THE IMPACT OF FLASH CRASHES ON ASSET CLUSTERS

Arthur Pollet
{arthur.pollet}@epfl.ch

Elias Naha
elias.naha@epfl.ch

ABSTRACT

Flash crashes, marked by rapid price swings, disrup markets and are often driven by algorithmic trading. The May 6, 2010 Flash Crash, which briefly wiped out $1 trillion in market value, highlights their significant impact. This study aims to examine how these events affect asset correlations and clustering, before, during and after the crashes. Using high-frequency trading data from 2010 for 29 assets, a flash crash detection method was constructed. Using the Hayashi-Yoshida estimator along with eigenvalue clipping, correlation matrices where constructed for the identified flash crashes. A Louvain clustering method was then implemented to identify asset clusters before, during and after the crashes. These methods were applied to uncover insights on the effect of flash crashes on asset correlations.

1 Motivation

1.1 Introduction

On May 6, 2010, at 14:32 and for about 36 minutes, the US financial markets experienced one of the most turbulent periods in their history, when stock index futures, options, and exchange traded funds, as well as individual stocks, experienced extreme price volatility accompanied by spikes in trading volume. [1] The S&P 500 Volatility Index increased by 22.5% and even if the major indices regained more than half of the lost value by the end of the day, the Flash Crash of 2010 took away approximately $1 trillion in the market value. [2] With different investigations concluding that high-frequency traders played a significant role in the crash.

More generally, a flash crash refers to an event where prices of the overall market or a particular stock decline rapidly then recover within a short period of time, typically caused by a rapid sell-off of securities. And with the emergence of digitalized trading, flash crashes today are usually triggered by computer algorithms, that can automatically sell large volumes at an extreme pace in reaction to economic news. [3] They occur frequently, but, unlike May 6, 2010, most are very small and do not make the news. [4]

In financial markets, the asset correlations a constantly monitored to construct efficient trading strategies. Where multiple factors affect these correlations. Assets with shared exposure to the same macroeconomic risks often exhibit similar behavior, while those within the same sector, driven by the same risk factors also move accordingly. [5] Correlations between assets are also subject to change over time due to various reasons, such as policy changes, investor behavior or market shocks. For example, during the 2008 financial crisis, many previously uncorrelated assets became highly correlated, increasing portfolio volatility and risk. On the other hand, during the 2020 Covid-19 pandemic, previously correlated assets became less correlated, creating more opportunities for diversification. [6]

It is common practice among investors to group highly correlated assets together to develop trading strategies. Commonly looking at historical data and using unsupervised learning methods. Given that market shocks influence asset correlations, and considering the flash crash as a major market shock, we ask ourselves two questions. Did the flash crash of 2010 have an impact on asset correlations, will we obtain different asset clusters before, during and after the events of the flash crash? We also ask ourselves, do smaller events of high activity on the market also impact asset correlations and the asset clusters?

1.2 Aim

Leveraging intraday trading datasets, we aim to answer our questions by investigating the effect of the 2010 Flash Crash on asset correlations, aiming to obtain asset clusters before, during and after the event. We also aim to find other periods of high activity, possibly tiny flash crashes, and also investigate their impact on asset correlations.

Bla bla compare, is flash crash just another small flash crash?

2 Dataset

Our dataset comprises daily trade and best bid/offer files for 29 publicly traded stocks, covering all 252 trading days of the year 2010. Each trade file is provided in a compressed .csv.gz format. The 29 stocks are:

'AAPL', 'AMGN', 'AXP', 'BA', 'CAT', 'CSCO', 'CVX', 'DOW', 'GS', 'HD', 'IBM', 'INTC', 'JNJ', 'JPM', 'KO', 'MCD', 'MMM', 'MRK', 'MSFT', 'NKE', 'PFE', 'PG', 'RTX', 'TRV', 'UNH', 'UTX', 'V', 'VZ', 'WBA', 'WMT', 'XOM'.

Each stock can be classified according to its sector on Nasdaq:

- Technology: 'AAPL', 'IBM', 'INTC', 'MSFT'
- Healthcare: 'AMGN', 'JNJ', 'MMM', 'MRK', 'PFE', 'UNH'
- Finance: 'AXP', 'GS', 'JPM', 'TRV'
- Industrials: 'BA', 'CAT', 'DOW, 'RTX', 'UTX'
- Telecommunications: 'CSCO', 'VZ'
- Energy: 'CVX', 'XOM'
- Retail: 'HD'
- Consumer Staples: 'KO', 'WBA'
- Consumer Discretionary: 'MCD', 'NKE', 'PG', 'V', 'WMT'

3 Methods

3.1 Preprocessing

The preprocessing pipeline starts from the bbo and trade files of every asset to obtain one joined parquet file per asset, for the entire year 2010. The steps are performed iteratively for one asset at a time.

The trade and bbo files are loaded as time series, indexed and sorted by date and time. The data is filtered to include only trades within the market's active hours, between 9:30 - 16:00. The trade and bbo files are then combined as one DataFrame for every pair of matching days using a full outer join on the index column, ensuring all data is included. The overlapping entries on the index are combined.

To address missing values, forward filling is applied on the bid price, bid volume, ask price, ask volume, trade price and trade volume columns. The DataFrames are saved in parquet format, to preserve data types. Finally, the joined DataFrames are concatenated into a single parquet file. At the end of the preprocessing, there is one joined trade bbo file for the entire year 2010 for every asset, meaning 29 DataFrames.

