Big data in Finance

Finance Research Group, IGIDR

July 25, 2014

Introduction

- Who we are?
- A research group working in:
  - Securities markets
  - Corporate governance
  - Household finance
- We try to answer policy questions using data and quantitative methods.

Case study

Paper: The causal impact of algorithmic trading on market quality
Authors: Nidhi Aggarwal & Susan Thomas
URL:
ifrogs.org/releases/ThomasAggarwal2014_algorithmicTradingImpact.html

The question

- Since 2000, escalating use of technology in trading on equities markets.
- AT now dominates exchanges worldwide. Concerns about liquidity, ‘flash crashes’, etc.
- Regulators all over the world are contemplating interventions on AT.
- In search of finding a market failure that justifies regulatory intervention, numerous researchers have asked: What is the effect of AT on liquidity and volatility?

Data

- Periods:
  - Pre co-lo: Jan '09 to Dec '09 (260 days)
  - Post co-lo: Jul '12 to Aug '13 (291 days)
- Criterion for securities selection: Study securities with at least 50 average daily trades in 2009 and 2012-13.
  This yields a set of 552 securities.
- Frequency used: Tick by tick trades and orders data.
- Data size analysed: 3.8 Terabytes of .csv text files.

Existing literature

1. A lot of the literature uses data from U.S. markets, which have highly fragmented liquidity.
   If AT adoption was taking place in different ways in different places, it becomes difficult to pin-point the starting point to measure the impact on the overall market.

Existing literature

1. A lot of the literature uses data from U.S. markets, which have highly fragmented liquidity.
   If AT adoption was taking place in different ways in different places, it becomes difficult to pin-point the starting point to measure the impact on the overall market.
2. Datasets often do not offer clear identification of AT. Without this, the measurement of AT activity is relatively weak.

Existing literature

1. A lot of the literature uses data from U.S. markets, which have highly fragmented liquidity.
   If AT adoption was taking place in different ways in different places, it becomes difficult to pin-point the starting point to measure the impact on the overall market.
2. Datasets often do not offer clear identification of AT. Without this, the measurement of AT activity is relatively weak.
3. Two issues that are worrisome:
   - Endogneity: If liquidity is a reason for ATs to choose to focus trading on a stock, and liquidity is an outcome to be measured, then which way does the causality flow?
   - Threats to validity: Was the change in market quality because of AT or other factors, such as macro-economics?

This paper

1. A *clean microstructure*: An exchange with 80% market share of all trading, one of the largest exchange in the world by transaction intensity.
2. Uses *an exogenous event*: Introduction of co-location services in Jan 2010, which directly affected AT.
3. *Data*: Every order explicitly tagged as "AT" or "non-AT" for every security at the exchange.

Consolidated trading

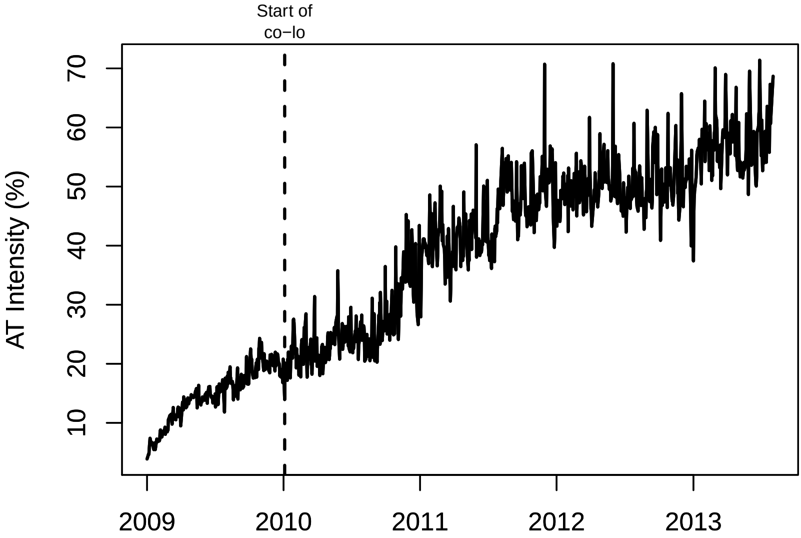
A big exchange by world standards

- In 2012 and 2013, NSE was the world's #1 exchange by number of trades on the equity market.
- The dollar value of these trades is small by world standards, but on this question, that is not important.

A natural experiment

- NSE launched co-location (co-lo) in January 2010.
- There was an S-shaped curve of adoption thereafter.
- This was an exogenous shock to AT intensity.

