	1,001,254	107,843,600	20,230,176
042711	7,864	170,699	121,812,440	4,785,688	6,591,250	86,525	1,229,467	121,293,800	24,724,096
042811	7,865	170,359	111,012,856	4,782,514	6,405,544	82,825	1,182,839	110,645,312	23,833,836
042911	7,862	170,776	96,483,920	4,536,216	6,328,429	84,580	939,308	96,036,240	19,341,172
050311	7,864	171,030	141,371,248	5,183,981	7,592,976	93,245	1,615,920	140,445,184	27,142,764
050411	7,865	170,721	167,210,880	5,869,078	8,882,914	112,567	1,594,911	165,956,960	32,207,802
050511	7,870	171,101	196,167,312	6,304,186	9,495,514	126,278	1,761,253	194,762,560	40,643,184
050611	7,868	171,179	201,538,896	6,002,656	8,523,668	108,614	1,877,863	200,631,904	41,009,252
average	7,857	172,564	130,020,490	5,244,410	6,772,661	89,291	1,283,991	129,806,590	26,323,171

Table 1: Number of Observations for a Daily File in ITCH data.

The large size is due to two reasons. The first one is the speed to add and cancel orders at the speed of 10^{-9} second. Table 1 demonstrates that on an average day, the data has around 300 million observations. The sample period is from April 15, 2011 to May 6, 2011. (A detailed explanation of the variables contained in the table can be obtained from the author.) What I want to highlight is that most of these data points are addition of orders (A and F type) and cancellation (D type). Many orders are added and then cancelled quickly because computer algorithms want to use them to detect hidden liquidity or use them to manipulate the price of the market. It is often blamed that these quickly cancelled or added orders are the causes of the Flash

crash and many mini flash crashes. However, it is extremely difficult to study these quickly cancelled orders due to the limitation in computing power. For example, the Security and Exchange Commission is now very concerned with a type of market manipulation called “quote stuffing”, that is, a high frequency trader can send millions of messages to the stock exchange, which in turn cause a jam in the exchange’s computer system. By doing that, the high frequency trader can cause a delay in the stock exchanges’ computer system and the high frequency trader can benefit from the delay he causes. However, detecting “quote stuffing” needs super computing power. The reason is natural: because the numbers of messages in quote stuffing is too high to delay the exchange’s trading system, it is almost impossible to detect such a pattern using desktop computer. Put in another way, if we can find “quote stuffing” in regular desktop computer, the “quote stuffing” we find may not be large enough to block the exchange system.

Because we have the most comprehensive dataset on trading, we first want to compare our dataset with the common dataset used by other people, and see what kind of bias or omission can be caused using more condensed or truncated data. Currently, the most comprehensive dataset used in academia is TAQ data. TAQ does not carry the information on order submission and cancellation; it only provides the information on executed trades. To make things worse, TAQ only provide the E type information for trades of size of 100 shares or above. This is due to the caveats in the regulation on trade report standard: trades less than 100 shares are not reported. We have found the evidence that this regulatory leakage has been utilized by sophisticated traders by slicing and dicing their orders into small pieces. Tables 2 provide an example. On May 15, 2008, from 9:45:39:548 to 9:45:42:459, 600 shares are bought and 600 shares are sold for the stock BRCM. However, the buy and sell are all in one share unit. This trade left no trace in the consolidated tape and TAQ data.

Sequence	Symbol	Hour	Minute	Second	Millisecond	Shares	BuySell	Price	Type
1	BRCM	9	45	39	548	1	B	26.93	HN
2	BRCM	9	45	39	550	1	S	26.92	HH
3	BRCM	9	45	39	553	1	B	26.93	HN
4	BRCM	9	45	39	555	1	S	26.92	HH
5	BRCM	9	45	39	558	1	B	26.93	HN
6	BRCM	9	45	39	560	1	S	26.92	HH
.....									
1197	BRCM	9	45	42	452	1	B	26.93	HN
1198	BRCM	9	45	42	455	1	S	26.92	HH
1199	BRCM	9	45	42	457	1	B	26.93	HN
1200	BRCM	9	45	42	459	1	S	26.92	HH

Table 2: A sequence of 600 alternating buy and sell trades with each other. All of them are exempt from the requirement of the consolidated tape. The 600 buy trades are generated by high frequency traders

taking liquidity from the non high frequency trader (HN type). The 600 sell trades are generated by one high frequency trader take liquidity from another high frequency trader (HH type).