3.2 Flash crash detection

3.2.1 Response function definition

Flash crashed are identified using the response function in trade time. The function is defined as follows [7]:

$$R(\tau) = \mathbb{E}\big[s_n(m_{n+\tau} - m_n)\big]$$

with τ the time lag, m_n the mid-price, p_n the mid-price, defined as the average between the bid-price and the ask-price and s_n defined as follows:

$$s_n = \text{sign}(p_n - m_n)$$

3.2.2 Response function profile on 2010-05-06

To identify flash crashes, the response function is computed for every 15 minute interval of a trading day, for τ equal to 1000. A sanity check is performed on 2010-05-06 to identify the 2010 Flash Crash, which is successfully identified between 14:30 and 15:15 observing the response function profiles of different assets.

The assets where the response function profile for 15 minute intervals doesn't allow to identify the 2010 Flash Crash are not kept for further analysis. This is because the goal of the study is to identify and compare asset clusters before, during and after the 2010 Flash Crash, and compare these with asset clusters obtained in the same manner from smaller flash crashes. A total of 7 assets are removed for further analysis. An asset is kept if during the day of the Flash Crash, the 15-minute interval response function with the highest absolute value corresponds to the interval 14:30 - 15:15.

3.2.3 Portfolio construction

To identify smaller flash crashes, a portfolio is constructed. The portfolio weights per asset are defined by the inverse of the sum of the median of observed response function absolute values, per 15 minute intervals, over a fixed period of time. The weights are then normalized so their sum is equal to 1. With different magnitudes in response function per asset, this portfolio allows for each stock to have equal impact on the portfolio. Creating a relevant portfolio for observing periods of simultaneous relative high values in response function for multiple assets.

The portfolio weight for each asset w_i is defined as:

$$w_i = \frac{1}{\sum_{j=1}^{N} Median\left(|R_j|\right)}$$

where:

- R_i is the response function for asset i over a 15-minute interval.
- Median $(|R_i|)$ is the median of the absolute values of the response function over a 15-minute interval i.
- N is the total number of 15-minute intervals available for the asset during the chosen fixed period of time.

This normalization ensures that the sum of all weights equals 1:

$$\sum_{i=1}^{K} w_i = 1$$

where:

- K is the total number of assets.

With a τ equal to 1000, some response functions show sudden volatility before τ reaching 1000 due to a lack of available data. The median, robust to outliers, gives less importance to these instances while still capturing if the response function absolute values are high.

3.2.4 Using the portfolio to detect smaller flash crashes

To identify the Flash Crash of 2010, the portfolio weights are calculated using the entire month of May as a fixed period of time, the median of the absolute value of the response function of the weighted portfolio is then calculated for every 15-minute interval. Meaning a numerical value is obtained for every 15-minute interval in the month of May. A sanity check is performed on the month of May, and the flash crash is successfully identified with a numerical value 4 times bigger than the numerical value associated to any other 15-minute period in the month of May.

To identify 15-minute intervals associated to smaller flash crashes, or instances where there is high activity on the market, the same method is applied iteratively to every month of 2010.

3.3 Correlation matrix estimation

3.3.1 Hayashi-Yoshida estimator

During periods of financial instability, such as flash crashes, accurately estimating covariance matrices is crucial for portfolio optimization. Traditional covariance estimators are generally unsuited due to high volatility and noise. To address these challenges, we implement the Hayashi-Yoshida (HY) estimator [8]. Additionally, eigenvalue clipping is employed to reduce noise and ensure numerical stability.

The HY estimator is designed to handle non-synchronous trading data by considering overlapping time windows. The cumulated covariance estimator is defined as:

$$\hat{C}^{HY} = \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \left(p_{1,t_{1,i}} - p_{1,t_{1,i-1}}\right) \left(p_{2,t_{2,j}} - p_{2,t_{2,j-1}}\right) K_{i,j} \tag{1}$$

with $p_{1,t_{1,i}}$: the mid-price of asset 1 at time $t_{1,i}$, $p_{1,t_{1,i-1}}$, the mid-price of asset 1 at the previous time step $t_{1,i-1}$, $p_{2,t_{2,j}}$, the mid-price of asset 2 at time $t_{2,j}$, $p_{2,t_{2,j-1}}$, the mid-price of asset 2 at the previous time step $t_{2,j-1}$, N_1, N_2 the number of observations for assets 1 and 2, and the kernel $K_{i,j}$ is given by:

$$K_{i,j} = \begin{cases} 1 & \text{if } \max(t_{1,i-1}, t_{2,j-1}) < \min(t_{1,i}, t_{2,j}) \\ 0 & \text{otherwise} \end{cases}$$

3.3.2 Equivalent Estimation

An equivalent estimation of the HY estimator is to multiply the returns r_{1,t_i} with the last known value.

3.3.3 Methodology

The following steps outline the computation of the covariance matrix for two assets:

1. Load intraday data for two assets over a fixed period.
2. Compute the mid-price for each asset.
3. Compute mid-price returns for both assets, retaining only non-zero returns.
4. Merge the two DataFrames using a full join (Polars).
5. Perform an in-place forward-fill to handle missing values.
6. Calculate the covariance estimator as the average product of the two mid-price returns. For the correlation estimator, divide by the respective standard deviations.

This method is applied iteratively for every pair of assets to obtain the full covariance matrix on the retained 22 assets.

3.3.4 Eigenvalue Clipping for Noise Reduction

Eigenvalue clipping is applied on the covariance matrix to reduce variance and mitigate noise. [9] The method adjusts the eigenvalues based on the Marčenko-Pastur (MP) distribution.

The MP distribution arises in random matrix theory and describes the asymptotic behavior of eigenvalues of large-dimensional sample covariance matrices. Specifically, when the number of assets N and the number of observations T both tend to infinity while their ratio $q = \frac{N}{T}$ remains constant, the eigenvalues of the sample covariance matrix $\hat{C}$ follow the MP distribution with support bounded by:

$$\lambda_{\pm} = \sigma^2 \left(1 \pm \sqrt{q}\right)^2$$

where σ^2 is the true variance of the underlying asset returns.

