Project HighLO: identifying anomalies in financial market data using particle physics tools

Axel Naumann, Lorenzo Moneta, Jonas Rembser, Joost Pennings, Tarek Alskaif, Marjolein Verhulst & Philippe Debie

June 29, 2023
Financial Markets

For example

- Stocks
- Futures and options
Challenge For Regulators and Researchers

- Effectively combatting manipulation in financial markets
- So much data, where to begin?
- How to convert *messages* in an understandable and analysable format?
- Irregular data
  - Traditional time series analyses difficult to apply
Project **High Energy Physics Tools in Limit Order Book Analysis** (HighLO)

- **Goal:**
  Interdisciplinary develop new methodologies to detect/identify manipulation

- **Data (so far):**
  - CME Group
  - All futures and options markets
  - 2015 ($\pm 30$ TB) and 2019-2020 ($\pm 300$ TB)
  - Time in **milli-** and **nanoseconds**
  - Limit orderbook and order data
## Key Team and Partners Project HighLO

www.highlo.org

### Key team

![Key team images]

### Affiliated Researchers

![Affiliated Researchers images]
Limit orderbook

- Marketplace showcasing all demand and supply for a specific product
- Limit orders
- Market orders

<table>
<thead>
<tr>
<th>Level</th>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>$55.00</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>$54.60</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>$53.50</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>$53.00</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>$50.00</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level</th>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$49.00</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>$48.00</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>$47.50</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>$46.00</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>$40.00</td>
<td>1</td>
</tr>
</tbody>
</table>
Why CERN?

- Data processing framework: ROOT
- 1EB data in ROOT format
- Scientifically analyse data; statistically correct
- Highly efficient and customizable
- Data storage support, machine learning, visualization
- Open source and supported for decades
- Love to look at masses of raw data!
## CERN vs. Limit Orderbook (LOB)

<table>
<thead>
<tr>
<th>CERN</th>
<th>LOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Look for <strong>unusual</strong> particles or for <strong>anomalies</strong> that do not fit the Standard Model</td>
<td>Look for <strong>unusual</strong> trading behavior or for <strong>anomalies</strong> that fit manipulative trading</td>
</tr>
</tbody>
</table>
First step: visualizations

Traditional timeseries analysis uses regular intervals in data

- E.g. seconds, minutes, hours, days, etc.

But our data is not regular

- Market activity is irregular
- Sometimes 20 updates in the same millisecond!

Traditional timeseries research aggregates to make it regular: lose a lot of information (data)!
The new visualization: Dec 2015 T-Bond

Limit Order Book ZBZ5 2015-11-12 (09:00 - 10:00)
Traditional vs. new visualization

Limit Order Book ZCZ5 2015-08-12 (09:00 - 12:00)
... the new visualization

Limit Order Book ZCZ5 2015-08-12 (09:00 - 12:00)
Visualizing Manipulation: the JPMorgan Case

JPMorgan Market Manipulation

- 2008 – 2016
- Spoofing
- Benefitted $172,034,790
- $311,737,008 market damages
- Settlement $920 million
Market manipulation: spoofing

- "The illegal practice of bidding or offering with **intent** to cancel before execution." (Dodd-Frank Act, 2010)

<table>
<thead>
<tr>
<th>Level</th>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ask (sell)</td>
<td>$53.50</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>$53.00</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>$50.00</td>
<td>9</td>
</tr>
<tr>
<td>Bid (buy)</td>
<td>$49.00</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>$48.00</td>
<td>4+1</td>
</tr>
<tr>
<td>3</td>
<td>$47.50</td>
<td>12</td>
</tr>
</tbody>
</table>

Price drops because of selling pressure

<table>
<thead>
<tr>
<th>Level</th>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ask (sell)</td>
<td>$53.50</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>$53.00</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>$50.00</td>
<td>9</td>
</tr>
<tr>
<td>Bid (buy)</td>
<td>$47.50</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>$47.00</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>$46.00</td>
<td>7</td>
</tr>
</tbody>
</table>

Small genuine order

<table>
<thead>
<tr>
<th>Level</th>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ask (sell)</td>
<td>$53.50</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>$53.00</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>$50.00</td>
<td>9</td>
</tr>
<tr>
<td>Bid (buy)</td>
<td>$49.00</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>$48.00</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>$47.50</td>
<td>12</td>
</tr>
</tbody>
</table>

Relatively large spoof order

<table>
<thead>
<tr>
<th>Level</th>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ask (sell)</td>
<td>$53.50</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>$53.00</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>$50.00</td>
<td>9</td>
</tr>
<tr>
<td>Bid (buy)</td>
<td>$47.50</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>$47.00</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>$46.00</td>
<td>7</td>
</tr>
</tbody>
</table>

Large order is canceled
JPMorgan Case: all market activity

Ultra T-Bond September 2015 contract

08:45:40 and 08:46:10 AM CT
Layered Spoofing

U.S. Treasury Bond September 2009 contract

07:47:10 – 07:47:30 AM CT
High-Frequency Data Analysis using the Tools from High-Energy Physics
1. How to measure the impact of an event on a time series using knowledge and tools from particle physics?

