Introduction to market microstructure and heterogeneity of investors

Fabrizio Lillo

https://fabriziolillo.wordpress.com

Dipartimento di Matematica, Università di Bologna
Porta di Piazza San Donato 5, 46126 Bologna, Italy

Varenna, July 20-21, 2018
Financial markets

• Financial markets allow two classes of agents to meet:
  – Entrepreneurs: who have industrial projects but need funding
  – Investors: who have money to invest and are ready to share profits and risks of the projects

• Therefore financial markets are systems where a large number of investors interact through trading to determine the best price for a given asset.

• From this point of view financial markets can be seen as a collective evaluation system.
Trading Technology Evolution

Phone

Financial Calculators

Spreadsheets

Screen Trading

Networking, Internet, Cloud Computing

Algo Trading

The New York Stock Exchange today

Trading Pit
Financial Data Science

The availability of large datasets is changing our understanding of the financial industry, opening new business opportunities, creating new sources of risk, also at the system level.

**New technological, conceptual, methodological challenges**

- Huge speed of data production: methods and technology
- Extracting information from complex sources (for example texts or tweets)
- Combining datasets of different origin
- From macro to micro: classification and prediction
- FinTechs (Cryptocurrencies, Blockchain, P2P lending, etc)
- Beyond individuality: the role of interaction among financial entities
- **Stakeholders**
  - Banks
  - Financial intermediaries
  - Regulators (Central Banks)
  - ….
## Financial Data

Extract informations and making prediction
Combine different kind of data

<table>
<thead>
<tr>
<th>Market data</th>
<th>Non public data</th>
<th>Public data</th>
</tr>
</thead>
<tbody>
<tr>
<td>price, volatility, balance-sheet,</td>
<td>portfolios composition, trading decisions of</td>
<td>News, Twitter, Blogs</td>
</tr>
<tr>
<td>agency rating, interbank lending</td>
<td>investors, companies relationship</td>
<td></td>
</tr>
</tbody>
</table>

Extremely fast production of data
Non uniform sampling
Aggregation at different time scales
The 5 V’s of Big Data:

**Volume**
- Terabyte
- Records/Arch
- Tables, Files
- Distributed

**Variety**
- Structured
- Unstructured
- Multi-factor
- Probabilistic
- Linked
- Dynamic

**Velocity**
- Batch
- Real/near-time
- Processes
- Stream

**Veracity**
- Trustworthiness
- Authenticity
- Origin, reputation
- Availability
- Accountability

**Value**
- Statistical
- Events
- Correlations
- Hypothetical

**Variability**
- Changing data
- Changing model
- Linkage
Velocity (and Volume)
Automated trading: a precursor

**Figure:** Cover of the 1971 article by Fischer Black on Financial Analyst Journal
Financial Markets Today

➢ Low Latency
➢ Co-Location & Proximity
➢ High Performance Computing
➢ Systemic Instability

- Data with nanosecond (!) resolution
- $10^6$ market events per day per stock
Market microstructure

• Market microstructure “is devoted to theoretical, empirical, and experimental research on the economics of securities markets, including the role of information in the price discovery process, the definition, measurement, control, and determinants of liquidity and transactions costs, and their implications for the efficiency, welfare, and regulation of alternative trading mechanisms and market structures” (NBER Working Group)
Floor market

- Old market structure
- Now used only in US commodity futures markets (Chicago Board of Trade, New York Mercantile Exchange, Chicago Mercantile Exchange)
- Little transparency on data
- Bund futures at London International Financial Futures Exchange (LIFFE) and Eurex (1997)
Dealer market

- A dealer is an intermediary who is willing to act as a counterparty for the trades of his customers
- Foreign exchange, corporate bond, swap markets
- Customers cannot typically place limit orders
- A large customer can have relationships with many competing dealers
- Low transparency: quotes in response to customer inquiries and not publicly available
- Interdealer trading is also important (for dealer inventory management). Limit order book for FX
- Dealers facilitate large (block) trades. Upstairs or off book market
Dynamics of Limit Order Book