In practical applications, where N is comparable to T, the sample covariance matrix $\hat{C}$ is significantly influenced by noise. The MP distribution provides a theoretical benchmark to distinguish signal from noise by identifying the range within which the bulk of the eigenvalues lie due to random fluctuations. Eigenvalues that lie outside this range are considered to carry genuine information about the underlying asset correlations.

3.3.5 Methodology

The clipping process involves the following steps:

1. Determine the empirical eigenvalues $\lambda_i^{(e)}$ from the covariance matrix.
2. Determine Clipping Threshold λ_+ with the formula in 3.3.4. And compute the clipped eigenvalues with the formula:
$$\lambda_{\text{clip},i} = \begin{cases} \lambda_i^{(e)} & \text{if } \lambda_i^{(e)} \geq \lambda_+ \\ \varepsilon & \text{otherwise} \end{cases}$$
3. Compute ε:
$$N_{\text{clip}} = \#\{i \mid \lambda_i^{(e)} < \lambda_+\}$$
Given the trace condition $\text{Tr}(C) = N$, with C the estimated covariance matrix, solve for ε:
$$\sum_{i=1}^{N} \lambda_i^{(e)} + N_{\text{clip}}\varepsilon = N$$

4. Construct clipped covariance matrix with the formula:

$$C_{\text{clip}} = V_0 \, \text{diag}(\lambda_{\text{clip}}) \, V_0^\top$$

where V_0 is the matrix with eigenvectors of the estimated covariance matrix C. Finally, normalize to obtain the correlation matrix:

$$C_{\text{clip},ii} = 1$$

where $C_{\text{clip},ii}$ are the diagonal elements of the clipped covariance matrix.

The flash crash detection method considers one 45 minute interval the day of the 2010 Flash Crash, in May and then considers one 15 minute interval for every other month. A covariance matrix is estimated and then clipped before the considered time interval, during the considered time interval and then after the given time interval for the entire day. As the 2010 Flash Crash has been considered between 14:30 - 15:15, and the smaller flash crashes are considered less important and only for a 15 minute interval. A total of 3 clipped matrices are computed for every interval obtained in Table 3 and on the day of the 2010 Flash Crash, totalling 36 matrices.

3.4 Louvain clustering algorithm

3.4.1 Louvain Algorithm Definition

The Louvain algorithm is a bottom-up, hierarchical clustering method that optimizes the modularity of a partition of the network, where the modularity is a measure to quantify the quality of a division of a network into communities. The process involves two main phases that are iteratively applied:

1. Initially, each node is assigned to its own community. The algorithm then iteratively considers moving each node to the community of its neighbors, choosing the move that results in the highest increase in modularity. This process continues until no further improvement can be achieved by moving any single node.
2. After the local optimization phase, a new network is constructed where each community identified in the first phase is represented as a single node. The edges between these new nodes are weighted by the sum of the weights of the edges between nodes in the corresponding communities. The algorithm then repeats the first step on this aggregated network.

This iterative process continues until no further improvements in modularity are possible.

3.4.2 Implementation of Louvain Clustering

The Louvain clustering was implemented using the Python `networkx` library and the `python-louvain` package. The implementation involves the following steps:

1. The adjacency matrix is prepared using the clipped covariance matrix as the basis for constructing the adjacency matrix used in the Louvain algorithm. A threshold θ is applied to the normalized adjacency matrix to filter out weak connections that may represent noise. The thresholding process is defined with:

$$A_{i,j} = \begin{cases} A_{i,j} & \text{if } A_{i,j} \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

where θ is the threshold value. Applying the method to the 2010 Flash Crash Day first, a threshold is calculated to obtain a significant number of clusters, this same threshold is then kept for every other clustering computation.
2. Begin with each asset in its own cluster.
3. For each asset, evaluate the potential improvement in modularity $\Delta Q_{i \to j}$ when moving the asset i to the community j of a neighboring asset. If the maximum ΔQ is positive, assign asset i to the community j that yields the highest ΔQ.
4. Repeat the iterative process for all assets until no further modularity improvements can be achieved.
5. Construct a new adjacency matrix representing the communities formed in the previous steps and repeat the clustering process on this matrix.

This method was applied to the 36 covariance matrix estimated in 3.3. Obtaining a total of 36 clusters. The threshold was chosen as $\theta = 0.1460$, giving 6 clusters for the time period before the 2010 Flash Crash, and is kept throughout every other cluster computation.

4 Results

4.1 Preprocessing

Each of the 29 assets has a final parquet DataFrame summarizing trade and bbo data for the year 2010. However, due to some empty trade and bbo files, some assets do not have records for every one of the 252 market days. The table below shows the number of trading days available for every asset.

Asset	Days Available	Asset	Days Available	Asset	Days Available
AAPL	222	AMGN	252	AXP	251
BA	252	CAT	252	CSCO	249
CVX	252	DOW	252	GS	252
HD	250	IBM	252	INTC	250
JNJ	252	JPM	251	KO	252
MCD	252	MMM	251	MRK	252
MSFT	252	NKE	252	PFE	251
PG	252	TRV	252	UNH	252
UTX	252	V	252	VZ	252
WMT	251	XOM	252		

Table 1: Number of Trading Days Available for Each Asset in 2010

4.2 Flash crash detection

4.2.1 Response function profile on 2010-05-06

When computing and plotting the response function for 2010-05-06, for AAPL, it is clear that the Flash Crash is successfully identified by the response function profile. With the value (Hour 14, Quarter 4) indicating the 14:45 - 15:00 time interval.

