Detecting intraday financial market states using temporal clustering

Dieter Hendricks Tim Gebbie Diane Wilcox

QuERILab
Quantifying Emergence, Risk and Information
in financial markets

School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg, South Africa

AASS Workshop, IMPA, 2016

Outline

1. Market microstructure and state representation
2. Identifying temporal system states
   - Unsupervised clustering approach
   - Giada-Marsili proposition with Noh ansatz
   - High-speed cluster configuration identification
   - Financial market intraday times as objects
3. Data and Results
   - Data
   - Temporal states, State Signature Vectors, Cluster size power-law fit, Estimated states, Transition probabilities
4. Conclusion

Market microstructure and state representation

- Financial markets as **complex adaptive systems** (Wilcox & Gebbie, 2015), prices and volumes as **measurable quantities**
- Market microstructure studies the **system evolution**, **behaviour** and consequences for **price formation** at the **lowest scale** (tick-level)
- Motivated by need for **efficient state representation** for automated trading agents in high-frequency markets to enable effective learning
- Agents faced with **high-volume**, **asynchronous** stream of financial market **data** from a real-time datafeed
- Can we find **persistent structure** in this streaming data, such that **meaningful learning** can take place for adaptive trading algorithms?

Outline

1. Market microstructure and state representation
2. Identifying temporal system states
   - Unsupervised clustering approach
   - Giada-Marsili proposition with Noh ansatz
   - High-speed cluster configuration identification
   - Financial market intraday times as objects
3. Data and Results
   - Data
   - Temporal states, State Signature Vectors, Cluster size power-law fit, Estimated states, Transition probabilities
4. Conclusion

Identifying temporal system states

Unsupervised clustering approach

- Clustering involves **grouping objects** according to **metadata** describing objects or their associations
- For unsupervised clustering, apply **super-paramagnetic ordering of q-state Potts model** for cluster identification (Blatt et al., 1996)
- Cost function is a Hamiltonian whose **low energy state** corresponds to a **cluster configuration** most compatible with the sample

$$H_g = -\sum_{s_i, s_j \in S} J_{ij}\delta(s_i, s_j) - \frac{1}{\beta}\sum_i h_i^M s_i$$

  where spins s_i can take on q-states and the short-range ferromagnetic interaction between spins is given by J_{ij}
- To parameterise the model, make Noh (2000) ansatz and use this to develop the **MLE approach** of Giada & Marsili (2001)

Outline

Identifying temporal system states

Giada-Marsili proposition with Noh ansatz

- Objects belonging to the same cluster share a common component,

$$\bar{x}_i = g_{s_i}\bar{\eta}_{s_i} + \sqrt{1-g_{s_i}^2}\,\bar{\epsilon}_i. \tag{1}$$

- If we take this as a **statistical hypothesis**, and assume $\bar{\eta}_{s_i}$ and $\bar{\epsilon}_i$ are Gaussian, this leads to the following probability density,

$$P\left(\{\bar{x}_i\}|\mathcal{G},\mathcal{S}\right) = \prod_{d=1}^{D}\left\langle \prod_{i=1}^{N}\delta\left(x_i(t) - g_{s_i}\bar{\eta}_{s_i} + \sqrt{1-g_{s_i}^2}\,\bar{\epsilon}_i\right)\right\rangle \tag{2}$$

  and the **maximum likelihood** for structure $\mathcal{S}$ can be written as (Giada & Marsili, 2001),

$$\mathcal{L}_c(\mathcal{S}) = \frac{1}{2}\sum_{s:n_s>1}\left(\log\frac{n_s}{c_s} + (n_s-1)\log\frac{n_s^2-n_s}{n_s^2-c_s}\right). \tag{3}$$

- Full derivation in Hendricks, Gebbie & Wilcox (2016b)

Outline

1. Market microstructure and state representation
2. Identifying temporal system states
   - Unsupervised clustering approach
   - Giada-Marsili proposition with Noh ansatz
   - High-speed cluster configuration identification
   - Financial market intraday times as objects
3. Data and Results
   - Data
   - Temporal states, State Signature Vectors, Cluster size power-law fit, Estimated states, Transition probabilities
4. Conclusion

Identifying temporal system states

High-speed cluster configuration identification

- Likelihood function in Equation 3 used as **objective function** in metaheuristic optimisation routine (genetic algorithm)
- Systematically **evaluate** candidate **configurations** ($\mathcal{S}$), converging towards **best approximation** of data descriptor
- Hendricks, Gebbie & Wilcox (2016a) demonstrate a high-speed **Parallel Genetic Algorithm** implementation in **CUDA**, using the SPMD architecture to enumerate the GPU thread hierarchy with population members for **concurrent application of genetic operators**
- A **low-cost**, **scalable**, **high-speed** implementation for **unsupervised cluster detection**