- **Buy limit orders**
  - $B_n$
  - $V_n^B$
  - Volume

- **Price**
  - $P_n$
  - $g_n^B$
  - $g_n^A$
  - $V_n^A$

- **Sell limit orders**
  - Volume

The diagram illustrates the distribution of buy and sell limit orders across the price range with specific bars representing their volume and spread.
Dynamics of Limit Order Book

\[ P_n A_n B_n \ v_n^A \ v_n^B \ \epsilon_n = +1 \]

- **Sell limit orders**: \( P_n A_n B_n \ v_n^B \)
- **Buy market order**: \( \epsilon_n = +1 \)
- **Volume**

The diagram illustrates the distribution of limit orders and a buy market order on the price axis. The order book shows the balance between buy and sell orders with volume bars representing the quantity of each order type at different price levels.
Dynamics of Limit Order Book

\[ P_{n+} A_{n+} B_{n+} v_{n+} B_{n+} v_{n+} A_{n+} \varepsilon_{n} = +1 \]

- Sell limit orders
- Price
- Volume
- Buy limit orders
- Buy market order \( \varepsilon_{n} = +1 \)
- Price
- Sell limit orders
Dynamics of Limit Order Book

- New limit order
- Buy limit orders
- Sell limit orders
- Volume
- Price
Dynamics of Limit Order Book
A snapshot of the book display

<table>
<thead>
<tr>
<th>Company name</th>
<th>Trade type indicator</th>
<th>Company code</th>
<th>Volume weighted average price of today's trading</th>
<th>Total of today's shares traded (order book only)</th>
<th>Previous day's closing price</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC Holdings</td>
<td></td>
<td>ABC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMS</td>
<td></td>
<td>200,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sector FT10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GBX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Currency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GBX = pence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GBP = pounds</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EUR = euros</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>International security number</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Highest &amp; lowest prices of the day on and off the order book</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total shares traded</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of buy orders at the best price</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yellow strip</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total volume of buy orders currently on the bid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buy market order volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volume at best bid price</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulative order book price &amp; volume information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base price – the uncrossing price or if no uncrossing price the next automatic trade</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buy order</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order price per share</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sell market order volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sell order</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volume at best offer price</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trade Hi</th>
<th>Open</th>
<th>Trade Lo</th>
<th>VWAP</th>
<th>Current</th>
<th>Curr Hi</th>
<th>Curr Lo</th>
</tr>
</thead>
<tbody>
<tr>
<td>530</td>
<td></td>
<td>517</td>
<td></td>
<td>520</td>
<td>530</td>
<td>520</td>
</tr>
<tr>
<td>524½</td>
<td></td>
<td>524</td>
<td></td>
<td>524½</td>
<td>+4 ½</td>
<td>+12 ½</td>
</tr>
<tr>
<td>523</td>
<td></td>
<td>523</td>
<td></td>
<td>523</td>
<td>520</td>
<td>520</td>
</tr>
<tr>
<td>522.86</td>
<td></td>
<td>521</td>
<td></td>
<td>521</td>
<td>526½</td>
<td>20,000</td>
</tr>
<tr>
<td>521.49</td>
<td></td>
<td>519</td>
<td></td>
<td>519</td>
<td>529</td>
<td>50,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BUY</th>
<th>SELL</th>
</tr>
</thead>
<tbody>
<tr>
<td>TVol 543,906</td>
<td>Tvol 702,746</td>
</tr>
<tr>
<td>Base 520</td>
<td>MOVol</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>20,000</td>
<td>10,000</td>
</tr>
<tr>
<td>524 - 525</td>
<td>524</td>
</tr>
<tr>
<td>525</td>
<td>10,000</td>
</tr>
<tr>
<td>10,000</td>
<td>2</td>
</tr>
<tr>
<td>526.00</td>
<td>525.00</td>
</tr>
<tr>
<td>526.74</td>
<td>526.89</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
<tr>
<td>525.49</td>
<td>525.34</td>
</tr>
<tr>
<td>525.34</td>
<td>525.74</td>
</tr>
</tbody>
</table>
From Ponzi et al. 2009
According to a 2012 study of the AMF, the resting time of orders sent to assets traded at the Euronext-Paris has the following properties:

- 45% of orders stay in the order book less than a second
- 26% of orders stay in the order book less than 100 millisecond
- 11% of orders stay in the order book less than 5 millisecond
- 4% of orders stay in the order book less than 1 millisecond
- What is the informative content of these orders?

