

Updates Since Last Week

Done Last Week

- Adding some metrics for number of clusters
- Adding code for co-occurrence network of graph
- Adjusting graph visualizations

Next Week Plans

- Start working on Kalman like approach to exploit the graph structure
- try DDTW by considering time intervals relevant to a news article

Enhancing Equity Predictions Using Informational Signals

Group Members: Bella Macaluso - Elizabeth Yang - Sourav Vemulapalli - Aditiya Palliyil - Joseph Jabbour

Github repo:: <https://github.com/bour278/info-signal-analysis>

Table of Contents



1- General Overview

2- Data Sources

3- Methodology

4- Limitations

General Overview

-  **Target:** Enhance equity predictions using informational signals
-  **Methods/Tools:** - Derivative Dynamic Time Warping (DDTW) - Louvain/Leiden Community Clustering - Kalman Filtering - Markov Random Fields

General Overview



DATA COLLECTION

- Kaggle
- Github
- Web Scraping

EDA

- Correlation between news and prices jump
- Optimal number of clusters

MODELS

- DDTW distance graph
- Co-occurrence matrix
- Kalmann Filtering

VISUALIZATION

- Network Clusters
- Time-series Clusters
- Prediction vs Enhanced Prediction using confidence bands

METRICS

- Cut metric / modularity for clustering
- Inter/Intra cluster variance
- MSE for enhanced predictions

Data Sources

- **Kaggle:** [Daily OHLC data for US-based equities](#)

Date	Open	High	Low	Close	Volume	OpenInt
1984-09-07	0.42388	0.42902	0.41874	0.42388	23220030	0
1984-09-10	0.42388	0.42516	0.41366	0.42134	18022532	0

Data Sources

- **Github:** [Reuters Financial Dataset](#)

```
-- Samsung aims to double its smartphone sales in Africa in 2014
--
-- Wed Nov 13, 2013 2:29am EST
-- http://www.reuters.com/article/2013/11/13/us-africa-samsung-idUSBRE9AC08620131113
```

```
CAPE TOWN (Reuters) - Samsung Electronics expects to supply half of the smartphones sold in Africa this year and aims to double these sales on the continent in 2014, an executive said.
```


Data Sources

- **Scraping:** [New York Times News Archive](#)

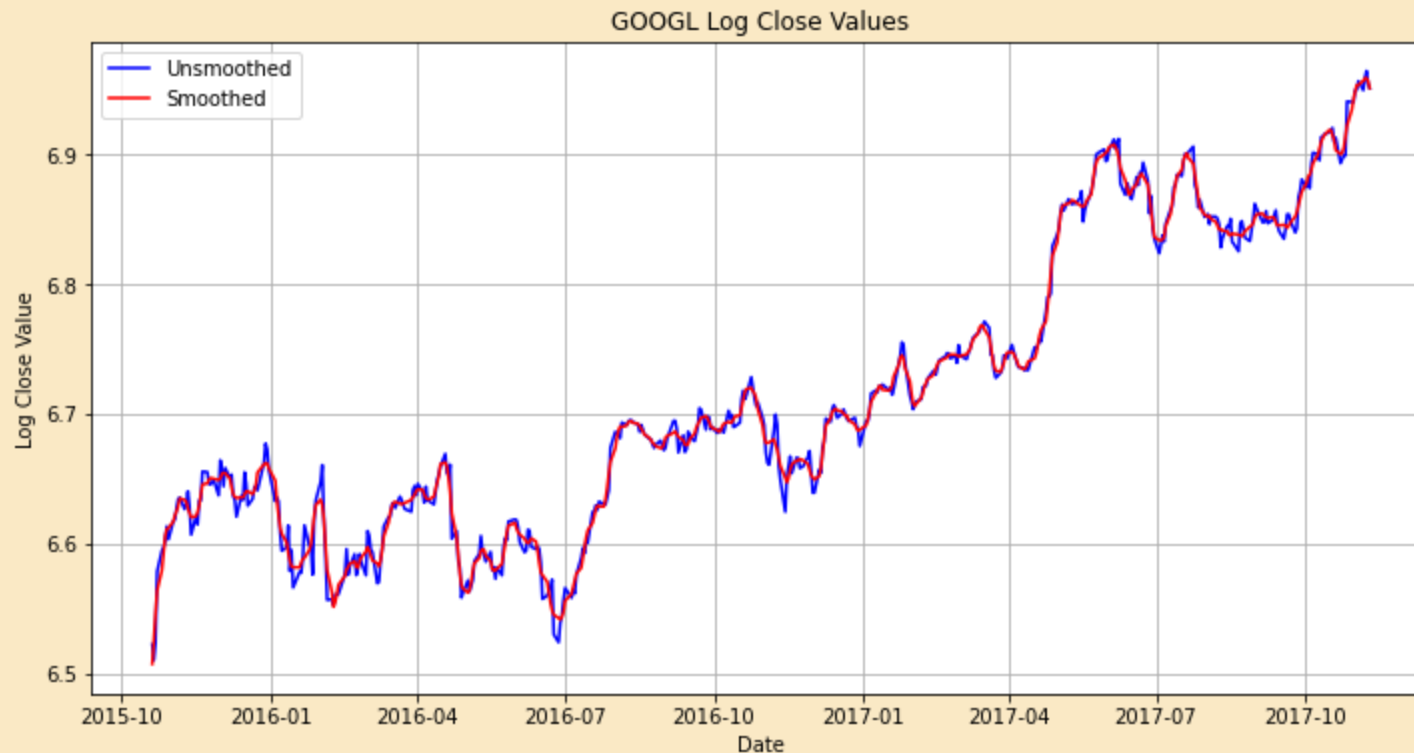
Chadwick Boseman Played Black Icons, Found Fame With 'Black Panther'
11:20 PM ET