Outline

Identifying temporal system states

Financial market intraday times as objects

- **Generic** data generative model **ansatz** conducive to any problem where **Gaussian** object and cluster innovations are reasonable
- Marsili (2002) grouped **days** according to **closing price** performance to identify temporal states
- Propose that **intraday temporal regimes** of financial markets are characterised by feature performance of stocks
- Using **trade price**, **trade volume**, **spread** and **volume imbalance** features of TOP40 stocks on the JSE, we find clusters of **60-min, 30-min, 15-min and 5-min periods**
- Determine whether clustering at varying **calendar time scales** reveals interesting **hierarchy** of system behaviour
- Reduces significant amount of high-frequency information into a **tractable representation** for intraday learning

Identifying temporal system states

State Signature Vectors

- Temporal clustering reveals *ex-ante* grouping of periods
- How do we incorporate this into an **online learning algorithm**? i.e. determine the state we are **currently** in
- Analyse **average feature performance** of stocks within identified temporal clusters and extract *State Signature Vector* (SSV) as state descriptor
- Online feature vectors are **easy to compute** from streaming market datafeeds in financial markets, conducive to near-real-time detection
- Need to determine which identified states are **significant**, i.e. likely to persist such that meaningful learning is possible
- Use best **power law fit** to cluster size to identify likely persistent states
- Use **Euclidean distance** of (current) online feature vector to set of SSVs for state index assignment

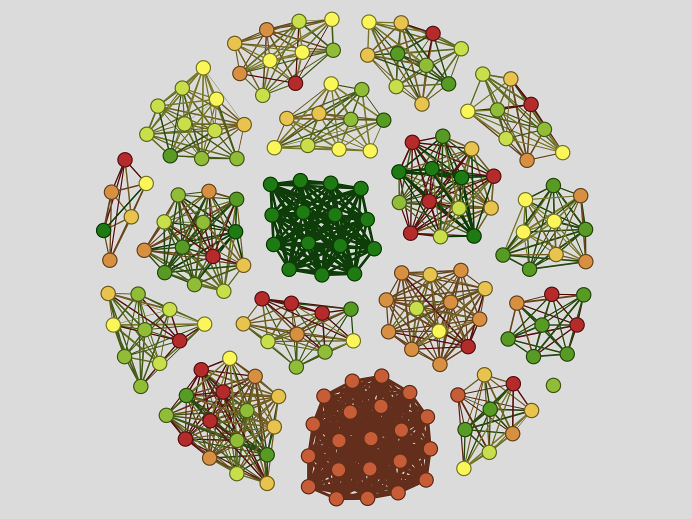
Outline

- Tick-level **trades** and top-of-book **quotes** for **42 stocks** on the **Johannesburg Stock Exchange** (JSE) from **1 November 2012 to 30 November 2012**
- Raw data sourced from Thomson Reuters Tick History
- Stored in **MongoDB noSQL database** on QuERILab server and integrated into MATLAB development environment

Outline

Identifying temporal system states

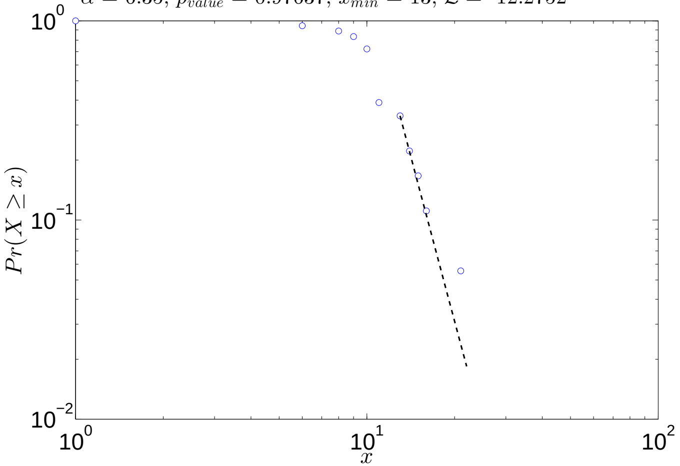
60-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012, GREEN=morning, YELLOW=lunch, RED=afternoon