Exhibit 1: Quote Stuffing: Heineken, 2nd May, 2011

Exhibit 2: Quote Stuffing: Telefonica, 10th August, 2012

Source: Credit Suisse AES Analysis
Endogenous instabilities
the Flash Crash: May 6, 2010

Just a normal day at the nation's most important financial institution...

I've got a share that could really excel.
Excel? Really excel?
Sell?
Sell?
Sell!
Sell!
Sell!
Sell!

This is madness! I can't stay here anymore.

Good bye!

Buy!
Buy!

I've got a share that could really excel.

Sell?
Sell!
Sell!

Buy?
Buy!

Good bye!

Endogenous instabilities

Momentary Lapse

Stock markets plunged suddenly yesterday afternoon and gained speed as computer programs prevented losses. But almost as quickly, the market recovered much of the decline.

Source: Bloomberg

Close 10,520.32
-3.2%
Sometimes liquidity evaporates...

**Figure:** Limit Order Book dynamics from Mercato Titoli di Stato (MTS. Colored horizontal lines correspond to quotes from different participants. The mid price (black horizontal lines) separates the bid and ask side. No trades reported.

Exogenous instabilities
(Twitter Flash Crash, 23rd April 2013)
The limit order book during the Twitter Flash Crash

Figure: Source Nanex
Modelling the dynamics of systemic instabilities

- A jump is a price movement which is abnormally large with respect to the statistical properties of prices in the recent past
- 140 highly liquid stocks traded in the US financial markets
- High frequency jumps -> one minute price changes

Focus on
- The self and cross excitation of instabilities (jumps) among stocks
- The role of exogenous drivers (financial news) versus the endogenously generated instabilities
- How the financial markets have changed in the last fifteen years. Role of High Frequency Trading, Algorithmic Trading, and market reforms (RegNMS, MiFid, etc)
Data and jumps identification

- 140 highly liquid stocks in the US equity markets in 2001-2013
- One minute returns
- Jump of **Threshold** $\theta$:
  \[
  \frac{|r|}{\sigma} > \theta
  \]
- For volatility estimation we use the realized bipower variation (Barndorff-Nielsen and Shephard (2003, 2004))
  \[
  \hat{\sigma}^2_{bv,t} = \mu_1^{-2}|r_i||r_{i+1}| = \mu_1^{-2} \alpha \sum_{i>0} (1 - \alpha)^{i-1}|r_{t-i}||r_{t-i-1}|,
  \]
  with $\mu_1 = \sqrt{\frac{2}{\pi}} \approx 0.797885$ and $\alpha = 0.032$ (gives 50% of weight to the closest 22 observations).
- **Multiplicity** $M$ of a cojump: number of stocks simultaneously jumping (i.e. in the same minute).
Total number of jumps

The total number of jumps and the number of single asset jumps has actually declined in recent years.
Historical evolution of market instabilities
Flash Crash (6th May 2010)
Evolution of systemic instabilities

![Graph showing event frequency over years with different thresholds for instability.]
Exogenous events are less than 50% of all instabilities
Modelling instabilities with Hawkes processes

Probability per unit time of an event (jump)

A point process $N_t$ is called a Hawkes process if it is a linear self-exciting process, defined by the intensity

$$\lambda(t) = \mu + \sum_{t_i < t} \nu(t - t_i) = \mu + \sum_{t_i < t} \alpha e^{-\beta(t-t_i)}$$

where $\mu$ is a deterministic function called the base intensity and $\nu$ is a positive decreasing weight function.

earthquakes
Hawkes process: definition

The counting process $N(t)$ describes the number of events detected until time $t$.