Japan
Abe Will Resign as Japan's Prime Minister, Citing His Health
10:17 PM ET

Politics
Thousands March on National Mall, Continuing Racial-Justice Push
10:11 PM ET

Methodology - Pre-Processing

- **Savitzky-Golay Filtering:** removing noise from historical time series data using polynomial interpolation at fixed-length window



Methodology - DDTW Clustering

- **DDTW:** algorithm finding shortest path distance between 2 time series using dynamic programming approach
- **Graph Representation** Adjacency matrix is built from pairwise DDTW distances between each pairs of equities

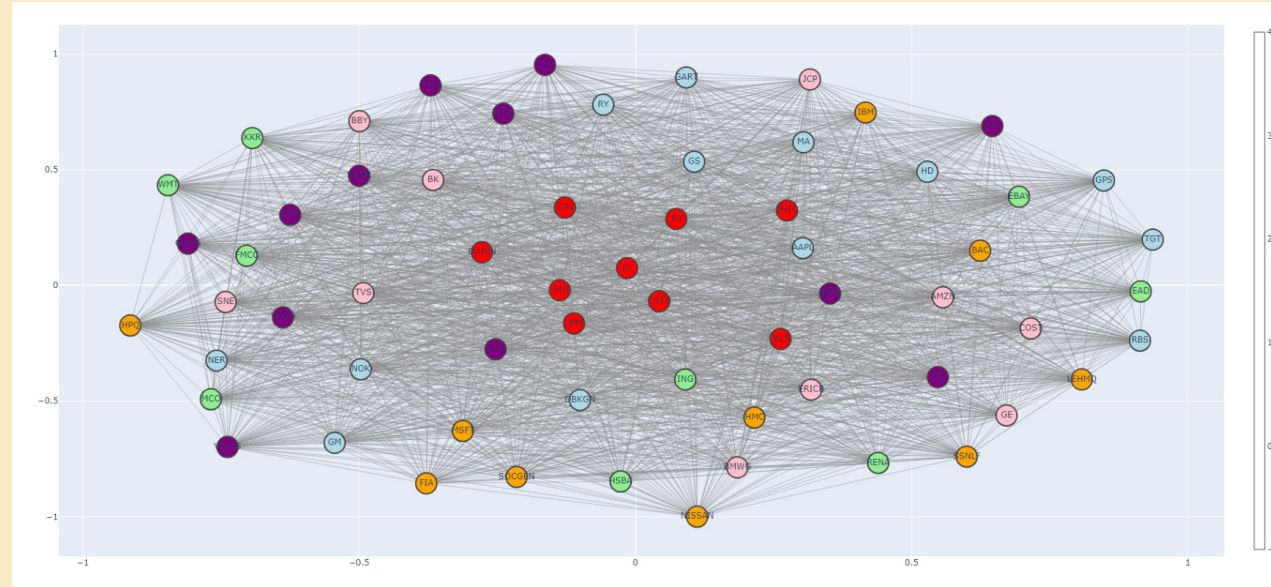
Methodology - DDTW Metrics