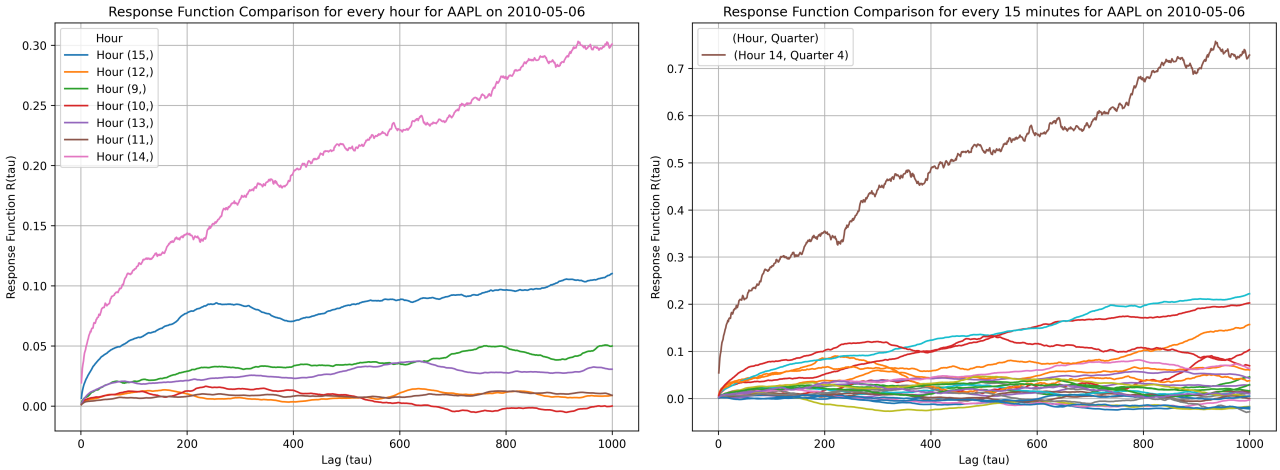

Figure 1: Response function profile of AAPL asset for 2010-05-06

The 7 assets that were removed for further analysis are : 'GS', 'JNJ', 'MCD', 'MSFT', 'NKE', 'PFE', 'UNH'. Their response profiles on 2010-05-06 are plotted below.

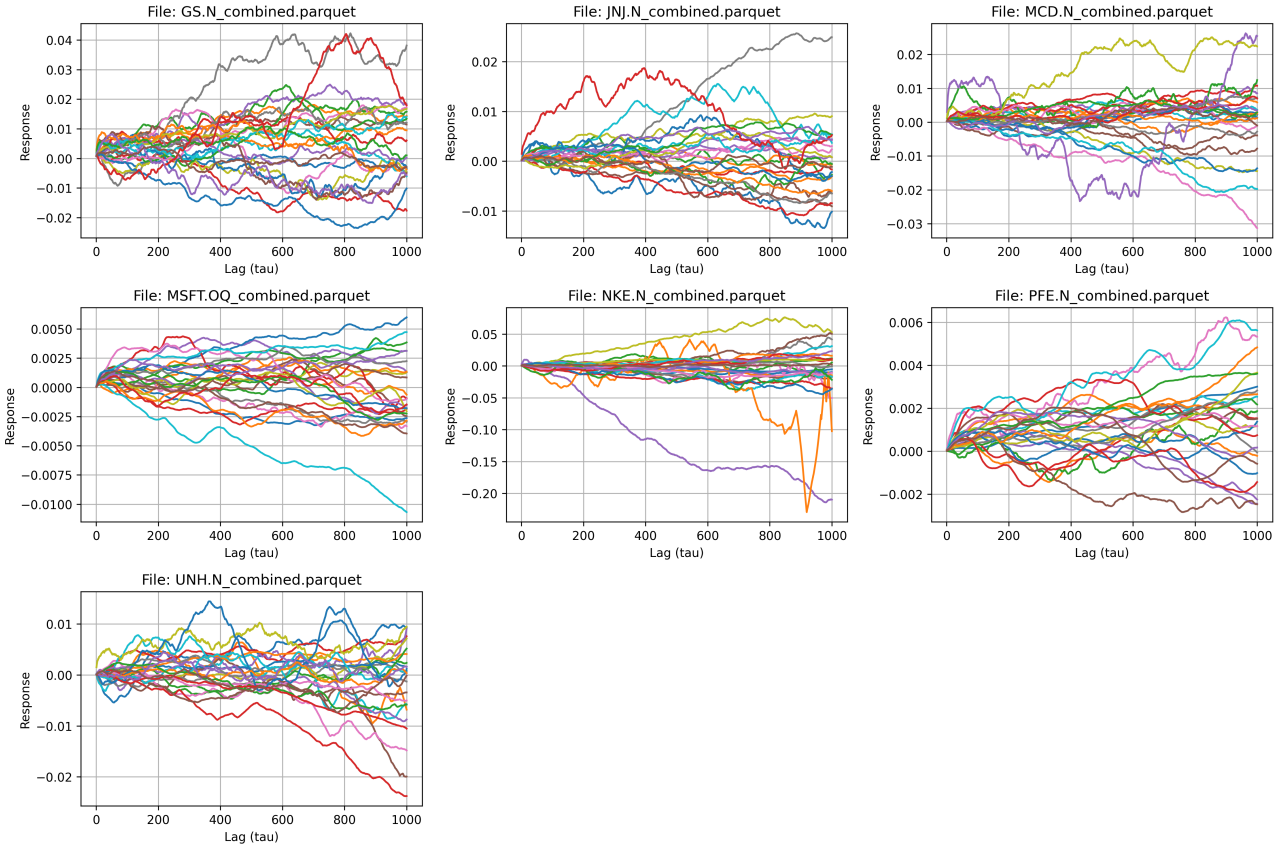

Figure 2: Response function profile of removed assets for 2010-05-06

A table is provided to show, for each asset, to which time period the peak in response function absolute value is associated to.