AT intensity between 2009-13

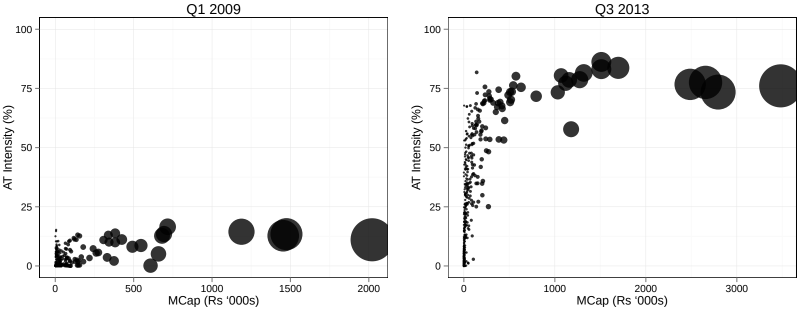


Issues in establishing causality

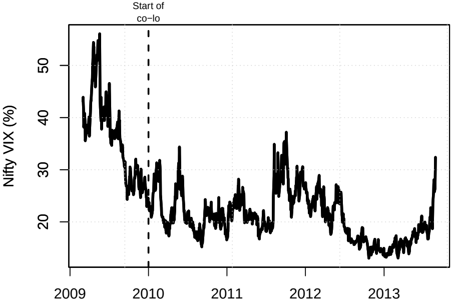
AT adoption at the firm level

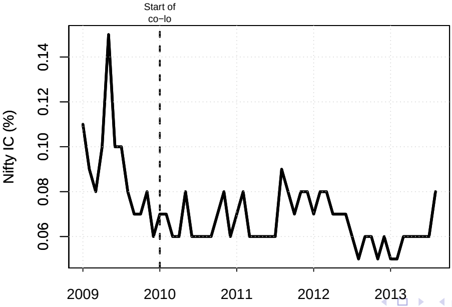
- Trading in some firms tend to become more AT while trading in some firms do not.
- Highly liquid firms tend to be more AT.
- There is the danger of selection bias here.

Variation in adoption of AT



Threats to validity: changes in macroeconomic conditions





I. Research design we use

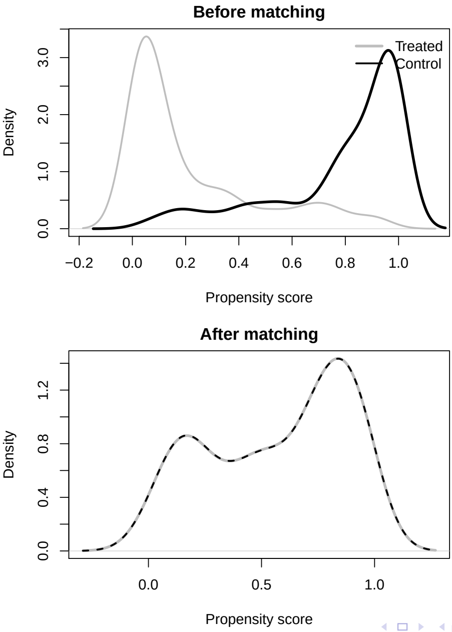
Matching at the security level

- We identify firms that got low AT adoption and firms that got high AT adoption.
- Use propensity score matching (PSM) to identify a matched sample.
- These are firms that are a lot like each other – but there was an almost experimental allocation where one group got the treatment of a surge in AT but the other group did not.

Obtaining set of matched firms

- After the launch of co-lo:
- Define
  - ‘Treated’: securities with Δ AT $> 70^{th}$ percentile value – 16.50% (276 firms)
  - ‘Control’: securities with Δ AT $< 30^{th}$ percentile value – 5.39% (276 firms)
  - Leave out firms in the middle.
- Propensity score matching:
  - Covariates: average daily values of market cap, price, floating security, turnover, number of trades (for the year 2009)

Density of the propensity score, before and after matching



Matching on macroeconomic conditions

- We capture changes in macroeconomic conditions by changes in the volatility of the market index (Nifty).
- We then match dates in the period before and after co-lo on volatility.
- This yields a set of dates in both periods which are alike in macroeconomic conditions.

Matching dates on macro-economic conditions

- Pick dates in the post co-lo period when market volatility matched the levels in the pre co-lo period (using Mahalnobis distance).
- This gives a set of 59 dates in each period that are alike.

Final sample characteristics

- Starting sample: Observations on 552 securities; Period of 260 days before co-lo and 291 days after co-lo.
- After matching on security level co-variates: 91 securities with high AT and 73 securities with low AT.
- After matching on macro-economic conditions: 59 days before co-lo and after co-lo.

II. Empirical setting

Market quality measures

- Liquidity
  1. Transactions costs
     1.1 QSPREAD (in %): (best ask - best sell)× 100 / mid-quote price.
     1.2 Impact cost (IC, %): execution cost of a market order at a size of Rs 25,000 relative to the mid-quote price.
  2. Depth
     2.1 TOP1DEPTH (in Rs.): Rupee depth available at the best bid and ask prices.
     2.2 TOP5DEPTH (in Rs.): Cumulated Rupee depth available at top five best bid and ask prices.
     2.3 DEPTH (# of shares): Average of the outstanding buy side and sell side number of shares.
     2.4 |OIB| (in %): Difference in buy and sell side depth as a percentage of the total depth, on average.