Project 1: Odd-lot Bias in Academic Data

Two natural questions arise once we know that TAQ data is only a truncated sample. First, what is the magnitude of missing trades in TAQ data? Second, what are the consequences for academic research and policy making using truncated data? We find that the median number of missing trades per stock is 19%, but for some stocks, the missing trades are as high as 66% of total transactions. This truncation may lead to significant bias or spurious inferences for academic research using TAQ data. Figure 1 demonstrates the time series of the missing trades in consolidated tape and TAQ data.

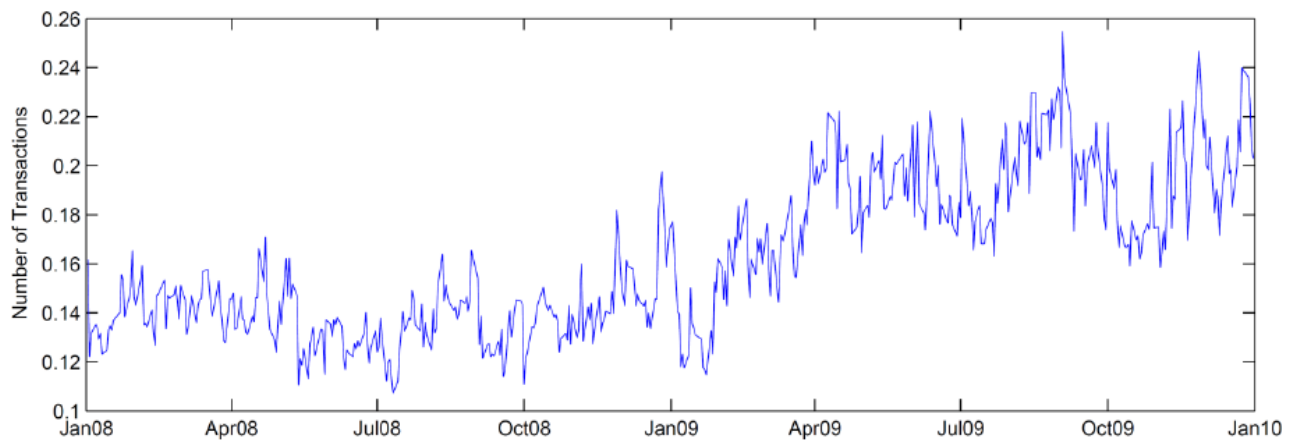


Figure 1: Number of Missing Trades in the Consolidated Tape and TAQ Data

More importantly, we find evidence that the missing trades are not mainly from small, native traders but traders with private information. Table 1 demonstrates the weighted price contribution, the standard measure of trade informativeness in the literature. According to this measure, an order size bucket is informative if its weighted price contribution is higher than its market share (Barclay and Warner (1993), Chakravarty (2001) Choe and Hansch (2005) and Alexander and Peterson (2007)). The classic result in Barclay and Warner (1993) is that orders less than 500 shares actually contribute negatively to price discovery, implying that they are coming from small, uninformed traders. The result has changed significantly in the high frequency trading era. Table 1 show that 80% of price discovery happens in trades of 100 shares or below. More importantly, 30% of the price discovery is related to the missing trades.