Asset	Max Absolute Value of response function	Time period
GS	0.04227120705381676	9:45-10:00
JNJ	0.025712589073635623	10:45-11:00
MCD	0.031307007786427206	13:30-13:45
MSFT	0.010680812288466813	12:15-12:30
NKE	0.22944444444444292	9:30-9:45
PFE	0.006235316407730764	9:30-9-45
UNH	0.02378717779246499	13:15-13:30

Table 2: Max Absolute Values and Corresponding Keys for Assets

For the asset NKE, the response function profile shows a higher value during the 2010 Flash Crash, however it also shows a high peak for the opening of the market day, 9:30 - 9:45, so the stock is not kept to avoid extra noise in the cluster computation for the time period before the 2010 Flash Crash further on in the analysis.

4.2.2 Using the portfolio to detect smaller flash crashes

Using the method described in 3.2.4 for the month of May, the highest value obtained is 0.0414 and corresponds to the 15-minute interval: '2010-05-06', 14:45-15:00. The distribution of the highest 30% values obtained during the month of May is plotted below:

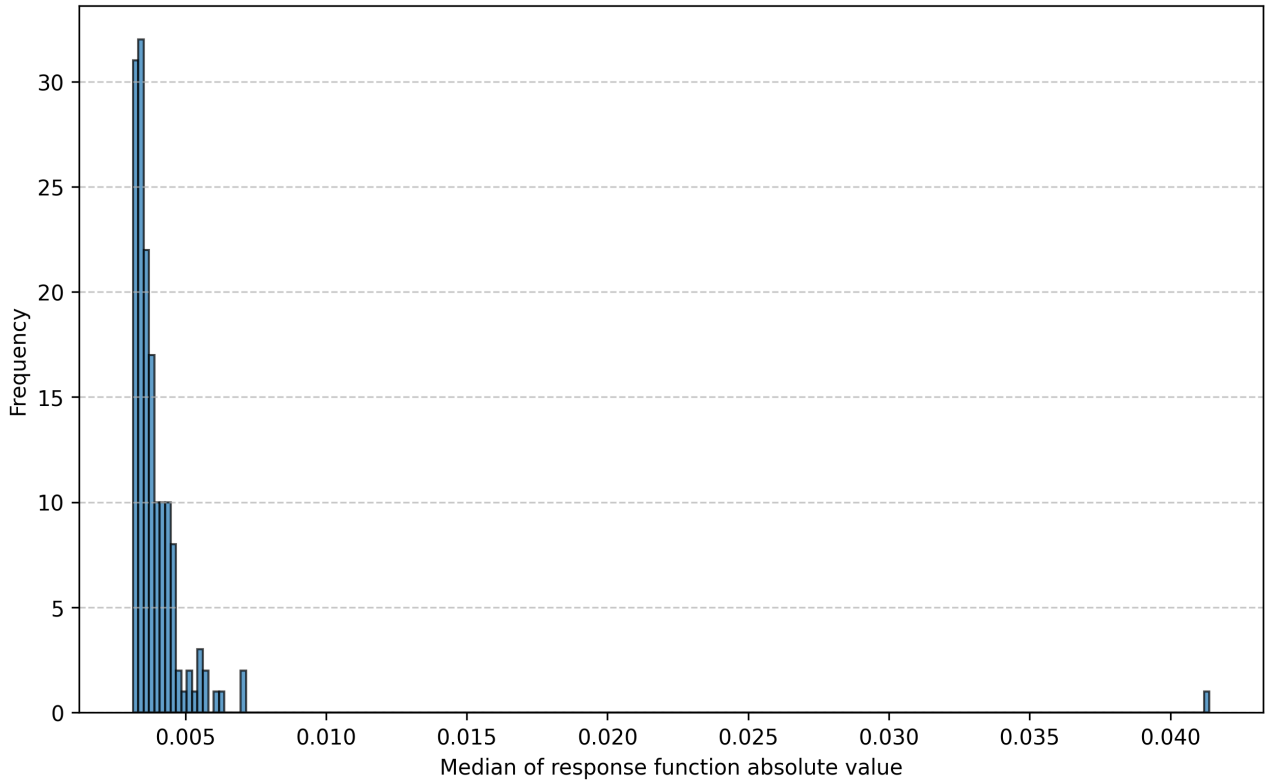

Figure 3: Distribution of top 30% median absolute values of response function per 15-minute interval of weighted portfolio for month of May

The distribution clearly shows an outlier for the Flash Crash, being more than 4 times higher than the second highest value obtained for the portfolio in the month of May. Applying the method to every other month of the year, the following 15-minute intervals are obtained, corresponding to periods where the constructed portfolio's median absolute value of the response function is highest for every month. Many highest values obtained where for the 9:30-9:45 intervals, when the market opens, these cases have been excluded because they are impractical for further analysis, when computing asset clusters before the event. In the case where the highest value obtained is in 9:30-9:45 are replaced with the highest values not in the same time period that month. The periods obtained, as well as the numerical values associated are shown in the table below:

Date	**Interval**	**Value**	**Month**
2010-01-27	14:00 - 14:15	0.0077	01
2010-02-03	13:00 - 13:15	0.0078	02
2010-03-08	12:30 - 12:45	0.0074	03
2010-04-05	10:00 - 10:15	0.0112	04
2010-06-11	13:15 - 13:30	0.0067	06
2010-07-23	12:45 - 13:00	0.0062	07
2010-08-20	12:45 - 13:00	0.0073	08
2010-09-23	14:45 - 15:00	0.0077	09
2010-10-28	13:00 - 13:15	0.0074	10
2010-11-26	13:15 - 13:30	0.0119	11
2010-12-31	12:00 - 12:15	0.0066	12

Table 3: Max Absolute Values and Corresponding Keys for Assets

The values found are never as high as for the 2010 Flash Crash. Plotting the distributions of the top 30% for every month, indicates that these intervals might still be significant.