A family of point processes defined via the intensity function $\lambda(t|\mathcal{F}_t)$

$$\lambda(t|\mathcal{F}_t) = \lim_{\delta \to 0} \mathbb{E} \left[ \frac{N(t + \delta) - N(t)}{\delta} | \mathcal{F}_t \right]$$

$$= \mu + \int_{-\infty}^{t} \phi(t - s) dN(s) = \mu + \sum_{t_i < t} \phi(t - t_i)$$

where

- $\mu$ is a positive constant baseline intensity (Poisson component)
- the kernel $\phi(t)$ is a positive and causal function in $L_1$ (i.e. $\phi(t) = 0$, $\forall t < 0$)
- stability condition $\|\phi\|_1 \equiv \int_{0}^{\infty} \phi(t) dt < 1$
- $n \equiv \|\phi\|_1 = degree \ of \ endogeneity.$

Applied to earthquakes, epidemiology, genomics, crime, finance, etc.
Hawkes processes

**Figure:** Left. Example of a simulated univariate Hawkes process. A blue triangle signals the occurrence of a count. A single exponential kernel was employed with $\mu = 1.2$, $\alpha = 0.5$, $\beta = 0.9$. Right. Branching structure of the Hawkes process (top) and events on the time axis (bottom). This picture corresponds to a branching ratio equal to $n = 0.88$. (from Filimonov and Sornette, PRE 2012).
Multivariate Hawkes process

\[ \lambda_i(t) = \mu_i + \sum_{j=1}^{D} \int_{-\infty}^{t} \phi_{ij}(t-s) dN_j(s) = \mu_i + \sum_{i=1}^{D} \sum_{t_j < t} \phi_{ij}(t-t_j) \]

- \( dN_j(s) = \sum_{t_j < s} \delta(s-t_j) ds \)
- \( \mu_i \) is a positive constant baseline intensity
- the kernels \( \phi_{ij}(t) \) are positive and causal functions in \( L_1 \)

For each component:

\[ \Lambda_i = \mu_i + \sum_{j=1}^{D} \Lambda_j \| \phi_{ij} \| \]

Hence:

- \( \mu_i \) is the immigrant intensity of type \( i \) events.
- \( \frac{\Lambda_j}{\Lambda_i} \| \phi_{ij} \| \) is the fraction of type \( i \) events "triggered" by type \( j \) events.
- \( \| \phi_{ij} \| \) is the average number of type \( i \) event triggered by a type \( j \) event.

The process \( N(t) \) is stationary if the spectral radius of the matrix

\[ \| \Phi \| = \{ \| \phi_{ij} \| \} \]

is strictly smaller than one.
Estimation of Hawkes processes

- **Maximum likelihood estimation**
  Given a functional form of the kernel $\phi$, the log-likelihood reads:

\[
\ln L = -\int_0^T (\lambda_\theta(s))ds + \int_0^T \ln \lambda_\theta(s)dN(s)
\]

- **Non parametric estimation**
  The second order statistic

\[
g(t)dt \equiv \mathbb{E}[dN(t)|dN(0) = 1] - \delta(t) - \mathbb{E}[\lambda(t)] dt
\]

is sufficient, thanks to the following

**Theorem**

*The kernel $\phi$ is the only causal solution of the Wiener-Hopf integral equation*

\[
g(t) = \phi(t) + g \ast \phi(t) \quad \forall t > 0
\]

where $\ast$ indicates convolution.
The challenge: highly dimensional Hawkes processes

- Multidimensional Hawkes processes are able to describe self and cross excitation between signals, but are difficult to calibrate

- Factor Hawkes models (Bormetti et al 2013)

- Parametric multidimensional Hawkes processes for the multiplicity of events (Calcagnile et al 2015)

\[ \lambda_i(t) = \lambda_i^0 + \sum_{j=1}^{M} \sum_{s_j < t} \alpha_{ij} e^{-\beta_{ij}(t-s_j)} \]