Trade size category	$WPC_{\text{return change}}$	$WPC_{\text{price change}}$	Shares of Trades	Shares of Volume
<100	0.306	0.354	0.158	0.034
100	0.504	0.497	0.54	0.281

200	0.053	0.041	0.117	0.121
300	0.022	0.01	0.041	0.065
400	0.014	0.016	0.025	0.053
500	0.013	0.006	0.024	0.062
100-500	0.652	0.615	0.791	0.633
501-900	0.018	0.012	0.027	0.099
901-1900	0.016	0.011	0.017	0.109
1901-4900	0.006	0.008	0.005	0.078
4901-9999	0.001	0.000	0.001	0.028
501-9999	0.042	0.031	0.051	0.313
≥ 10000	0.001	0.000	0.000	0.019

Table 2: Price Discovery, Share on Number of Trades and Volume for Each Size Category. $WPC_{\text{return change}}$ and $WPC_{\text{price change}}$ are the weighted price contribution measured by return change and price change. Both variables are measures of order informativeness or price discovery.

The missing trades propose significant challenges to use TAQ data. From 1993, 182 articles published in the top three finance journals use TAQ data. The research covers a wide range of fields such as asset pricing (see, e.g. Sadka(2006), Easley, Hvidkjaer and O'Hara (2002), among others), behavioral finance (see, e.g. Malmendier and Shanthikumar (2007), Barber, Odean and Zhu (2009), among others), corporate finance (see, e.g. Nimalendran Ritter and Zhang (2007), Krigman, Shaw and Womack (1999), Chen, Goldstein and Jiang (2007), among others) and real estate (see, e.g. Gentry, Kemsley and Mayer (2003)). The increased magnitude of missing trades in recent years presents significant challenges to a number of empirical measures designed in the past. One is order imbalances, which is often defined as the number of buys minus the number of sell. With missing trades in the data, we may easily sign a trading day with a buy imbalance using TAQ data when the true imbalance is sell imbalance. The other one is the \$5,000 cut-off for the individual trades. The literature started by Lee and Radhakrishna (2000) uses trades of \$5,000 as a proxy for trades from individuals. My current research demonstrates two challenges for this approach. First, small trades may not come from small traders. More importantly, the fact that TAQ data truncates all the observations less than 100 shares implies that we observe 0 individual trading for any stock with a price of \$50 or above. We can also see mechanical patterns of individual trades led by the \$5,000 cut-off. For example, for some stock with price \$49, we may see some individual trades, however, all the individual trades disappear when the price rises above \$51. The pattern is generated through data truncation. The magnitudes of these two problems are huge. Figure 2 shows that stocks with price higher than \$50 can constitute 70% of market value of all U.S. common stocks.

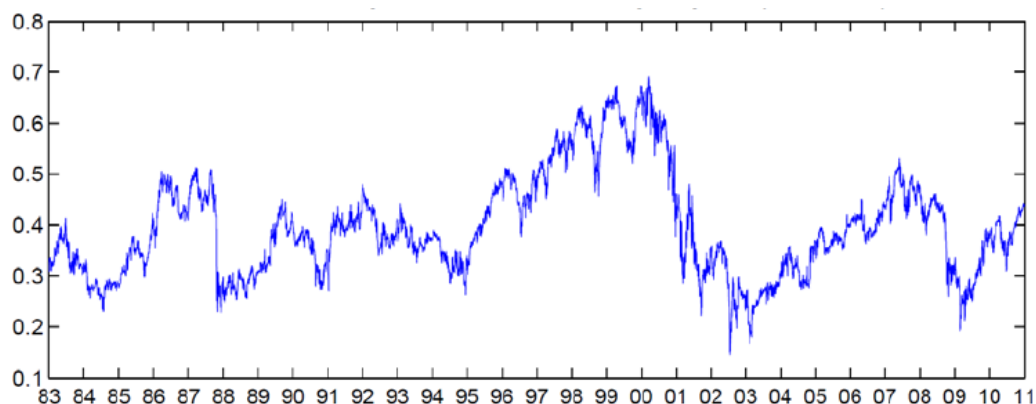


Figure 2: Percent of the Total Market Value for Which We Will Draw the Wrong Inference That There Is No Individual Trades. Although every stock should have trades from individuals, we would infer a stock has 0 individual trading for a day using the \$5,000 cut-off for individual trades once its price is above 50. This figure demonstrates the market value of those stocks as a percent of total market value from 1983 to 2011.