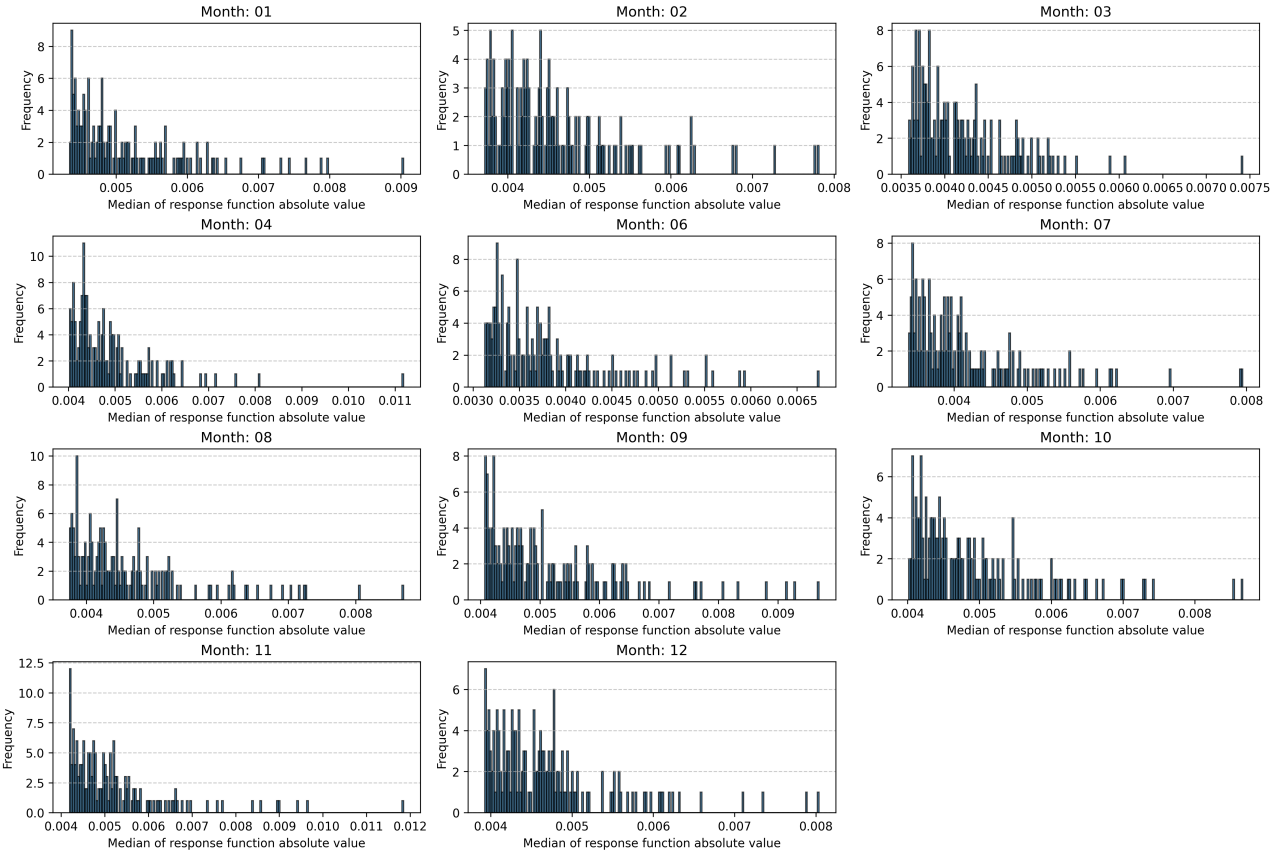

Figure 4: 15-minute intervals with highest value for weighted portfolio

4.2.3 Asset clusters obtained

Running the clustering method on the 2010 Flash Crash, for the time interval 14:30 - 15:15. We get the following clusters:

Cluster Type and Group	Stock Symbols
Before Clusters: 4	AAPL, AXP, CVX, HD, IBM, INTC, TRV
Before Clusters: 1	AMGN, DOW, UTX, XOM
Before Clusters: 2	BA, PG, VZ
Before Clusters: 3	CAT, V, WMT
Before Clusters: 5	CSCO, JPM, KO
Before Clusters: 0	MMM, MRK
During Cluster: 4	AAPL, BA, CAT
During Cluster: 1	AMGN, JPM, KO, WMT
During Cluster: 2	AXP, CVX, UTX
During Cluster: 3	CSCO, HD, INTC
During Cluster: 5	DOW, VZ, XOM
During Cluster: 0	IBM, MRK, V
During Cluster: 6	MMM, PG, TRV
After Cluster: 0	AAPL, AMGN, BA, CAT, CSCO, IBM
After Cluster: 1	AXP, CVX, DOW, INTC, KO, TRV, V
After Cluster: 2	HD, JPM, MMM, MRK, PG, UTX, VZ, WMT, XOM

Table 4: Clusters for flash crash day

With the following distributions of domains per cluster before, during and after the interval 2010-05-06 14:30 - 15:15.