Power law fit for cluster size

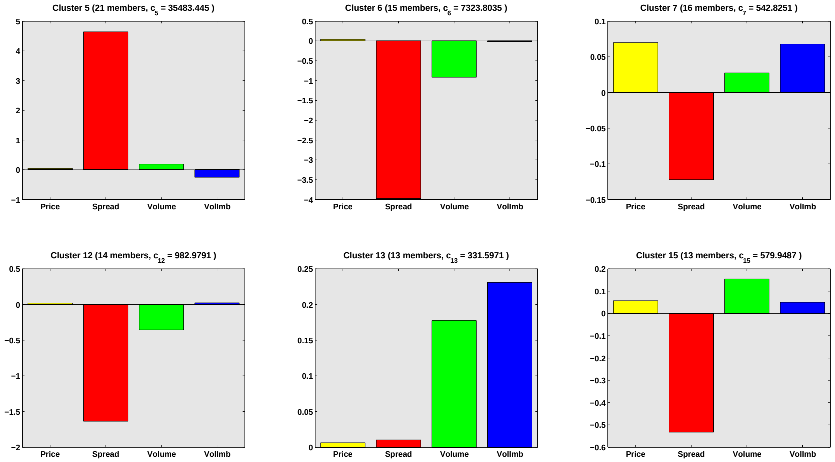
60-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012

Test for **cluster size** power law fit: *60-min periods*
Power law distributional form ($p(x) \sim x^{-\alpha}$) vs empirical data
$\alpha = 6.35$, $p_{value} = 0.97637$, $x_{min} = 13$, $\mathcal{L} =$ -12.2752



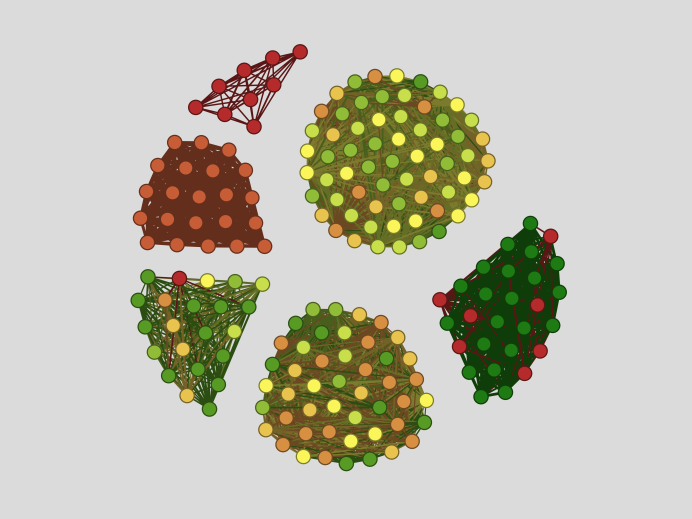
State Signature Vectors

60-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012



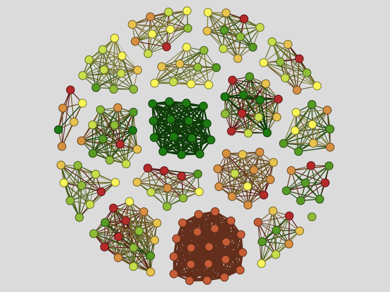
Estimating temporal states using feature vectors