Project 2: Odd lots, Return Predictability and Information

The preliminary result of project 1 has shown that odd-lot trades are informative because the informed traders can use odd-lot trades to hide their information. Then the next question is how long this information advantage will last. We want to test the hypothesis that some types of odd-lots can predict the next day's return. We will focus on the trades of 99 shares, which are only 1 share below the minimum reporting requirement and see whether these trades can predict next day's return.

To do that, we need to search the entire dataset to look for the trades of size 99. As daily data can have 300 million observations, this work needs supercomputing. We will generate a panel data of the numbers of buy and sells in 99 shares for each stock on each day. Then we will follow the research on the daily return predictability (Pan and Potesman (2006) and Boehmer, Jones and Zhang (2008))

Project 3: Mini Flash Crash and Flash Crash

We got a dataset from Nanex on 18,000 mini-flash crashes that occurred from 2006-2010. These are the events that look like the Flash Crash except that they are smaller in magnitude. We hope the analysis of the mini-crashes will help us understand the causes and consequences of the flash crash happened on May 6, 2010. The flash crash is certainly more significant than a mini-crash,

but it is hard to draw any statistical inference from the Flash Crash because it only happened once. A repeated Flash Crash remains a possibility, albeit with a different catalyst.

As I mentioned in the introduction, it takes SEC several months to process several hours of data in flash crash. Analyzing 18,000 mini-flash crashes needs to process more lengthy data and thereby more computing power.

The questions we want to ask are

1. What is the cross sectional and time series pattern of mini-crash?

We have already found some pattern for mini-flash crash: mini-flash crash is more likely to happen at the beginning and the end of a trading day. Surprisingly, mini-flash crash is more likely to attack stocks with larger size, volume and lower spread.

2. What can be used to predict mini-flash crashes?

There have already been a number of explanations for the flash crash such as VPIN, volatility and news. However, it is very hard to test different hypothesis because the Flash Crash only happen once. In addition, a more important question is how to predict and prevent the next flash, which may be led by a different reason than the one that happened on May 6, 2010. We want to test the ability of these competing explanations for the Flash Crash to explain the mini flash crashes.

Potential Impact

I expect our research will result in several papers in leading journals. The preliminary results using our startup allocation on Blacklight have already aroused interests of academics, regulators and practitioners. I have received invitations from three universities to present the preliminary results, one of which is outside the United States (Vienna Graduate School of Finance, one of the leading research institutes in Europe and the world). I also received invitation for seminars from practitioners (Goldman Sachs Group) and stock exchange (NASDAQ). Just when I am preparing this proposal, one of my preliminary results was featured in Traders magazine. Here is a quote from the article when the magazine interviews Jim Toes, president and chief executive officer of the Security Traders Association:

*“He said the academic paper has already had a big impact on his organization’s members. ‘I think the industry is familiar with the report,’ Toes said. ‘And I haven’t heard anybody dispute the data.’”*³

³ Paper Points to Rise of Odd Lots, Traders Magazine, October 3, 2011

The paper and this article have also raise the attention of regulators. I am now discussing the details about a possible presentation in front of the U.S. Securities and Exchange Commission (SEC) .

Computing Methodology

Our analysis is divided into three steps: downloading, conversion and analysis. We need MATLAB and STATA for our analysis and we have already installed these two software in the Blacklight supercomputer. We receive one ITCH file each day and the estimation below applies to each daily file. The average size of the daily data is 10-12 GB, and each blade of the Blacklight is able to load up to two-day's data into memory simultaneously. Whole data are loaded into memory prior to processing to save file I/O time, and the loading/unloading still requires an enormous amount of file I/Os.

Download: We wrote a Linux shell script to download multiple data files from NASDAQ OMX server to PSC simultaneously in the background. The data, which are in binary format, was generously provided by NASDAQ to University of Illinois through a contract. The simultaneous downloads are packed into a single job. It approximately takes 1--3 days to download one-month data. Each month consists of about 20-25 trading days and the daily data size is 5-7 GB. So the size of a month of data is about 100-175 GB.

