# Stock Market Analysis using Hierarchical Agglomerative Algorithm

**TEAM NAME: Five Analysts**

**TEAM NUMBER: 2**

**TEAM MEMBERS:**
Anishka Vaitla
Ramesh Kumar
Tarun Kumar Sahu
Varivashya Poladi
Yogangi Tiwari

## Introduction

This project develops an advanced analytical model to visualize and interpret the clustering patterns of stock market fluctuations during periods of global financial crisis. By constructing a network in which individual stocks are represented as nodes and the correlations between their price movements as weighted edges, this model uncovers critical insights into market dynamics and interconnectedness under stress.

In financial networks, nodes are linked by edges that represent correlations above a certain threshold, with higher edge weights indicating stronger co-movement. These weights, derived from normalized correlation coefficients, allow for a nuanced view of stock relationships. Applying the Agglomerative Generative Model (AGM) to these network reveals clusters—groups of stocks that exhibit synchronized behaviour—which become especially significant during crises. Understanding these clusters can be instrumental for retail investors seeking to navigate volatile markets.

Focusing on the 2008 Global Financial Crisis and the 2020 COVID-19 pandemic, our analysis contrasts the market structures immediately before and after each event, tracking the effects of the crisis until they subside. By examining shifts in network clusters, we provide a comparative view of how market relationships intensify or dissolve in response to economic turbulence.

This network-based approach to analyzing stock correlations offers a valuable framework for identifying patterns and relationships that may inform investment strategies during times of crisis.

## Datasets Used:

We have chosen the S&P 500 stock index as our primary dataset. We extracted essential attributes for S&P 500 stocks, including ticker, date, open, close, high, low, and volume, from NASDAQ and NYSE stock indices. The data we use spans from 2007 to 2023, but we will be specifically looking at the data from the years 2007-2012 and 2018-2023, covering both major crises that we aim to analyze.

|  | 2007 - 2012 | 2018 - 2023 |
|---|---|---|
| No. of NASDAQ files | 1509 | 1509 |
| No. of NYSE file | 1519 | 1509 |
| No. of rows in each NAS file | 820 | 804 |
| No. of rows in each NYSE file | 1088 | 1164 |

# Methods

We have chosen to specifically focus on the time periods of 2007-2012 (encompassing the 2008 crisis – The Global Financial Crisis) and 2018-2023 (encompassing the 2020 crisis – The Great Lockdown Crisis), covering both crisis and non-crisis zones.

We are first creating the graphs of the networks by choosing appropriate nodes and edges to represent the financial networks and clustering various stocks.

## Calculation Of Edges:

- **Calculate Daily Returns**: For each stock, daily returns are computed using the formula:

  $$\text{Return}_t = (\text{Price}_t - \text{Price}_{t-1}) / \text{Price}_t$$

  where $\text{Price}_t$ is the stock's closing price on day t.

- **Compute Pearson Correlation**: For each pair of stocks, the Pearson correlation coefficient is calculated over a specified time window (say, 1 year) for their respective daily returns. The Pearson correlation coefficient formula for two stocks' returns X and Y is:

  $$P_{x,y} = \text{Cov}(X,Y) / \sigma_x\sigma_y$$

  where Cov(X,Y) is the covariance of returns between stock X and stock Y, and $\sigma_x$ and $\sigma_y$ are the standard deviations of returns corresponding to the stocks X and Y, respectively.

- **Set Correlation Threshold**: Often, a threshold is applied to the correlation values. Only pairs with a correlation above a certain level (e.g., 0.6 or 0.7) are connected by edges to focus on significant relationships.

**Experiments Done:**

- The reason we chose to make the edge weights equivalent to the Pearson correlations between the Prices is that the correlation between two stocks measures how closely they move together which provide insights into their co-movement behaviour. A high positive correlation indicates that the stocks tend to increase or decrease in value simultaneously, while a negative correlation indicates that they move in opposite directions.

- Our aim is to cluster stocks such those within a community are interdependent and economically linked in the market. By assigning correlations as edge weights, the graph exactly captures these relationships, helping identify clusters of stocks that tend to move together.
- We experimented with using other edge weights such as Spearman correlation, but we found that Pearson correlation worked best.

## Construction Of the Correlation Matrix Construction

- **Pairwise Correlation Calculation**:

  Using Pearson correlation, pairwise correlations are computed between the daily returns of all stock pairs within each time window. The Pearson correlation coefficient for two stocks X and Y with returns over n days is:

$$\text{Correlation}_{X,y} = \sum_{i=1}^{n} \frac{(Xi-X)(Yi-Y)}{(n-1)\sigma X\sigma Y}$$

- **Threshold Application**:

  A correlation threshold was applied to filter out weak correlations, retaining only significant relationships for the network structure. For instance, correlations above 0.7 might indicate a strong relationship, making an edge between nodes.

  **Experiments Done:**

  We have tried using thresholds of 0.5, 0.7, 0.8 and 0.9, and found that the threshold of 0.7 yields the best results.

- Window Size:

  In our experiment we are choosing window size 21 days. Generally, 63 days window would be sufficient for normal market conditions, but our focus is to observe changes of clusters during the crisis.