60-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012

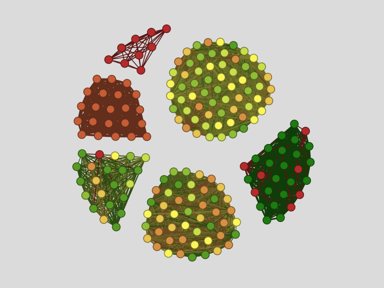


Actual vs Estimated temporal states

60-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012



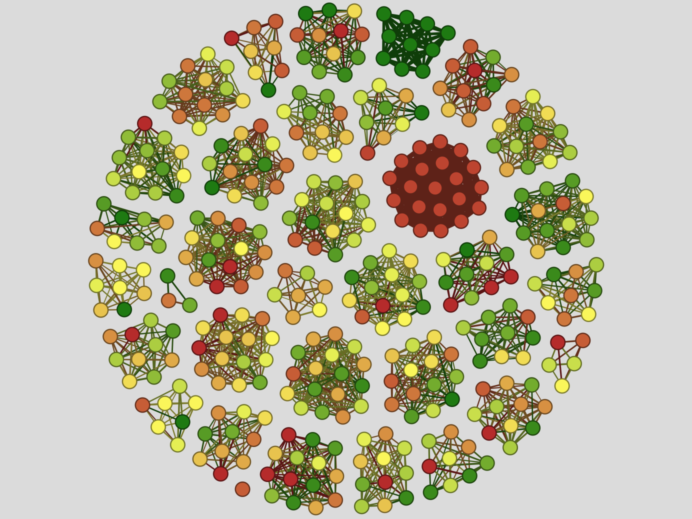
(a) Actual 60-min states



(b) Estimated 60-min states

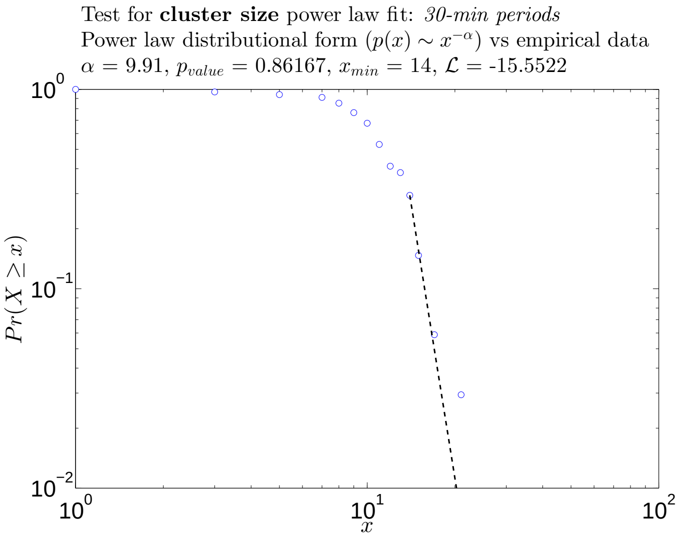
Identifying temporal system states

30-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012, GREEN=morning, YELLOW=lunch, RED=afternoon



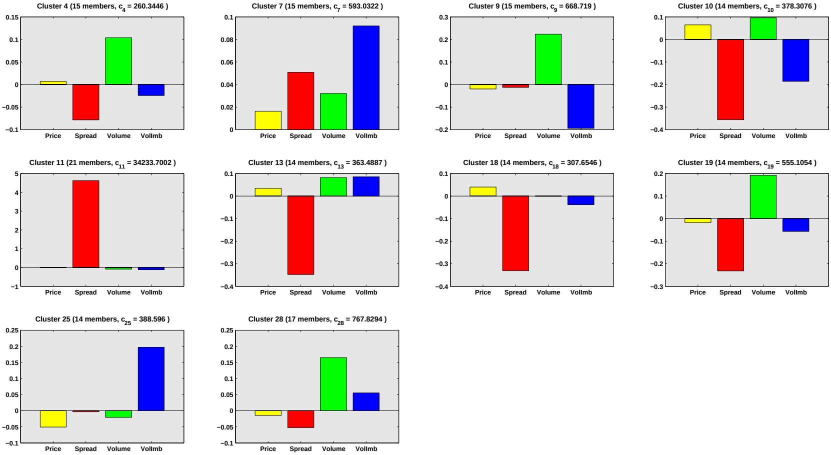
Power law fit for cluster size

30-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012



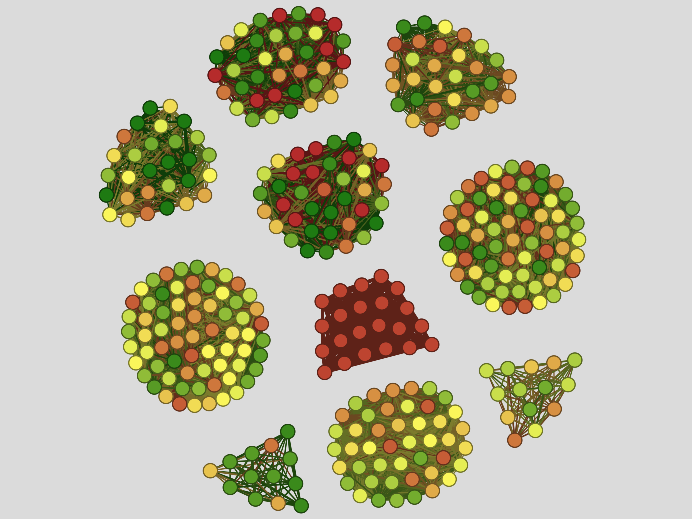
State Signature Vectors

30-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012



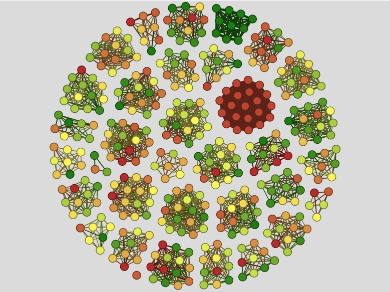
Estimating temporal states using feature vectors