A large fraction of instabilities and of market activity (> 99%) is endogenously generated
High-multiplicity is self-exciting
Inferred kernel of the Hawkes process measuring the **endogenous** feedback effect on trading in the EUREX market

- clear evidence of a "market reaction time" at 300μs
- 300μs is the average co-location roundtrip time!
- Interaction below market reaction time: HF strategies by same trader (e.g. order splitting)
- Effect on cancellations! order re-positioning by market makers

M. Rambaldi, E. Bacry, F. Lillo 2017
Variability
WE LIVE IN A DYNAMICAL WORLD . . .

Volatility (standard deviation) of price returns

Tradeoff between:
- More accurate estimate (large windows)
- More timely estimate (small windows)
- Error: $1/\sqrt{T}$
WHICH IS ALSO NOT ALWAYS OBSERVABLE AND NOISY

trades (old 2000 data)

What is the price now?
Correlation among stocks is present
Problem 1: estimating high frequency correlations

Estimating (and forecasting) correlations between asset returns at ultra-high frequency is important but complicated by

- Microstructure noise (e.g. bid-ask bounce)
- Asyncronicity

Bund 10Y - Bobl 5Y - 40 days - 9-11 am

Epps effect (source Bacry)

Many estimators robust to these effects have been proposed: Hayashi and Yoshida (2005), Barndorff-Nielsen, Hansen, Lunde, Shephard (2011), Ait-Sahalia, Fan, and Xiu (2010), Corsi, Peluso and Audrino (2014), and many others.
Consider a multivariate local level model with time-varying (t.v.) parameters:

\[ Y_t = X_t + \epsilon_t, \quad \epsilon_t \sim \text{NID}(0, H_t) \]

\[ X_{t+1} = X_t + \eta_t, \quad \eta_t \sim \text{NID}(0, Q_t) \]

where \( Y_t \), \( X_t \) are \( \mathbb{R}^n \) and \( H_t \) is diagonal. Assuming that \( H_t \) and \( Q_t \) are known, the Kalman filter allows to:

1. Write down the log-likelihood in closed form
2. Provide optimal estimates of the latent process \( X_t \)
3. Easily handle missing observations

Thus, one can think of incorporate microstructure effects into \( \epsilon_t \) and treat asynchronicity as a missing value problem, in a similar fashion to Corsi, Peluso and Audrino (2014) [1].

However, the dynamics of \( H_t \) and \( Q_t \) is unknown. We want to answer the following question:

Is it possible to estimate \( X_t \) and \( \{H_t, Q_t\} \) using the Kalman filter?

\( H_t \) and \( Q_t \) are stochastic but, given observations up to time \( t - 1 \), their value at time \( t \) is completely known.
Recovering the efficient price

Figure: Observed log-price of Citigroup in a 5 minutes time window on 02/01/2014 and its filtered estimate provided by the LLSD
Dynamical Correlation Model

Capture fast, real-time changes of volatilities and correlations: assessing the real-time impact of macro-news announcements

Standard deviations (volatilities)

Cross correlations between stocks
Veracity
How news affect the trading behavior of different categories of investors in a financial market

Research questions

- What is the relative role of endogenous and exogenous factors affecting trading behavior of agents?
- Endogenous factors: price returns and volatility.
- Exogenous factors: number of news and sentiment via semantic analysis of news.
- Is there a difference in the importance of these factors between different categories of investors (e.g. households, companies, governmental, or financial institutions)?
The Finnish database

- Central register of shareholdings for Finnish stocks and financial assets in the Finnish Central Securities Depository.

- **Six main categories**: non-financial corporations, financial and insurance corporations, general governmental organizations, non-profit institutions, households, and foreign organizations.

- Foreign investors can choose to use nominee registration, giving aggregate results. Our focus is mainly on Finnish investors.

- We consider the stock Nokia in the period Jan. 2, 2003 - Dec. 30, 2008 (1,510 trading days).

- The time resolution is one day.

**Table:** Summary of the number of investors (# ids), the number of transactions (N), and the exchanged volume (V). Volume is given in millions of shares.