2. What is the impact of large limit order submissions on the transaction price?
Impact Plot – The Concept

Idea:

Measure the changes in a time series at fixed intervals before or after an event.

Example:

What is the average price change 50ms after a limit order submission?
## Data

<table>
<thead>
<tr>
<th>E-mini S&amp;P 500 futures</th>
<th>Crude oil Futures</th>
<th>US T-Bond Futures</th>
</tr>
</thead>
<tbody>
<tr>
<td>CME Globex code</td>
<td>ES</td>
<td>CL</td>
</tr>
<tr>
<td>Data range</td>
<td>2019/07/01 – 2020/07/01</td>
<td></td>
</tr>
<tr>
<td>Number of messages</td>
<td>4.9B</td>
<td>14.8B</td>
</tr>
<tr>
<td>Raw data</td>
<td>656 GB</td>
<td>1 140 GB</td>
</tr>
<tr>
<td>In ROOT</td>
<td>46 GB</td>
<td>80 GB</td>
</tr>
</tbody>
</table>
Impact plot – The Construction

1. Collect a set of triggers
   → Large limit order submissions
2. For each trigger, extract a time series
   → 200ms before to 200ms after the trigger
3. Overlap these time series (align the triggers)
4. For each time delay, build a distribution
   → 401 distributions, sequence of distributions
Impact plot – A Diagram

Trigger

M1 → M2 → M3 → M4 → M5 → M6 → M7 → M8 → M9 → M10 → M11 → M12 → ...

M1 → M2 → M3 → M4 → M5
M3 → M4 → M5 → M6 → M7
M7 → M8 → M9 → M10 → M11
Impact plot – E-mini S&P 500
Compare the impact plot against average market conditions

- The signal impact plot: As described
- The background impact plot: Randomly drawn message

→ Pairwise statistical testing, per time slice
→ A sequence of p-values, one for each delay
Impact plot – E-mini S&P 500
Impact plot – Crude Oil
Impact plot – US T-Bonds
Summary

Advantages compared to an impact analysis using VAR models

1. No data fitting or user-chosen modelling parameters
2. Better interpretation and visualisation
3. Linear computational cost and trivial multi-threading
4. Measure past and future time correlations

→ Interdisciplinary collaboration led to new techniques
An Agent-Based Financial Simulation of the Futures Commodity Market
Manipulation Detector Verification

Possible verification methods:

▪ Testing on proven cases (= labels) → Bias to existing work

▪ Spoof the market to generate labels → Not feasible

▪ Spoof in a simulation

→ Agent-based modelling using BioDynaMo
Research Questions

1. What is the accuracy of our detection method?
2. Who is hurt due to the spoofing?

“The economy needs agent-based modelling”
Next Steps
Next Steps for Project HighLO

- Interdisciplinary overview of spoofing
- Market manipulation/anomaly detection:
  - Spoofing: Expert vs. Machine
  - Detection tool & real-time monitoring
- Agent-based simulation:
  - BioDynaMo
  - Simulate financial markets
- Energy market collaboration
  - Dutch Authority for Consumers and Markets
  - ACER (EU)
Next Steps for Project HighLO (2)

Need:
- Interaction academics and practitioners
- Feedback and input from regulators, exchanges, surveillance agencies
- Reactive and proactive

**Goal:** International Expert Group on Market Surveillance
International Expert Group on Market Surveillance (IMS Group)

- 18 regulatory agencies across the world
- Combine science with industry through co-creation

Goal

- Combine academic research with expert knowledge to:
  - Tackle market surveillance challenges
  - Create consistency in regulation, definitions and interpretations of the law
  - Develop new surveillance methods and tools
Topics:

- Past and current surveillance systems
- Surveillance in energy markets
- Spoofing case
- Spoofing identification tools and metrics
- Challenges of algorithmic trading surveillance
- Research done by Project HighLO

Kick-off at CERN: March 15-16, 2023
Questions?

axel.naumann@cern.ch
marjolein.verhulst@wur.nl
philippe.debie@wur.nl

www.highlo.org