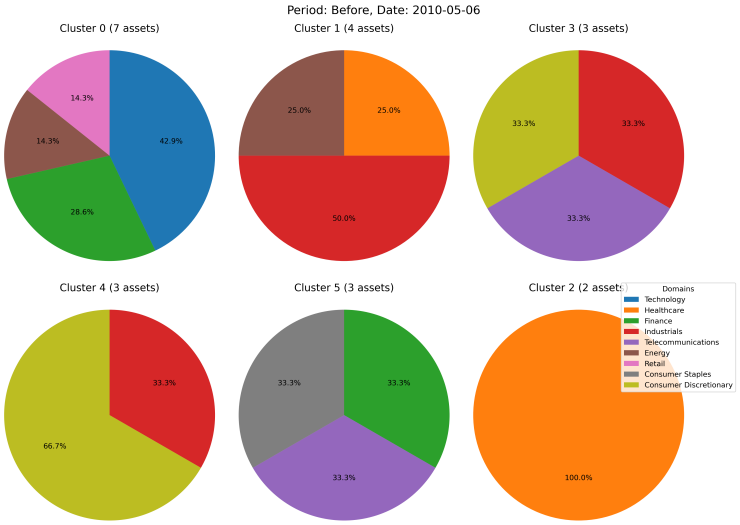

Figure 5: Domain distribution per cluster before Flash Crash

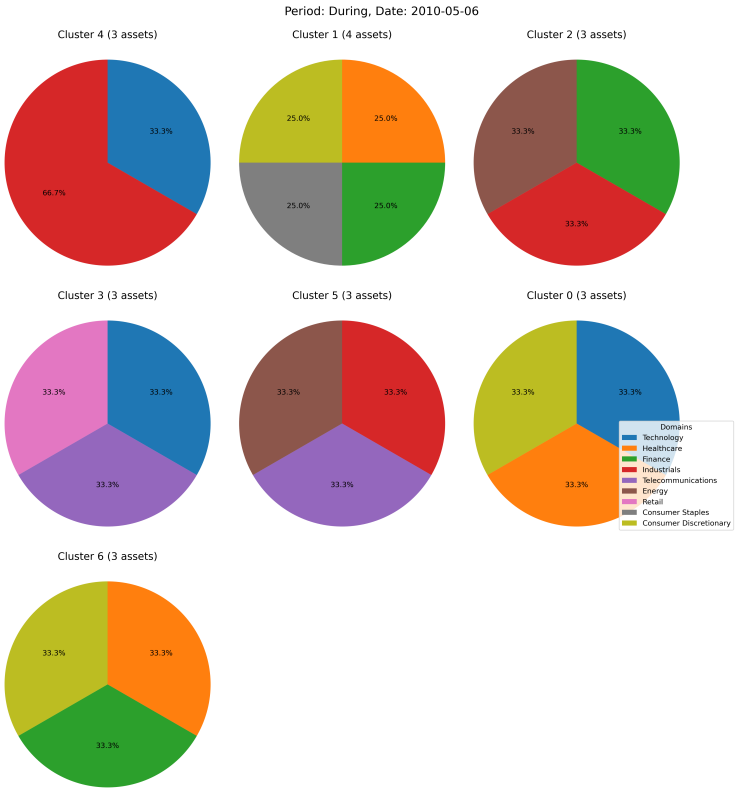

Figure 6: Domain distribution per cluster during Flash Crash

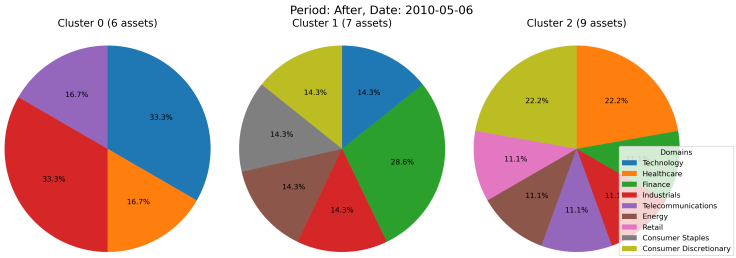

Figure 7: Domain distribution per cluster after Flash Crash

For the different periods (before, during, after) the time interval, we plot the distributions of cluster sizes.

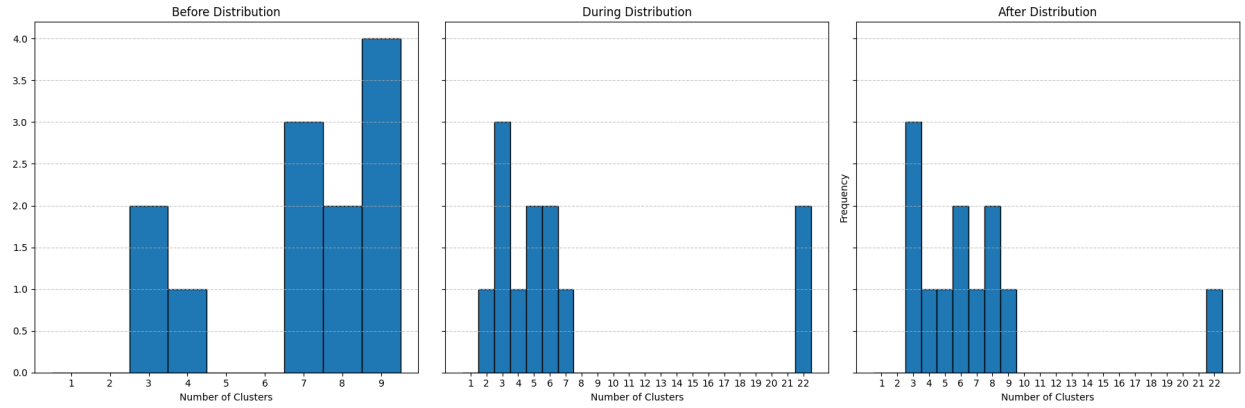

Figure 8: Distribution of cluster sizes considering all time intervals, before, during and after time interval

5 Discussion

5.1 Results interpretations

For the preprocessing, the final dataset obtained is almost complete with the exception of a select few assets that are missing data for not more than 3 days, and with the exception of the AAPL asset, where data is only available for 222 out of 252 days. We can conclude safely that there are no significant gaps in the data available, which could hinder the rest of the analysis.