- **k-means optimal number of clusters:** For this case we used the *silhouette score method* to computer the optimal number k of clusters. The best k was achieved at $k=2$.
- **inter-variance of numner of the graph** Metric to determine how efficient the clustering method computed by $\frac{\sum_i^K n_i ||c_i - \bar{x}||^2}{K}$ where c_i represent the centroid of the i^{th} cluster and \bar{x} is the global mean of the graph.

Methodology - News Co-occurrence

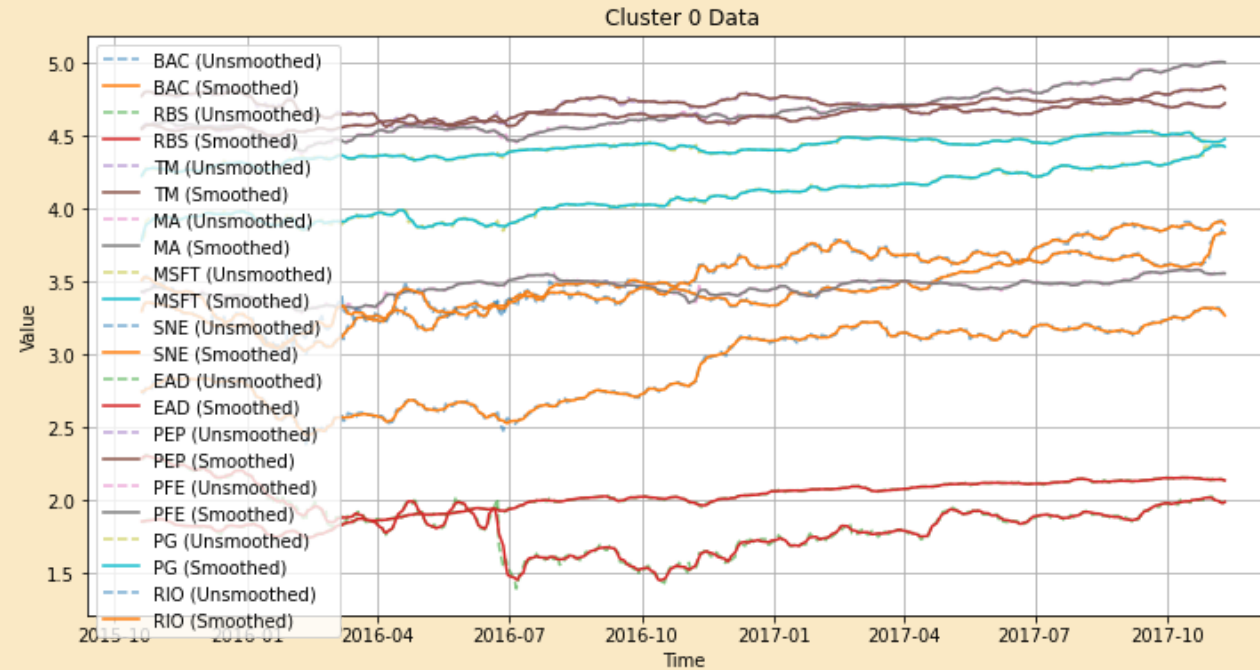
- **News co-occurrence matrix:** Matrix A where $A_{i,j}$ corresponds to the number of news articles where stock i appeared with stock j .
- **Louvain Clustering** Community detection algorithm that helps retrieve clusters in a graph and does not require setting the optimal number of clusters before running the algorithm.