<table>
<thead>
<tr>
<th>Category</th>
<th># ids</th>
<th>N</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Companies</td>
<td>8,396</td>
<td>1,009,226</td>
<td>4,825</td>
</tr>
<tr>
<td>Financial</td>
<td>392</td>
<td>4,079,174</td>
<td>21,402</td>
</tr>
<tr>
<td>Governamental</td>
<td>124</td>
<td>39,278</td>
<td>1,985</td>
</tr>
<tr>
<td>Non profit</td>
<td>922</td>
<td>21,778</td>
<td>248</td>
</tr>
<tr>
<td>Households</td>
<td>129,952</td>
<td>1,555,096</td>
<td>1,993</td>
</tr>
<tr>
<td>Foreign</td>
<td>1,405</td>
<td>789,552</td>
<td>7,685</td>
</tr>
<tr>
<td>Total</td>
<td>141,190</td>
<td>7,494,104</td>
<td>38,138</td>
</tr>
</tbody>
</table>
The Thomson Reuters database and the sentiment proxy

- Headlines of the NewsScope archive of news released in English by Thomson Reuters.
- We extract all headlines in English language labeled with at least one Nokia Reuters Instrument Code → 11,484 unique headlines.
- We consider only the headlines during European trading hours (from 8.00 am to 4.30 pm UTC time).
- We construct a sentiment proxy using the number of positive and negative words present in each headline. Positive and negative words are detected by using the General Inquirer from the Harvard psychosocial dictionary.

**Figure:** Average daily pattern of the arrival rate of news on the Nokia company. The rate is measured in number of headlines per minute.
Investigated variables

**Investor variables**
- For each day we classify each agent in buyer (B), seller (S), or buyselling (BS) by using the $q(i, t)$ function defined above.
- $N^K_B(t)$, $N^K_S(t)$, and $N^K_{BS}(t)$ are the number of investors of category $K$ classified at day $t$ as buyers, sellers or buysellers, respectively.
- From these variables we obtain
  \[ N^K(t) = N^K_B(t) + N^K_S(t) + N^K_{BS}(t) \]
  \[ \Delta N^K_A(t) = N^K_B(t) - N^K_S(t) \]
  \[ \Delta N^K_R(t) = \frac{N^K_B(t) - N^K_S(t)}{N^K(t)} \]
  → number of investors of category $K$
  → excess of buyers of category $K$
  → relative excess of buyers of category $K$

**Endogenous variables**
- Daily return
- Daily volatility (range)

**Exogenous variables**
- Number $H(t)$ of Nokia headlines
- Absolute and relative sentiment of the news in a given day
  \[ S_A(t) = G(t) - B(t) \]
  \[ S_R(t) = \frac{G(t) - B(t)}{G(t) + B(t)} \]
  where we use the number of positive ($G(t)$) and negative ($B(t)$) words in the headlines.
Figure: From top to bottom the figure shows the time series of the number of Nokia headlines $H(t)$, the daily volatility $Vol(t)$ of Nokia stock, and the time series of $N^K(t)$ for the category of Financial investors and for the category of Households investors.
Figure: From top to bottom the figure shows the time series of the relative sentiment indicator $S_R(t)$, the daily return $Ret(t)$ of Nokia stock, and the time series of $\Delta N^K_R(t)$ for the category of Financial investors and of Households investors.
Number of news is correlated with volatility, $\text{Corr}[H, Vol] = 0.501$.

We fit

$$\hat{N}^K(t) = \alpha_H \hat{H}(t) + \alpha_{Vol} \hat{Vol}(t) + \epsilon(t)$$

where $\hat{x}$ is the standardized versions with zero mean and unit variance of $x$.

