Machine Learning and Electronic Markets

Andrei Kirilenko
Commodity Futures Trading Commission

This presentation and the views presented here represent only our views and do not necessarily represent the views of the Commission, Commissioners or staff of the Commodity Futures Trading Commission.
The Flash Crash - May 6, 2010
What did people think?

A survey conducted by Market Strategies International in June 2010 reports that over 80 percent of U.S. retail advisors believe that

“overreliance on computer systems and high-frequency trading”

were the primary contributors to the volatility observed on May 6.
Modern financial markets: machines (algorithms) against machines

- How do machines trade in electronic markets?
- How are the prices discovered?
- How robust are the markets populated by machines?

These are empirical questions.
Data: E-mini S&P 500 futures contract

Trades exclusively on the CME Globex electronic trading platform.

Highest dollar trading volume among U.S. equity index products.

Contributes the most to price discovery of the S&P 500 index.

Account-level, transaction-by-transaction data for all regular futures transactions over several days or weeks of data (e.g., August 2009).

Size: 7.2 million transactions during the month of August 2009 between 9:30 a.m. and 4 p.m. EST.

During August 2009: 26950 trading accounts of 346 brokers.
Approach: Machine-learning methods

Biclustering and Plaid models:
Can we identify different (algorithmic) trading strategies?

Hidden Markov Chain (Markov Switching) models:
Can we analyze the contribution of different (algorithmic) strategies to price discovery, volatility, liquidity?

Network analysis:
Can we gain insights into the dynamics of anonymous electronic markets?
What is biclustering?

A class of clustering algorithms that simultaneously clusters the rows and columns of a matrix to find homogenous submatrices.
The Static Plaid model

A regression-based clustering model which features a series of additive layers that comprise the matrix.

\[ \mu_k + \alpha_{ik} + \beta_{jk} \]
Time Series Variation

The Plaid model was originally designed for a single, static data matrix.

We observe a time series of such matrices.

The naïve method of ignoring time becomes overwhelmed with transient patterns.

Can we identify persistent trader groups and the important covariates that separate them consistently over time?
Dynamic Data: Cross-Section and Time Series
Smooth Plaid Models:

Instead of standard regression, we introduce penalties to filter out transient patterns and detect only the most persistent groupings/features.
Smooth Plaid Results:

The smooth plaid model clusters traders into five broad groups:

14 high frequency traders
271 slower market makers
7126 opportunistic traders
254 fundamental traders (buyers and sellers)
8021 small traders
Trading Volume

Estimated Log Average Volume (# Contracts)

- HFT
- Market Maker
- Opportunist
- Fundamental
- Noise

Day1, Day2, Day3, Day4, Day5, Day6, Day7, Day8, Day9, Day10, Day11, Day12, Day13, Day14, Day15, Day16, Day17, Day18, Day19, Day20, Day21
Net Position

Estimated Net Position (# Contracts Bought - # Contracts Sold)
Given the (reverse-engineered) identities of the trader types below, can we analyze the price discovery process?

14 high frequency traders
271 slower market makers
7126 opportunistic traders
254 fundamental traders (buyers and sellers)
8021 small traders

The financial econometrician’s favorite occupation:

Forecasting returns!
Let’s start with a simple cut:
High frequency traders (HF) and everyone else (LF).
Remember, retail advisors think that HF traders caused the Flash Crash!

HFTs appear very different.

<table>
<thead>
<tr>
<th>Group</th>
<th>Median Avg P&amp;L (daily)</th>
<th>Median TPS (daily)</th>
<th>Median Trades (month)</th>
<th>Median Volume (month)</th>
<th>Mean (SD) Inventory (daily)</th>
<th>Mean (SD) Trade Size (month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Frequency</td>
<td>22768.25</td>
<td>0.761</td>
<td>373853</td>
<td>1498920</td>
<td>5.77 (90.73)</td>
<td>4.26 (1.68)</td>
</tr>
<tr>
<td>Low Frequency</td>
<td>-6</td>
<td>0.001</td>
<td>77</td>
<td>132</td>
<td>-0.01 (69.00)</td>
<td>2.69 (2.63)</td>
</tr>
</tbody>
</table>
How do high frequency and low frequency traders trade with each other?