Results - Log Close Graph Cluster

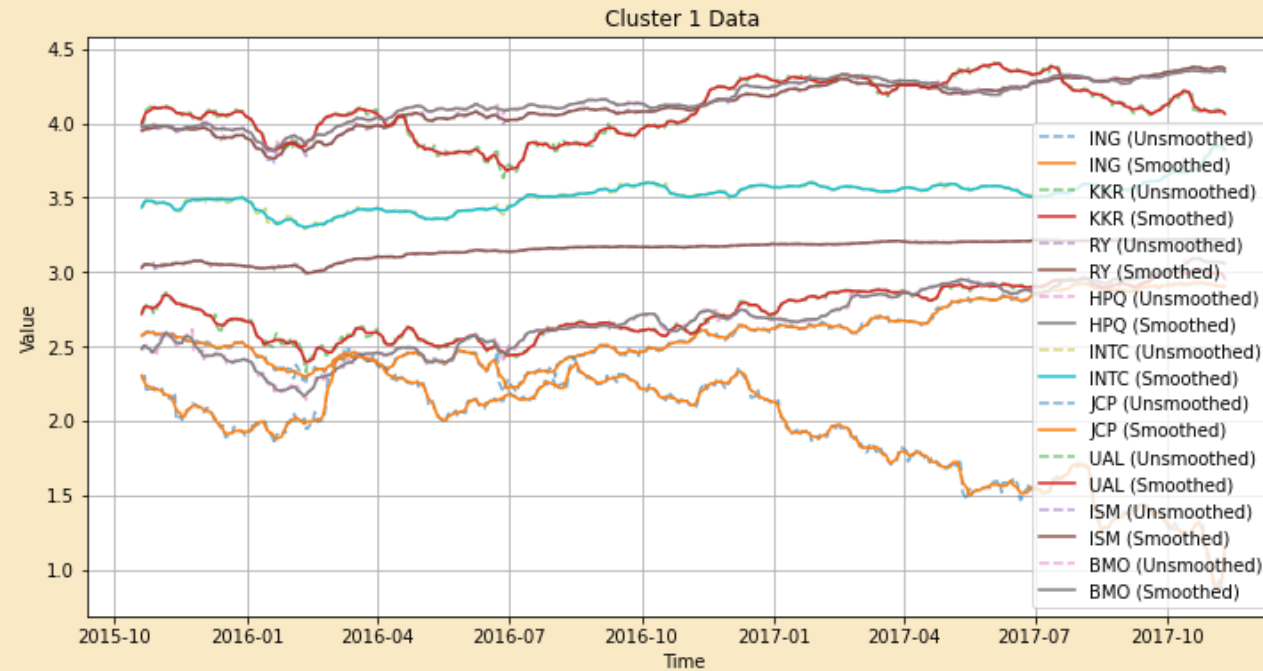


[rendered html fo the graph](#)