Table: Summary of the results of the linear regression of the number $N^K$ of trading investors versus the news intensity signal $H$ and the volatility proxy $Vol$. The number in parentheses are the 5%-95% confidence intervals under Gaussian hypothesis and by using bootstrap analysis. The last two columns show the results of the partial correlation analysis.

| Investor category | $\alpha_H$ | $\alpha_{Vol}$ | % variance of residual of $N^K$ | $\rho(N^K, H|Vol)$ | $\rho(N^K, Vol|H)$ |
|-------------------|------------|----------------|--------------------------------|-------------------|-------------------|
| Companies         | 0.271 (0.229,0.313) | 0.517 (0.475,0.559) | 51.8 % | 0.309 | 0.534 |
| bootstrap         | (0.205,0.335)         | (0.437,0.597)         |      |      |      |
| Financial         | 0.195 (0.149,0.242) | 0.479 (0.433,0.526) | 63.8 % | 0.207 | 0.461 |
| bootstrap         | (0.125,0.264)         | (0.407,0.558)         |      |      |      |
| Governmental      | 0.238 (0.183,0.292) | 0.192 (0.138,0.246) | 86.0 % | 0.215 | 0.180 |
| bootstrap         | (0.164,0.303)         | (0.119,0.262)         |      |      |      |
| Non profit        | 0.319 (0.269,0.369) | 0.270 (0.220,0.320) | 73.9 % | 0.305 | 0.264 |
| bootstrap         | (0.249,0.394)         | (0.199,0.344)         |      |      |      |
| Households        | 0.226 (0.188,0.263) | 0.627 (0.589,0.664) | 41.4 % | 0.289 | 0.651 |
| bootstrap         | (0.165,0.285)         | (0.554,0.697)         |      |      |      |
| Foreign org.      | 0.158 (0.109,0.207) | 0.442 (0.393,0.492) | 70.9 % | 0.160 | 0.416 |
| bootstrap         | (0.094,0.224)         | (0.374,0.517)         |      |      |      |
Regression results and partial correlation analysis: sentiment and returns

- Sentiment is correlated with returns, $\text{Corr}[S_A, Ret] = 0.155$ and $\text{Corr}[S_R, Ret] = 0.118$ (statistically significant).

- We fit

$$\hat{\Delta N^K_R}(t) = \alpha_{S_R} \hat{S}_R(t) + \alpha_{Ret} \hat{Ret}(t) + \epsilon(t)$$

where $\hat{x}$ is the standardized versions with zero mean and unit variance of $x$.

**Table:** Summary of the results of the linear regression of the relative difference $\Delta N^K_R$ between buying and selling investors versus the relative sentiment indicator $S_R$ and the stock return $Ret$. The number in parentheses are the 5%-95% confidence intervals under Gaussian hypothesis and by using bootstrap analysis. The last two columns show the results of the partial correlation analysis.

| Investor category | $\alpha_{S_R}$ | $\alpha_{Ret}$ | % variance of residual of $\Delta N^K_R$ | $\rho(N^K_R, S_R | Ret)$ | $\rho(N^K_R, Ret | S_R)$ |
|-------------------|----------------|---------------|---------------------------------|------------------------|------------------------|
| Companies         | 0.055 (0.014, 0.095) | -0.610 (-0.650, -0.569) | 63.3 % | 0.0685 | -0.6056 |
| Financial         | 0.018 (-0.025, 0.062) | -0.520 (-0.564, -0.477) | 73.1 % | 0.0212 | -0.5170 |
| Governmental      | 0.021 (-0.029, 0.071) | -0.179 (-0.230, -0.129) | 96.8 % | 0.0215 | -0.1782 |
| Non profit        | 0.025 (-0.025, 0.075) | -0.175 (-0.225, -0.125) | 96.9 % | 0.0256 | -0.1738 |
| Households        | 0.068 (0.026, 0.110) | -0.565 (-0.608, -0.523) | 68.4 % | 0.0811 | -0.5615 |
| Foreign org.      | 0.030 (-0.017, 0.077) | -0.400 (-0.446, -0.353) | 84.2 % | 0.0323 | -0.3970 |
The activity of governmental and non profit organizations is very poorly explained by return and news sentiment. Of the two factors, return plays clearly a major role.