Market quality measures (contd..)

- Volatility
  1. Price risk, RVOL: Standard deviation of five-minutes returns.
  2. Price risk, RANGE: Difference in highest and lowest mid-quote price in a five-minutes interval.
  3. Liquidity risk, LRISK: Standard deviation of IC in five-minutes intervals.
- Efficiency
  1. VR: Ratio of 10-min variance of returns to 5-min returns
  2. KURTOSIS: Value of kurtosis in a five minute interval (absolute value).

What we find

Estimation using a Difference-in-Difference regression with matched securities and matched dates.

$$\begin{aligned}\text{MKT-QUALITY}_{i,t} &= \alpha + \beta_1 \text{AT-DUMMY}_i + \beta_2 \text{CO-LO-DUMMY}_t + \\ &\quad \beta_3 (\text{AT-DUMMY}_i \times \text{CO-LO-DUMMY}_t) + \epsilon_{i,t}\end{aligned}$$

	β_3	Expected sign
QSPREAD	-0.35^{+}	$-$
IC	-0.80^{+}	$-$
|OIB|	-14.34^{+}	$-$
DEPTH	-0.08	$+$
TOP1DEPTH	0.09	$+$
TOP5DEPTH	0.25^{*}	$+$
|VR-1|	-0.03^{+}	$-$
KURTOSIS	$\mathbf{6.81}^{+}$	$-$
RVOL	-2.88^{+}	$-$
RANGE	-19.86^{+}	$-$
LRISK	-0.02^{+}	$-$

What we find, contd.

- Kurtosis is the incidence of extreme returns. Does higher kurtosis mean more flash crashes?
- We analyse how frequently:
  1. Traded prices move by 2%, 5% or 10%
  2. In a period of 5 minutes

  before co-lo and after co-lo.
- What we find:

in %

	Pre co-lo		Post co-lo	
	High-AT	Not	High-AT	Not
TWO-EXCESS	33.35	33.46	29.36	36.84
FIVE-EXCESS	5.21	5.65	5.30	7.85
TEN-EXCESS	1.01	0.91	1.42	1.29

Some more facts

Are ATs consumer or providers of liquidity?

- A well-accepted hypothesis is that ATs trade at the cost of non ATs. They are assumed to take away liquidity, and do not supply it.

Are ATs consumer or providers of liquidity?

- A well-accepted hypothesis is that ATs trade at the cost of non ATs. They are assumed to take away liquidity, and do not supply it.
- We investigate this hypothesis. We define:
  - AT liquidity demand: % of trades that were *initiated* by ATs irrespective of who provided the liquidity.
  - We calculate out of total trades:
    - **AT2AT**:% of AT trades where ATs were liquidity suppliers.
    - **nAT2AT**:% of AT trades where non-AT supplied liquidity.
  - Separate and similar calculations for Non ATs.
- Done for the period between Jan 2013 to Dec 2013.

The facts

Overall cash market:

in %

	Mean	Median	SD	Min	Max
AT-DEMAND	37.45	**37.75**	3.86	11.22	48.28
AT-SUPPLY	39.92	**40.11**	5.17	8.76	53.24
AT2AT	18.60	**18.80**	3.47	1.42	29.24
AT2nAT	21.33	**21.34**	2.03	7.34	25.91
nAT2AT	18.85	**18.91**	1.19	9.80	21.75
nAT2nAT	41.23	**40.92**	5.41	29.60	81.44

Nifty stocks

in %

	Mean	Median	SD	Min	Max
AT-DEMAND	47.12	**47.30**	4.21	17.12	58.41
AT-SUPPLY	56.12	**56.36**	5.34	18.31	67.98
AT2AT	28.95	**29.13**	4.36	3.64	40.89
AT2nAT	27.17	**27.21**	2.08	14.67	32.86
nAT2AT	18.17	**18.22**	1.38	13.48	22.56
nAT2nAT	25.71	**25.27**	4.92	15.96	68.20

Case study: Conclusion

- The world has shifted from manual to computer-supported trading in an extremely short time.
- A major new phenomenon that requires analysis.
- All the regulators of the world are interested.
- Rapidly growing literature.
- Four identified flaws: (a) Fragmented microstructure (b) No clear identification in data infrastructure (c) Lack of exogenous change in AT and (d) Problems of causal identification.
- Our research design addresses these four problems.
- Main result: AT is good for market quality, but a) no significant impact on the depth though, b) no evidence in support of increase in flash crashes.

Thank you

anand.chirag@gmail.com
http://www.ifrogs.org/