<table>
<thead>
<tr>
<th></th>
<th># of transactions</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Frequency ➔ Low Frequency</td>
<td>3326581</td>
<td>0.460438</td>
</tr>
<tr>
<td>Low Frequency ➔ High Frequency</td>
<td>1568438</td>
<td>0.21709</td>
</tr>
<tr>
<td>High Frequency ➔ Low Frequency</td>
<td>1574267</td>
<td>0.217897</td>
</tr>
<tr>
<td>High Frequency ➔ High Frequency</td>
<td>755538</td>
<td>0.104575</td>
</tr>
</tbody>
</table>
Transition matrix

<table>
<thead>
<tr>
<th></th>
<th>LF $\rightarrow$ LF</th>
<th>LF $\rightarrow$ HF $\rightarrow$ LF</th>
<th>HF $\rightarrow$ HF</th>
</tr>
</thead>
<tbody>
<tr>
<td>LF $\rightarrow$ LF</td>
<td>79.52%</td>
<td>20.17%</td>
<td>0.31%</td>
</tr>
<tr>
<td>LF $\rightarrow$ HF $\rightarrow$ LF</td>
<td>21.26%</td>
<td>68.38%</td>
<td>10.35%</td>
</tr>
<tr>
<td>HF $\rightarrow$ HF</td>
<td>1.73%</td>
<td>42.70%</td>
<td>55.57%</td>
</tr>
</tbody>
</table>

The transition matrix is quite stable:
- Convergence in 1 to 1½ hours in calendar time or
- 80,000 to 100,000 transactions in transaction time
Forecasting returns:

\[ r_t = X^n \cdot r_t \quad (1) \]

\[ r_t = r_t^* + u_t \quad (2) \]

Markov Model:

\[ r_t^* = X \cdot r_{t-1}^* + \xi_t \]

1. Use 1½-hours of data to construct the transition matrix
2. Update the transition matrix (from time zero)
3. Update the transition matrix with rolling window

Naïve model: Transition probability is 1/3.

Random Walk.
$R^2 = 0.11$: We can forecast one-period returns!
Why are we able to forecast returns?

High Frequency Traders really stand out in the data.

Their actions are almost deterministic.

They contribute a predictable component to prices/returns.

That’s why we can forecast returns.

How do prices form in electronic markets?
How do prices form in anonymous electronic markets?

In limit order markets, the process of finding market clearing prices is fundamentally different from that of a standard demand-supply auction.

The standard market-clearing price arises from an aggregation of supply and demand schedules for all market participants.

In limit order markets, there is neither a uniform market-clearing price nor a time when the limit order market clears.

The price discovery process in a limit order market can be thought of as a sequential aggregation process.

This process can be represented as a graph of bilateral executions among all market participants - a trading network.
Price discovery (matching) represented as a trading network.
## Actual trading networks

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CEN</strong></td>
<td>0.920</td>
<td>-0.092</td>
</tr>
<tr>
<td><strong>AVDEG</strong></td>
<td>2.027</td>
<td>3.269</td>
</tr>
<tr>
<td><strong>SDDEG</strong></td>
<td>8.134</td>
<td>5.960</td>
</tr>
<tr>
<td><strong>INOUT</strong></td>
<td>-0.470</td>
<td>-0.101</td>
</tr>
<tr>
<td><strong>AI</strong></td>
<td>0.353</td>
<td>0.559</td>
</tr>
<tr>
<td><strong>CC</strong></td>
<td>0.001</td>
<td>0.049</td>
</tr>
<tr>
<td><strong>LSCC</strong></td>
<td>0.014</td>
<td>0.016</td>
</tr>
<tr>
<td><strong>E</strong></td>
<td>75</td>
<td>103</td>
</tr>
<tr>
<td><strong>Returns</strong></td>
<td>0.059</td>
<td>-0.019</td>
</tr>
<tr>
<td><strong>Range</strong></td>
<td>0.059</td>
<td>0.059</td>
</tr>
<tr>
<td><strong>Volume</strong></td>
<td>1104</td>
<td>1343</td>
</tr>
<tr>
<td><strong>Duration</strong></td>
<td>0</td>
<td>12</td>
</tr>
</tbody>
</table>
Can network variables describe market dynamics?

Divide transaction data into consecutive periods of transaction time.

Construct networks for each period of transaction time.

Compute two sets of variables: Network variables for each network and Financial variables for each period of time.

See if the two sets of variables are statistically related.

Check for causality.
Financial networks: Definitions

An edge is defined as the occurrence of trading between two unique counterparties within a specified period of time.

The direction of an edge:
(Trading Network) IN for buy and OUT for sell.
(Liquidity Network) IN for aggressive and OUT for passive

Examples: A sold 3 contracts to B for $10.50
B initiated a transaction for 2 contracts with A at $10.20
Network Variables

Individual node: Centrality – How many trading partners do you have?
Centralization (CEN): Take centrality measures for each node and calculate how unequal is the whole degree distribution of edges.

Pair of nodes: Assortativity – Are those you trade with similar to you?
Assortativity Index (AI): Take four correlation coefficients and calculate a compound measure.

Triple of nodes: Clustering – Are those you trade with trade with each other?
Clustering Coefficient (CC): Ratio of closed triplets to connected triples.

Network components: How connected is the whole network?
Large Strongly Connected Component (LSCC): The largest subset of nodes such that any node can reach any other node by traversing edges.
Results

Network variables are statistically significantly related to financial variables (returns, volatility, volume, intertrade duration).