The flash crash detection method shows excellent performance when applied to the intraday data of the month of May. The 15-minute interval corresponding to the flash crash is identified as a clear outlier, with the value associated to that time interval being more than 400% higher than any other value associated to any time interval for the entire month of May. Based on this result, we decide to use the method to detect smaller flash crashes.

When observing the results obtained for the method on every other month of the year, it is clear no time interval shows values near the one found for the Flash Crash of May 6th. The top values are often associated to the first 15 minutes of when the market opens, suggesting that markets show more activity during the opening minutes than throughout the rest of the day. Although we cannot use them because it wouldn't be possible to calculate clusters before the time interval for the same day in these cases.

No obvious patterns can be drawn from the clusters obtained during the 2010 Flash Crash or from the distribution of asset domains within the clusters. However, it is clear that the assets do not behave in the same groups for the three separate time intervals. Suggesting that the flash crash did indeed have a strong effect on how assets are correlated. The low number of 3 clusters for the time period following the Flash Crash suggests a potential convergence in asset behavior, which follows the period of extreme volatility induced by the Crash. This aligns with stylized facts in financial markets, where periods of high volatility are often followed by periods of lower volatility, as markets stabilize and participants adjust to the post-crisis environment. This could also reflect reduced differentiation in asset performance as uncertainty diminishes and stability gradually returns to the market.

Observing the distribution of cluster sizes for all 12 time intervals retained by the flash crash detection method, including the 2010 Flash Crash, no clear tendencies can be drawn. We do observe an extreme case for 2010-11-26, which corresponds to the second highest value obtained by the method, where every asset belongs to its own cluster for periods during and after the small flash crash. This might be due to the clustering algorithm being unfit for this case, also potentially suggesting that the clustering method is faulty. The clusters obtained and domain of asset distribution within those clusters also do not show any clear pattern.

5.2 Challenges encountered

The preprocessing pipeline initially did not return such a complete dataset, with several trade and bbo files could not be loaded without errors, a method was devised to cast different types to a subset of columns to allow for these files to be used further for merging procedures.

The flash crash detection algorithm was first developed using the highest peaks and mean values of the response functions associated to time intervals, which did not allow for a clear identification of the 2010 Flash Crash because of

many response functions flaring up in the cases where insufficient data was available to compute them entirely until τ reached 1000.

The large size of data used also posed a challenge throughout the project, with long run times to obtain results, throughout all the steps of the project.

5.3 Suggested improvements

The use of more time intervals to plot the cluster size distributions could potentially result in clear observable differences between the distributions, or to conclude with certainty there are no clear observable differences. Although this comes with the risk of using time intervals that are not periods that can be defined as small flash crashes.

The flash crash detection algorithm currently assumes a uniform distribution throughout the year, assuming an instance can be found for every month. Using a clear threshold and having insights on the actual distributions of flash crashes through empirical data could refine the method.

Using intraday data on multiple years might also increase the number of significant flash crashes available for further analysis. The use of more assets that allow for the identification through their response function profiles will also bring more data for further analysis.

The use of a different clustering method, such as the Giada and Marsili clustering, might bring more significant clusters. This method offers advantages as it incorporates domain specific structures that might give more meaningful clusters, although it is more computationally expensive. [10]

6 Appendix

Git link: https://github.com/Eliasepfl/bigData

References

[1] Andrei Kirilenko. The flash crash: The impact of high frequency trading on an electronic market. https://www.epfl.ch/schools/cdm/wp-content/uploads/2019/03/Kyle-paper.pdf.

[2] CFI. 2010 flash crash. https://corporatefinanceinstitute.com/resources/equities/2010-flash-crash/, .

[3] Investopedia. What is a flash crash. https://www.investopedia.com/terms/f/flash-crash.asp#:~:text=The%20term%20flash%20crash%20refers,resulting%20in%20dramatic%20price%20declines.

[4] CFI. Flash crashes. https://corporatefinanceinstitute.com/resources/career-map/sell-side/capital-markets/flash-crashes/, .

[5] Nasdaq. Understanding correlation between asset classes. https://www.nasdaq.com/articles/understanding-correlation-between-asset-classes.

[6] Fastercapital. Relationship between different assets and their impact on your portfolio. https://fastercapital.com/content/Asset-Correlation-Analysis--How-to-Understand-the-Relationship-Between-Different-Assets
html#Key-Takeaways-and-Recommendations-for-Asset-Correlation-Analysis.

[7] Marc Potters Jean-Philippe Bouchaud, Yuval Gefen and Matthieu Wyart. Fluctuations and response in financial markets: the subtle nature of 'random' price changes. https://iopscience.iop.org/article/10.1088/1469-7688/4/2/007/pdf.

[8] Takashi Hayashi and Naohiro Yoshida. On covariance estimation of non-synchronously observed diffusion processes. https://doi.org/10.3150/bj/1116340299.

[9] Vincenzo Tola, Fabrizio Lillo, Mauro Gallegati, and Rosario N. Mantegna. Cluster analysis for portfolio optimization. https://www.sciencedirect.com/science/article/pii/S0165188907000462.

[10] Lorenzo Giada and Matteo Marsili. Algorithms of maximum likelihood data clustering with applications. http://dx.doi.org/10.1016/S0378-4371(02)00974-3.