Conversion: This part again consists of two sub-tasks.

- a) Raw binary datafile => text file using MATLAB. It takes 24-48 hours on 1 core of Blacklight to convert a single daily data file, not considering the waiting time in a queue since it varies depending on server status. The output text file is ~10 – 11 GB in size from one daily dataset. The conversions of one-month data (equivalently 20--25 days of data) is carried out using job script level parallelism, by packing separate serial conversion jobs (one per core) into a single PBS job and submitted to a single blade.. We are currently converting one month of data at a time using a single blade of Blacklight since this is providing optimal throughput. We found that the waiting time is more than double when two or more blades are asked for. Therefore, using one blade is most time efficient way due to the waiting
- b) Text file 10-11 gb => STATA file (using STATA) : It takes 2--4 hours to convert a single text file into a STATA format. The STATA file for a single daily dataset is ~7-8 GB. Only two of these conversion jobs are packed into a single PBS job due to limited memory. In contrast to the MATLAB conversion, our current STATA code for this step needs enough memory to hold the input, output and intermediate data simultaneously in memory. We found that a blade, with 128 GB of RAM, easily handles two STATA

conversions simultaneously, but the program sometimes crashed when three or more STATA conversions are allocated on a single blade.

We used to submit 5--10 PBS jobs simultaneously; each of them is committed to handling one-month or 20-25 trading days data. When the server was relatively free, it runs most of them at the same time so that it consumes around 5,000 SUs per day. When the server was busy, however, it ran only one or two of them, consuming less than 1,000 SUs.

Analysis: This step is also performed with STATA. The Totalview ITCH file data- contains the order entry, update and cancellation message. The most important step is to combine the entry, update and cancellation message and then construct the limit order book at each time. The limit order book contains the information of all the remaining orders on the book, and provides an overview of the market condition at each point. Then we need to compute a number of empirical measures such as bid-ask spread, the slope of the limit order book, which may take another day. These empirical measures, eventually, are at daily level with one observation for each stock and each day. For example, if we are interested in whether the trades of 99-shares can predict the return of next day, we need to global search the daily file to know the number of buy and sell trades of the size 99 for about 7000 stocks in U.S. The return predictability literature usually uses 24 or 60 month of data, which means that we need search about 500-1250 daily file (each year has about 250 trading days). This means that we need to search 3 TB to 8 TB of data for the 99 share trades. This amount of data can never be loaded into desktop computer. The finance department of University of Illinois has a server with 16 cores and 32GB of memory. However, the server can only analyze one day of data per day. This means that we can analyze 5 years of data in 3-4 years. (There are less trading days than calendar days.) Therefore, support from XSEDE is essential for this type of computing.

The computing resources consumed by the analysis step vary depending on which type of analysis is done. Basically, more sorting and merging lead to the consumption of more computing resources. For example, suppose we estimate intraday stock prices for every 5-minute interval. This program takes 1-2 days to complete since it does sorting and filtering frequently. Moreover, the job can take even longer if we were to estimate prices for smaller intervals such as 1 minute or 30 seconds. Our estimate is that we need 5-10 days to estimate the price for every minute for each stock and 10-20 days to estimate the price for each 30 seconds.

Resource Request and Justification

With the help of PSC's Blacklight supercomputer, we have already achieved some significant results that will not only affect academic literature, but also policy making. The computing for our data is very intense: we consume around 1,000-5000 SUs per day. Therefore, we would like to request 1.5 million SUs for the next year. As per our discussion with the Blacklight supercomputer, we apply for SUs for both PSC's Blacklight and SDSC's Gordon. The reason we want to apply for both supercomputers is because Blacklight is sometimes very crowded with 95% of resources been used. However, financial researches can sometimes be very time sensitive. Suppose some significant events such as the Flash Crash happens, we would like to know the reason as soon as possible. Therefore, we want to have two supercomputers to increase our speed of analysis. Both Blacklight and Gordon have unique capabilities that suit to our data-intensive computing needs. Below is a summary of our SU and storage request:

	Blacklight	Gordon
SUs	1,000,000	500,000
Storage	25 TB	15 TB

We request large storage place for the following reasons. The total data we have has already 8-10 TB in size. The data is only raw data, and we may generate even larger data in the analysis. One example is limit order book. The ITCH data contains order addition, execution and cancellation information. However, the most important thing is to use the data to generate the limit order book. To be more specific, we want to know how the market looks like at each point in time: how many orders want to buy at different prices and how many orders want to sell at different prices. Therefore, each order addition, execution and cancellation changes market condition. The order addition, execution and cancellation instruction only has a line of message, however, it changes how the market looks like. For example, suppose now the best price that wants buy is 20 with 1000 shares with that price, and the second best price is 20.01 with 1000 shares and the third best is 20.02 with 1000 shares. Then a 3,000-share market sell order come in, which transact against buy orders of three prices levels. This market sell orders leads to three changes to the limit order book: the depth of market of 20, 20.01 and 20.02 all reduce to 0. In addition, we see three different executions of 1,000 shares each at 20, 20.01 and 20.02. Therefore, a simple market order generates a bunch of changes to the whole market, and we need to update them in the data. The limit order book data is much larger in dimension than the raw data. We will do our best not to store the data unless it is absolute necessary, but we believe we still need 25 TB and 15 TB in both supercomputers to perform the analysis.

In addition, we would also like to request Extended Collaborative Support Services (ECSS) to help improve our workflow and increase the efficiency of our computations. The file conversions cause multiple large chunks of disk I/O during the processing of a single set of data. The same data is written to a file by MATLAB and then read back into memory by STATA. The effective use of Blacklight's memory file system or Gordon's flash memory may significantly reduce the processing time of each data set. ECSS can help us achieve this.

Team Members

My coauthors and team members cover a wide range in the stages of their academic career. We have the former President of American Finance Association, Professor Maureen O'Hara, an associate professor, and two assistant professors, five Ph.D. students and two undergraduate students. Among the five Ph.D. students, two have already used the project as part of their dissertation.

Here are qualifications of the team member in this project

Mao Ye (the principle investigator of the project) is an assistant professor at the University of Illinois at Urbana Champaign. He was the winner of the \$15,000 NASDAQ dissertation fellowship in 2009 through national completion. He holds a Ph.D. degree from Cornell University and the first part of his dissertation has published in the most recent issue of *Journal of Financial Economics*.⁴ His research using the Blacklight Startup package has already been featured in the news in *Traders Magazine*, which discussed a possible reform on reporting standard for U.S. trades based on the result of his research.

Maureen O'Hara (coauthor of the first project) is the Robert W. Purcell Professor of Finance at the Johnson Graduate School of Management, Cornell University. Professor O'Hara has served as the President of the American Finance Association, as the President of the Western Finance Association, and the current President of the Financial Management Association. She has recently stepped down as the Executive Editor of the Review of Financial Studies. Professor O'Hara is also Chairman of the Board of Directors of the NYSE listed firm Investment Technology Group, Inc.

Xiaoyan Zhang is an associate professor at Purdue University. She got her Ph.D. degree from Columbia University in 2002. Before she joined Purdue, she worked as an assistant professor of finance at Cornell University.

Julie Wu is an assistant professor of finance at the University of Georgia. She got her Ph.D. degree in finance in 2008 from Texas A&M University. Her dissertation on short sell got cited in the Security and Exchange Commission's summary in the study of the short selling.

Dongming Sun is a Sixth year Ph.D. student in finance at the University of Illinois, Urbana-Champaign.

⁴ "Is Market Fragmentation Harming Market Quality?" *Journal of Financial Economics*, 3, 459-474

Jaehoon Lee and Chen Yao are fifth year Ph.D. students in finance at the University of Illinois, Urbana-Champaign.

Jianglin (Dennis) Ding and Mo Liang are second year Ph.D. students in finance at University of Illinois, Urbana-Champaign.