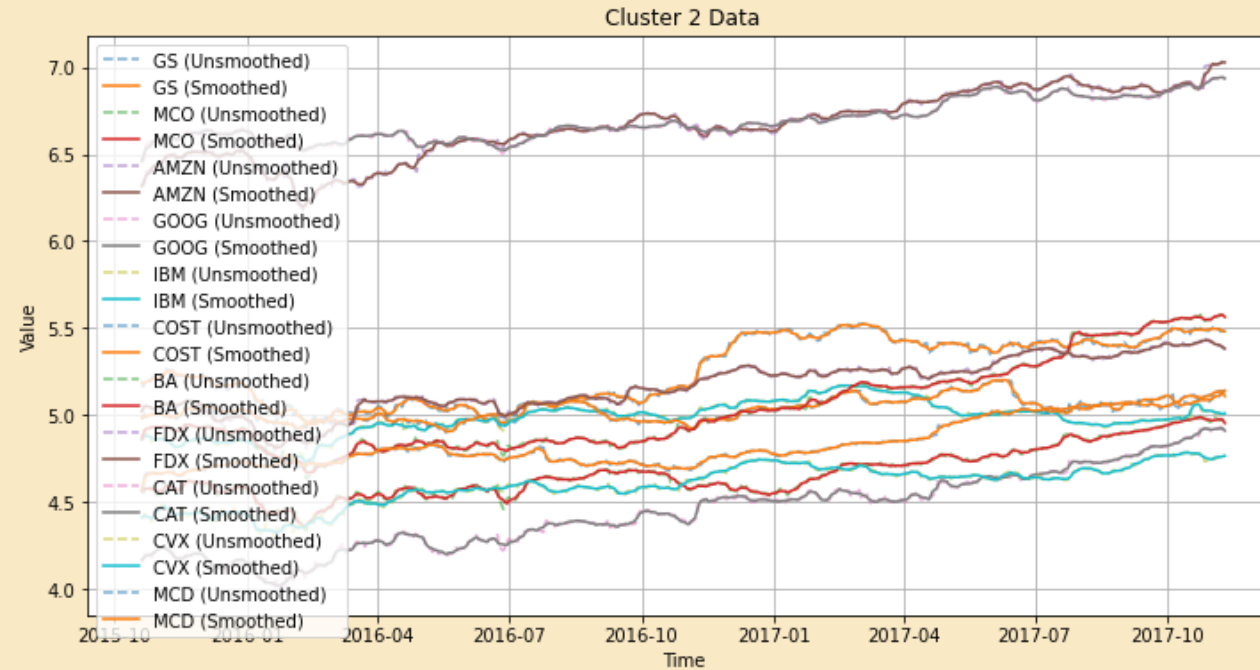
Results - Time Series Cluster 0



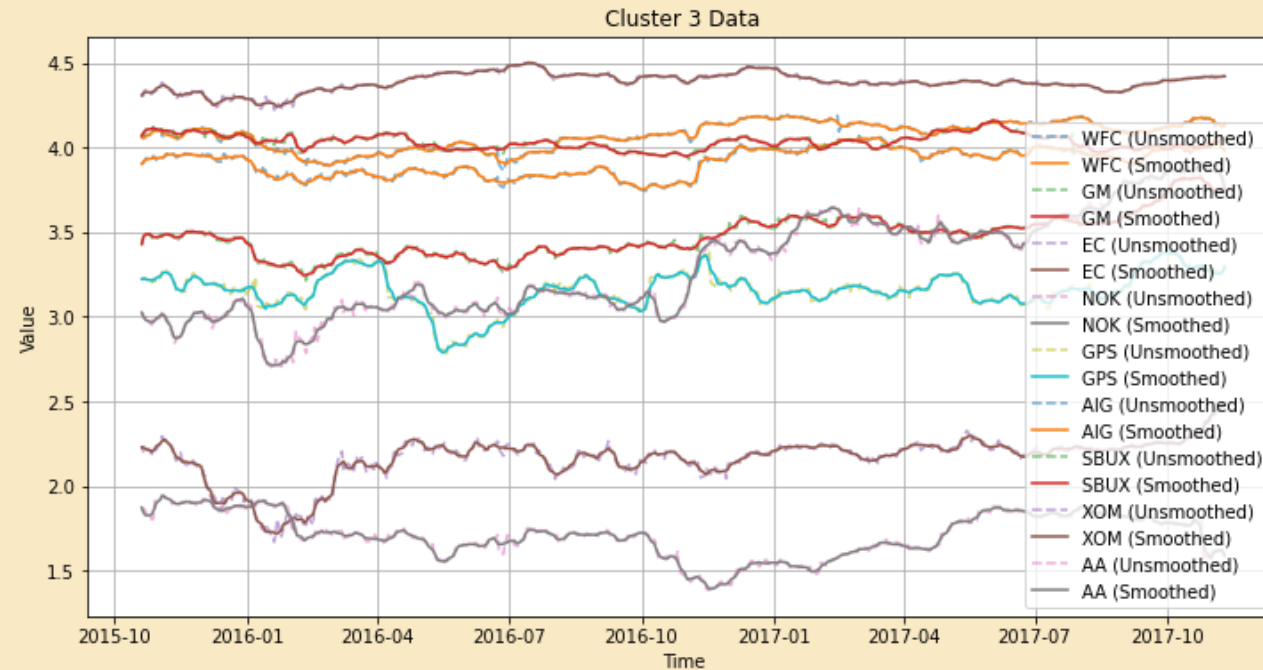
Results - Time Series Cluster 1



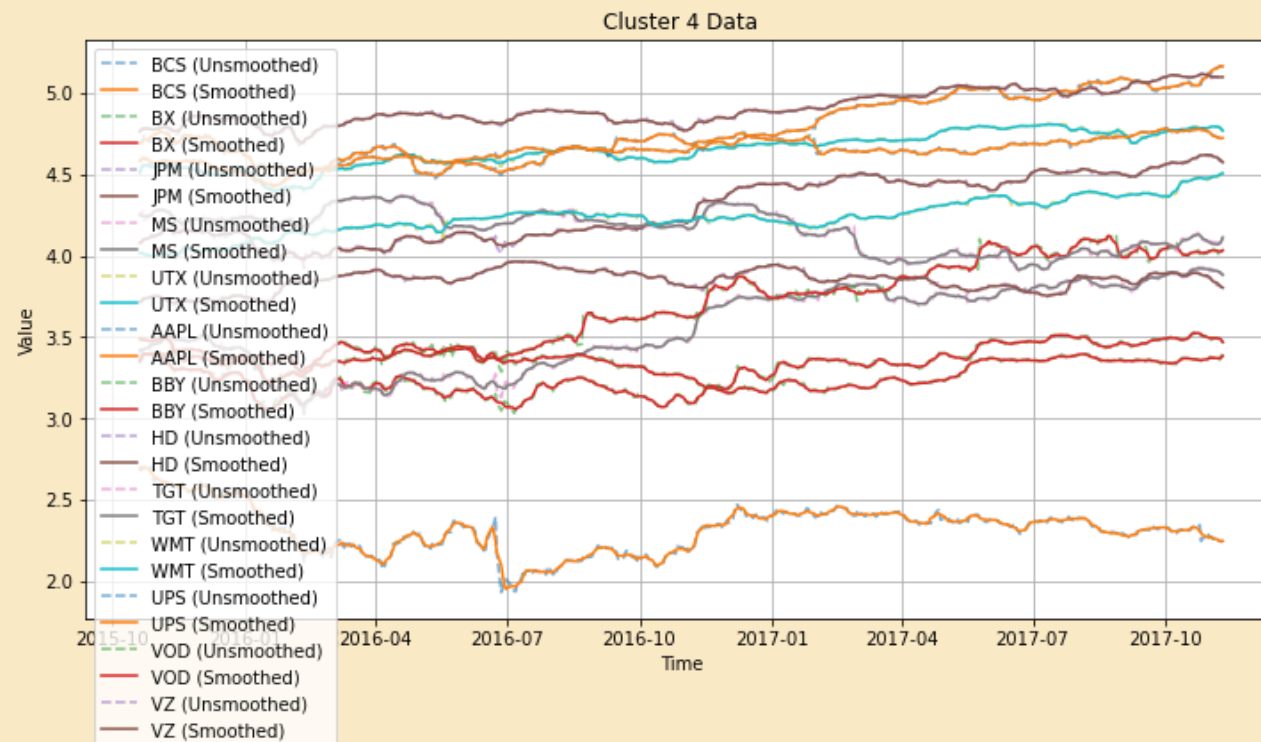