30-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012

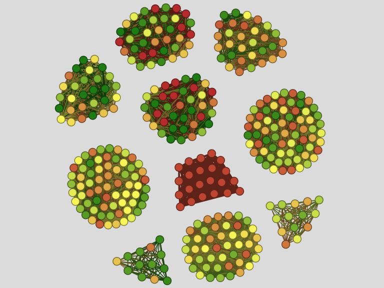


Actual vs Estimated temporal states

30-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012



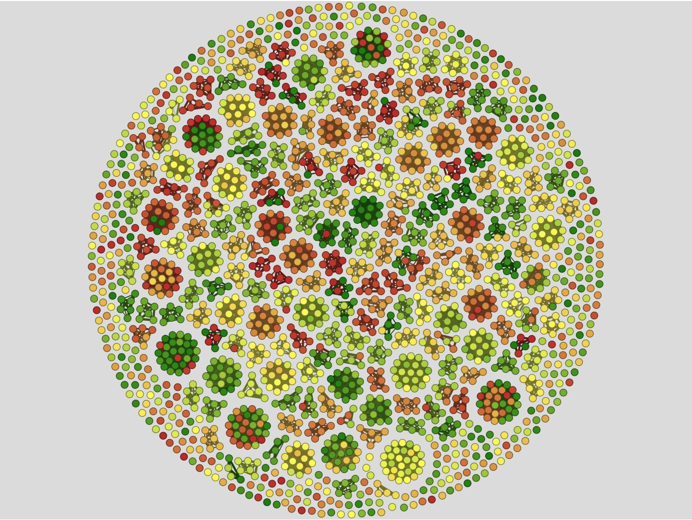
(a) Actual 30-min states



(b) Estimated 30-min states

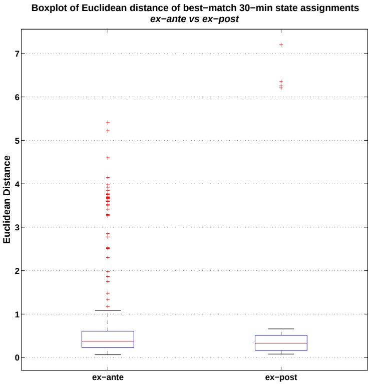
Identifying temporal system states

5-min clusters (TOP40 stocks; price, volume, spread, volume imbalance), 01-Nov-2012 to 30-Nov-2012, GREEN=morning, YELLOW=lunch, RED=afternoon

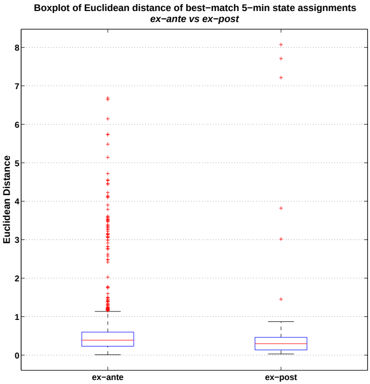


Stability of the online state assignment out-of-sample

ex-ante (01-Nov-2012 to 30-Nov-2012, same period used for SSV estimation) vs *ex-post* (03-Dec-2012 to 07-Dec-2012, one week after SSV estimation window)



(a) 30-min states



(b) 5-min states

Conclusion

- Provided a scheme for **unsupervised** determination of intraday financial market **states** and **online** state **detection** using **SSVs**
- Tractable **state representation** for learning agents in high-frequency trading domain
- Non-trivial **hierarchy** of clustering **behaviour** at varying time-scales reveals **scale-specific information** for learning
- One application is the online enumeration and refinement of an **empirical transition probability matrix**
- States appear to persist **long enough** to be exploited - needs to be verified
- Will consider clustering in **machine time** in further work

For Further Reading I

- D. Hendricks, D. Wilcox, T. Gebbie. *Detecting temporal financial market states using clustering*. Quantitative Finance (accepted, to appear 2016). (Pre-print: http://arxiv.org/abs/1508.04900)
- D. Hendricks, D. Wilcox, T. Gebbie. *High-speed detection of emergent market clustering via an unsupervised parallel genetic algorithm*. South African Journal of Science, vol. 112, no. 1/2, 2016.
- L. Giada, M. Marsili. *Data clustering and noise undressing of correlation matrices*. Phys. Rev. E, vol. 63, no. 061101, 2001.
- M. Marsili. *Dissecting financial markets: sectors and states*. Quantitative Finance, vol. 2, no. 4, pp. 297-302, 2002.
- M. Blatt, S. Wiseman, E. Domany. *Clustering data through an analogy to the Potts model*. Advances in Neural Information Processing Systems, pp. 416-422, 1996.
- D. Wilcox, T. Gebbie. *Hierarchical causality in financial economics*. Working paper, QuERILab, 2015 (Available at SSRN: http://ssrn.com/abstract=2544327)