Households and companies are those for which sentiment and returns have the best explanatory power of their trading action. Return is clearly more important, but sentiment has also some explanatory power, especially when one consider the relative imbalance between buyers and sellers.

For financial and foreign organizations the variance explained by the regressions is somewhat intermediate between the two pairs of categories above, but in general returns have a much higher explanatory power and sentiment plays a negligible role.

For companies, financial institutions, households and foreign organizations $\alpha_{Ret} < 0$ and large indicating that market polarization of trading actions is strongly anticorrelated with the Nokia return. The majority of single investors of these categories are therefore buying when the Nokia price goes down and selling when the price goes up.

On a daily time scale, news move investors to trade.

Most of the times the sentiment indicator is not significantly correlated with the imbalance between buyers and sellers.
Huge quantity of texts (news, blogs, social networks)

Extracting in automatic way
- features
- meaning
- sentiment

Natural Language Processing

Sentiment extraction

Are news predicting prices?

IBM Is Finding Revenue Opportunity in the Cloud

Anne Shields

October 20, 2016

Once a Tech Darling, IBM Has 18th Quarter of No Revenue Growth

PART 2 OF 5 Strategic Imperatives contribution rose to 40% of IBM's overall revenues in the previous part of this series, we looked at market expectations for International Business Machines' (IBM) fiscal 3Q16 results. Fiscal 3Q16 marked the 18th consecutive quarter of no revenue growth. However, despite no growth in overall revenues, IBM's Strategic Imperatives continued to post growth in 3Q16, as expected. In fiscal 3Q16, IBM's Strategic Imperatives, which refers to the company's investments in cloud, analytics, mobile, social, and security technologies, rose 16.0% to $8.0 billion. As you can see in the chart below, in the last
NEWS AND CLICKS TO PREDICT RETURNS

- Sentiment of public news have a very weak predictive power on price returns.
- Important news (e.g. earning announcements) have a stronger impact; however only when the content of the news was unexpected (see the Efficient Market Hypothesis).

- Is it possible to measure the “surprise” of a news by monitoring the number of clicks it receives in an internet news portal?
- Is the sentiment weighted by the surprise more predictive?

Market data: 100 high capitalization stocks traded in US equity markets, in 2012-2013.
- $V$, the traded volume in that interval of time,
- $R$, the logarithmic price return in the time scale,
- $\sigma$, the return absolute value, a simple proxy for the stock volatility.

News and click data: news published on Yahoo! Finance together with the time series of the aggregated clicks made by the users browsing each page.
- $C$, the time series of the total number of clicks in a time window,
- $S$, the sum of the sentiment of all news related to each company,
- **WS**, the sum of the sentiment of all news weighted by the number of clicks.

Ranco et al. PlosONE 2016
NOT ALL NEWS TRIGGER THE SAME ATTENTION

Complementary of the cumulative distribution function of the number of clicks a news receives for the ten assets with the largest number of news and the aggregate portfolio of 100 stocks.
Time evolution of the cumulative number of clicks per news in a time interval of five hours after the publication. We normalize the cumulated amount by a constant which corresponds to the total number of clicks received by a single news during the first week after publication. The news are grouped in deciles according to the total number of clicks they have received until October 2013 and the curves represent average values. Inset: estimated values and standard errors of the attention time scale obtained by an exponential fit of the decile curves.
Granger Causality tests at hourly scale between different variables. The white cells correspond to tests for which we do not reject the null hypothesis of no Granger causality at 5% significance level. A black cell corresponds to a statistically significant Granger causality.

**Weighting the sentiment of a news with the number of clicks it receives (i.e. the attention of users) highly increases its capability of forecasting price returns**
$FAKE$: EVIDENCE OF SPAM AND BOT ACTIVITY IN STOCK MICROBLOGS ON TWITTER

Figure 1: Sample tweet with the $AAPL$ hashtag.

(c) $NFLX$ (Netflix, Inc.).
(d) $TSLA$ (Tesla, Inc.).

Large fraction of tweet peaks on low cap stocks likely due to bots

Cresci et al. 2018