Network variable that quantifies centrality (star-shaped pattern) has a very high correlation with returns.

Network variable that quantifies the assortativity of connections (diamond-shaped pattern) has a high correlation with volatility.

Network variables strongly Granger-cause trading volume, realized volatility and intertrade duration, but are not Granger-caused by them.

Cool. But is all of this useful?
May 6, 2010 - The Flash Crash: E-mini S&P 500 Volume and Price
Trader Categories

These were constructed manually, but note the similarities!

- High Frequency Traders (16)
- Intermediaries (179)
- Fundamental Buyers (1263)
- Fundamental Sellers (1276)
- Opportunistic Traders (5808)
- Small Traders (Noise) (6880)

Did High Frequency Traders do it?
Trader Categories

**May 3**
- High Frequency Traders
- Opportunistic Traders and Intermediaries
- Fundamental Sellers
- Fundamental Buyers

**May 4**
- High Frequency Traders
- Opportunistic Traders and Intermediaries
- Fundamental Sellers
- Fundamental Buyers

**May 5**
- High Frequency Traders
- Opportunistic Traders and Intermediaries
- Fundamental Sellers
- Fundamental Buyers

**May 6**
- High Frequency Traders
- Opportunistic Traders and Intermediaries
- Fundamental Sellers
- Fundamental Buyers
HFTs do not accumulate inventory larger than 4500 contracts!
The Flash Crash: This Market is a Complex Adaptive System

13:32  A large fundamental seller initiates a sell program
13:42  HFTs reverse the direction of their trading (start selling)
13:45  “Hot Potato”: Lack of Fundamental and Opportunistic Buyers
13:45:28 - 13:45:33  5 second trading pause
13:45:33 – 13:45:58  Prices stabilize
13:46  Fundamental and Opportunistic Buyers lift prices up
14:08  Prices are at the 13:32 level
The Complexity of the Electronic Markets

- **Price feed**
- **Fill information (Order matched)**
- **Heartbeat**
- **Quote/Order acknowledgment**

**Exchange**
- Exchange sets limits
- FCM/broker-dealer sets limits
- Trading firm sets limits

**Firm/Dealer**
- Exchange pre-trade controls

**Proprietary/Vendor system**
- Risk limits set at exchange level
- Risk limits set at trading server
- Risk limits set at exchange level

**Risk Manager**
- Risk limits on GUI

**Drop Copy**
- Settlement Data
- Position

**Trading Firm**
- Order originated
- Log in

**Trading Server**
- Co-location Facility
- Internet / VPN access

**Clearing House Server**
- Risk limits monitored set at trading server
An Agent Based Simulation Model

- **Trade Speed**

- **Order Price Selection**

  - **Trader CDF**

  - Probability vs. Ticks from Best Bid/Ask

- **Inventory Constraints**

  - # Contracts

- **Order Size**

  - **Order Size CDF**

  - Probability vs. Number of Contracts
We Can Simulate The Flash Crash!

<table>
<thead>
<tr>
<th>Trader Types</th>
<th>Number of Traders*</th>
<th>Speed of Order</th>
<th>Inventory Constraints</th>
<th>Market Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Small</strong></td>
<td>200 (42%)</td>
<td>2 hours</td>
<td>None</td>
<td>1%</td>
</tr>
<tr>
<td><strong>Fundamental Buyers</strong></td>
<td>40 (8%)</td>
<td>1 minute</td>
<td>None</td>
<td>9%</td>
</tr>
<tr>
<td><strong>Fundamental Sellers</strong></td>
<td>40 (8%)</td>
<td>1 minute</td>
<td>None</td>
<td>9%</td>
</tr>
<tr>
<td><strong>Intermediaries</strong> (Market Makers)</td>
<td>6 (1.3%)</td>
<td>20 seconds</td>
<td>-4 – 4</td>
<td>10%</td>
</tr>
<tr>
<td><strong>Opportunistic</strong></td>
<td>183 (39%)</td>
<td>2 minute</td>
<td>-4 – 4</td>
<td>33%</td>
</tr>
<tr>
<td><strong>High Frequency</strong></td>
<td>3 (0.6%)</td>
<td>2 second</td>
<td>-20 – 20</td>
<td>38%</td>
</tr>
</tbody>
</table>

* The simulated market is 1/32 the size of the real market for computational tractability
More importantly: Can we design risk safeguards to prevent future Flash Crashes?
Conclusions

Regulators need to know what type of trading strategies are deployed in anonymous electronic markets, when and how.

Algorithmic traders (machines) need to know what type of trading strategies are deployed in anonymous electronic markets, when and how.

Investors need to know what type of trading strategies are deployed in anonymous electronic markets, when and how.

How do we find out?: Use machine-learning methods to learn about the machines.

Once we know, we can collectively come up with solutions to keep the markets liquid, fair, and free of abuse. Design safeguards against malfunctions.