Siraj Nyakhar is a junior in the Department of Computer Science at the University of Illinois, Urbana-Champaign.

Wandi Wang is a sophomore in the College of Business at the University of Illinois, Urbana-Champaign.

References

- Alexander, G. J., and M. A. Peterson, 2007, "An Analysis of Trade-Size Clustering and Its Relation to Stealth Trading," *Journal of Financial Economics*, 84, 435-471.
- Barber, B. M., T. Odean, and N. Zhu, 2009, "Do Noise Traders Move Markets?" *Review of Financial Studies*, 22, 151-186.
- Barclay, M. J., and J. B. Warner, 1993, "Stealth Trading and Volatility: Which Trades Move Prices?" *Journal of Financial Economics*, 34, 281-305.
- Boehmer, E., Jones, C. M., & Zhang, X. (2008). Which shorts are informed? *The Journal of Finance*, 63(2), 491-527.
- Brogaard, J. 2010, "High Frequency Trading and Its Impact on Market Quality," working paper, Northwestern University.
- Chakravarty, S. 2001, "Stealth-Trading: Which Traders' Trades Move Stock Prices?" *Journal of Financial Economics*, 61, 289-307.
- Chen, Q., I. Goldstein, and W. Jiang, 2007, "Price Informativeness and Investment Sensitivity To Stock Price," *Review of Financial Studies*, 20, 619-650.
- Choe, H., and O. Hansch, 2005, "Which Trades Move Stock Prices in the Internet Age?" working paper, Pennsylvania State University and Seoul National University.
- Chordia, T., A. Goyal, and N. Jegadeesh, 2011, "Buyers Versus Sellers: Who Initiates Trades and When?" working paper, Emory University and University of Lausanne.
- Chordia, T., R. Roll and A. Subrahmanyam, 2002, "Order Imbalance, Liquidity, and Market Returns," *Journal of Financial Economics*, 65, 111-130.
- Chordia, T., and A. Subrahmanyam, 2004, "Order Imbalance and Individual Stock Returns: Theory and Evidence," *Journal of Financial Economics*, 72, 485-518.

- Easley, D., S. Hvidkjaer, and M. O'Hara, 2002, "Is Information Risk A Determinant of Asset Returns?" *Journal of Finance*, 57, 2185-2221.
- Getry, W. M., D. Kemsley, and C. J. Mayer, 2003, "Dividend Taxes and Share Prices: Evidence From Real Estate Investment Trusts," *Journal of Finance*, 58, 261-282.
- Hasbrouck, J., and G. Saar, 2010, "Low-Latency Trading," working paper, Cornell University.
- Hvidkjaer, S. 2006, "A Trade-Based Analysis of Momentum," *Review of Financial Studies*, 19, 457.
- Krigman, L., W. H. Shaw, and K. L. Womack, 1999, "The Persistence of IPO Mispricing and the Predictive Power of Flipping," *Journal of Finance*, 54, 1015-1044.
- Lee, C., and B. Radhakrishna, 2000, "Inferring Investor Behavior: Evidence From TORQ Data," *Journal of Financial Markets*, 3, 83-111.
- Malmendier, U., and D. Shanthikumar, 2007, "Are Small Investors Naive About Incentives?" *Journal of Financial Economics*, 85, 457-489.
- Nimalendran, M., J. R. Ritter, and D. Zhang, 2007, "Do Today's Trades Affect Tomorrow's IPO Allocations?" *Journal of Financial Economics*, 84, 87-109.
- O'Hara, M., and M. Ye, 2011, "Is Market Fragmentation Harming Market Quality?" *Journal of Financial Economics*, 3, 459-474.
- Pan, J., & Poteshman, A. M. 2006. The information in option volume for future stock prices. *Review of Financial Studies*, 19(3), 871.
- Sadka, R., 2006, "Momentum and Post-Earnings-Announcement Drift Anomalies: The Role of Liquidity Risk," *Journal of Financial Economics*, 80, 309-349.