Results - Time Series Cluster 2



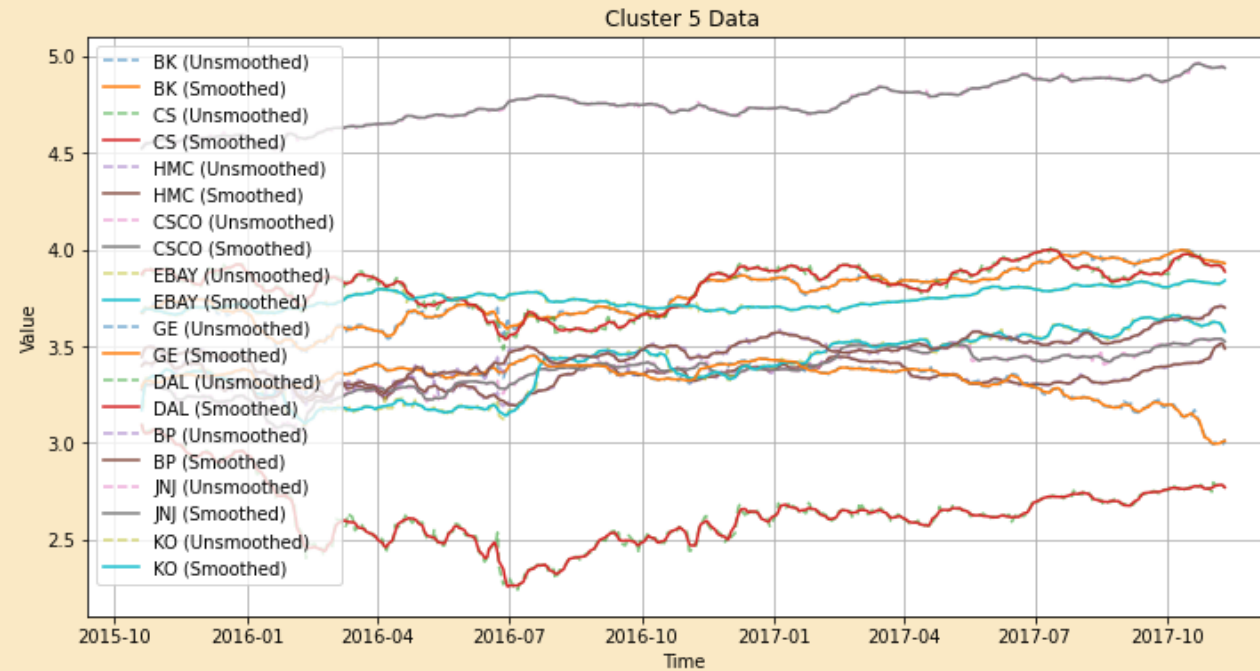
Results - Time Series Cluster 3



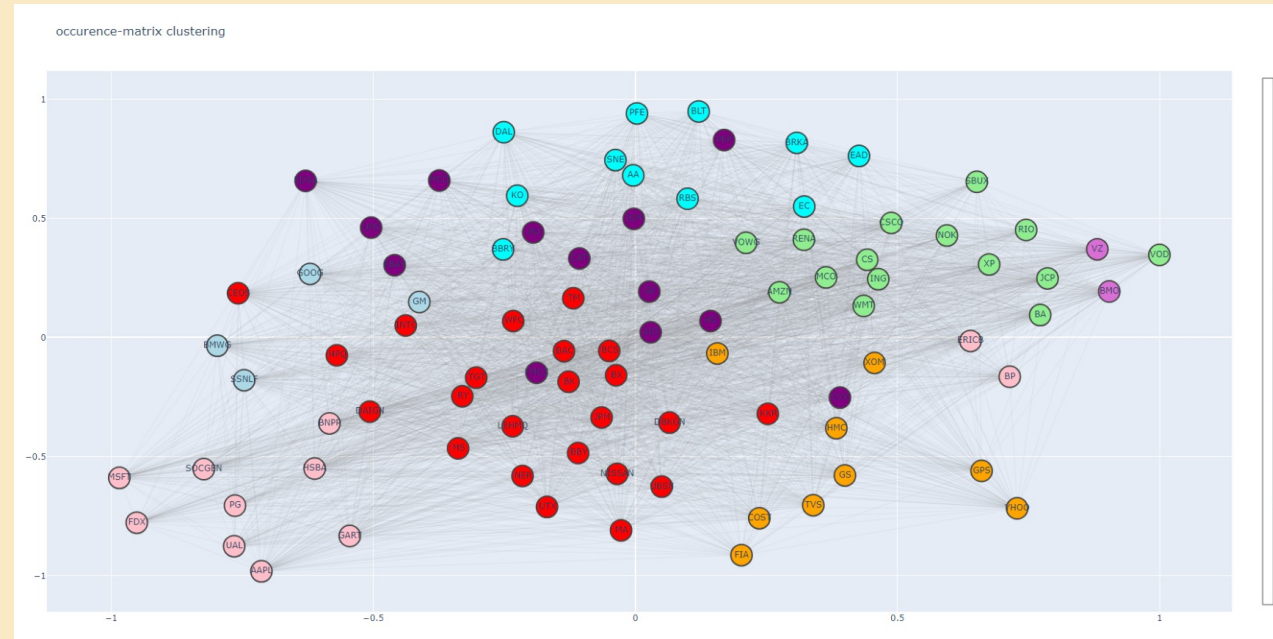
Results - Time Series Cluster 4



Results - Time Series Cluster 5



Results - Co-occurrence Network



rendered html for the graph

Limitations (More to be found 🕒)

- Limited tick data
- Computationally expensive to build graph for long time series