# Clustering Algorithms

## Agglomerative Generative Model:

The formula for the **AGM** distance between two clusters A and B is:

$$D(A,B) = \frac{1}{|A|\cdot|B|} \sum i \in A \sum j \in B \quad d(i,j)$$

where:

- |A| and |B| are the number of elements in clusters A and B, respectively.
- d(i,j) is the distance between points i and j, where i∈Ai  and j∈B.

This method tends to produce more balanced clusters, especially in cases where the clusters have similar size and density.

### Distance Matrix Construction

From the correlation matrix, **distance matrix** is created through the following:

$$D_{ij} = \sqrt{2(1 - \text{Pij})}$$

where $P_{ij}$ is pearson correlation coefficient.

### Hierarchical Representation:

The distance matrix give rise to linkage matrix which is used to represent the hierarchy clusters through dendrograms.

## Measuring of Cluster Quality:

**C**luster quality refers to how well the data points in a cluster are grouped together based on certain characteristics or features. It can be measured by modularity, conductance and coverage.

Conductance: Measure quality of the cut (C, C'). Given a graph of node i in G with adjacency matrix A the conductance C is:

$$\phi\text{ (C)} = \sum \frac{i\in c,\ j\in c\ Aij}{\min\{A(C), A'(C)\}}$$

where A(C) = Σ i ∈C Σ j ∈V Ai, j = Σi ∈C d(i), where d(i) is degree of ndoe i in G. Hence th conductance of graph G is given by ΦG = minC ⊂V Φ(C). It provides the goodness of a community.

Coverage : The graph coverage indicates how many vertices are assigned to clusters (i.e., the number of assigned vertices divided by the total number of vertices in a graph.

$$\text{Coverage} = 100* \frac{intra{-}community\ edges}{Total\ number\ of\ edges}$$

Modularity: It measures the strength of sets of a network into modules i.e. communities. It is defined as:

$$Q = \sum_{i=1}^{C}(eij - aix^2)$$

where eij is the fraction of edges with one end vertices in community Ci and the other in community Cj and ai = ki/2E is the fraction of ends of edges that are attached to vertices in community i.

## Louvain Algorithm:

- The Louvain algorithm is an algorithm that is widely used for identifying clusters (or communities) in graphs, where nodes within the same community are densely connected, while connections between communities are sparser.

- The Louvain algorithm works iteratively to optimize modularity which is a measure that quantifies the density of edges within communities compared to edges between communities.

- We have experimented with Louvain algorithm in place of AGM and obtained similar results in terms of quality of clusters.

# Results

The results we obtained are as follows.

## 1:

The results obtained from running the AGM on the data corresponding to a time frame of one year are given below:

| Year | No. Of Nodes | No. Of Edges | No. Of clusters | Av. Clustering Coefficient |
|---|---|---|---|---|
| 2007 | 371 | 5129 | 239 | 0.355 |
| 2008 | 371 | 24343 | 48 | 0.684 |
| 2009 | 371 | 12102 | 73 | 0.541 |
| 2010 | 371 | 1253 | 94 | 0.241 |
| 2011 | 371 | 26100 | 51 | 0.697 |
| 2012 | 371 | 9163 | 219 | 0.464 |
| 2018 | 371 | 6358 | 157 | 0.051 |
| 2019 | 371 | 5348 | 125 | 0.078 |
| 2020 | 371 | 54325 | 85 | 0.890 |
| 2021 | 371 | 13880 | 180 | 0.581 |
| 2022 | 371 | 1499 | 71 | 0.430 |
| 2023 | 371 | 3498 | 265 | 0.505 |

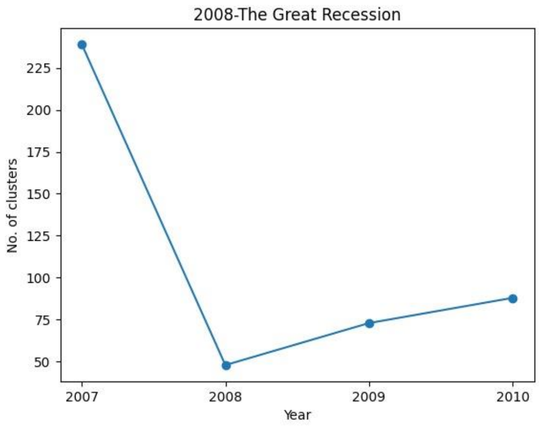
The results obtained from running the AGM on the data corresponding to a time frame of one month are given below:

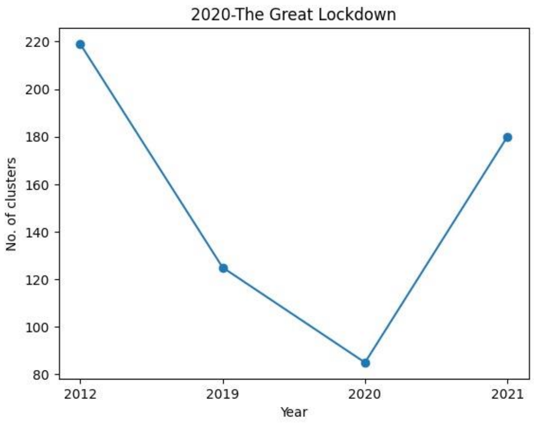
| Month/Year | No. Of Nodes | No. Of Edges | No. Of clusters | Av. Clustering Coefficient |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| 01/2008 | 371 | 5129 | 239 | 0.355 |
| 02/2008 | 371 | 24343 | 48 | 0.684 |
| 03/2008 | 371 | 10695 | 66 | 0.523 |
| 04/2008 | 371 | 2423 | 93 | 0.317 |
| 05/2008 | 371 | 3625 | 92 | 0.383 |
| 06/2008 | 371 | 7862 | 77 | 0.508 |
| 07/2008 | 371 | 4469 | 80 | 0.426 |
| 08/2008 | 371 | 9654 | 69 | 0.541 |
| 09/2008 | 371 | 18553 | 59 | 0.647 |
| 10/2008 | 371 | 20453 | 54 | 0.672 |
| 11/2008 | 371 | 29957 | 51 | 0.436 |
| 12/2008 | 371 | 20618 | 55 | 0.650 |

| Month/Year | No. Of Nodes | No. Of Edges | No. Of clusters | Av. Clustering Coefficient |
|---|---|---|---|---|
| 01/2020 | 371 | 1996 | 92 | 0.348 |
| 02/2020 | 371 | 12605 | 67 | 0.582 |
| 03/2020 | 371 | 34703 | 42 | 0.755 |
| 04/2020 | 371 | 20188 | 55 | 0.648 |
| 05/2020 | 371 | 13794 | 65 | 0.569 |
| 06/2020 | 371 | 23709 | 52 | 0.681 |
| 07/2020 | 371 | 2714 | 87 | 0.380 |
| 08/2020 | 371 | 1829 | 94 | 0.300 |
| 09/2020 | 371 | 6979 | 66 | 0.512 |
| 10/2020 | 371 | 4239 | 75 | 0.431 |
| 11/2020 | 371 | 8729 | 68 | 0.524 |
| 12/2020 | 371 | 1140 | 98 | 0.278 |

Below are the graphs plotted corresponding to the number of clusters made by the algorithm for the given threshold over the years.

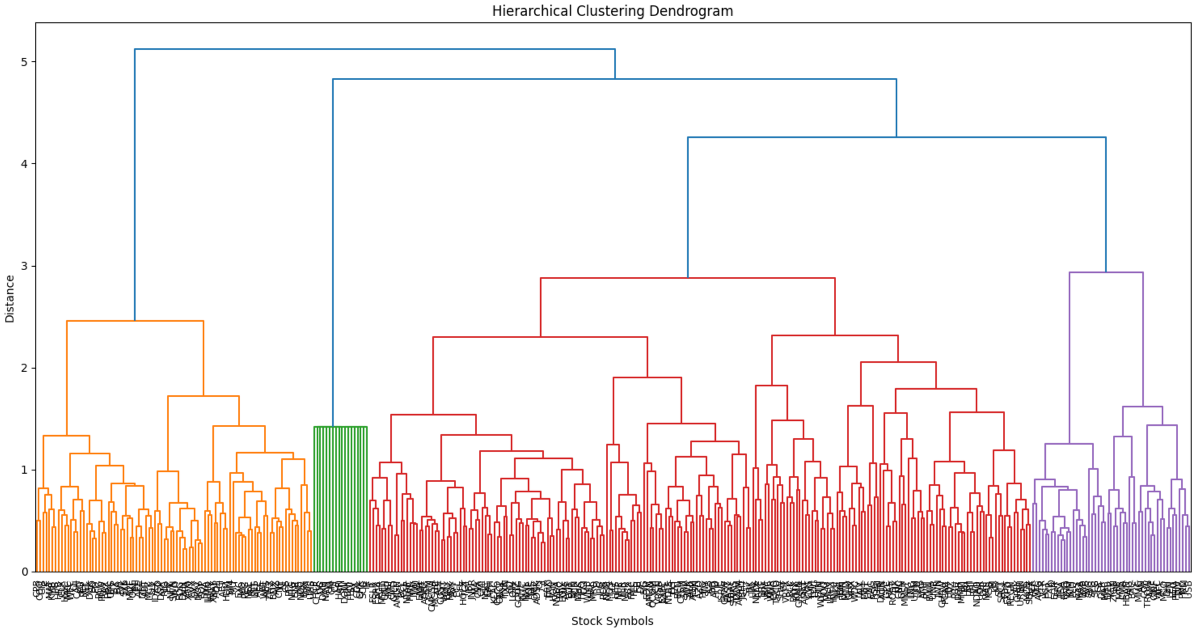




The observations we made from this data is that when there is a major crisis, the number of clusters formed by the AGM significantly decreases. As we can see, there is a clear local minima whenever there is a major financial crisis. Further, the rate at which the market recovery occurs is reflected in the slope of the curve after the crisis occurs. We infer from these results that there is generally a local dip in the graphs whenever the market crashes or is unstable.

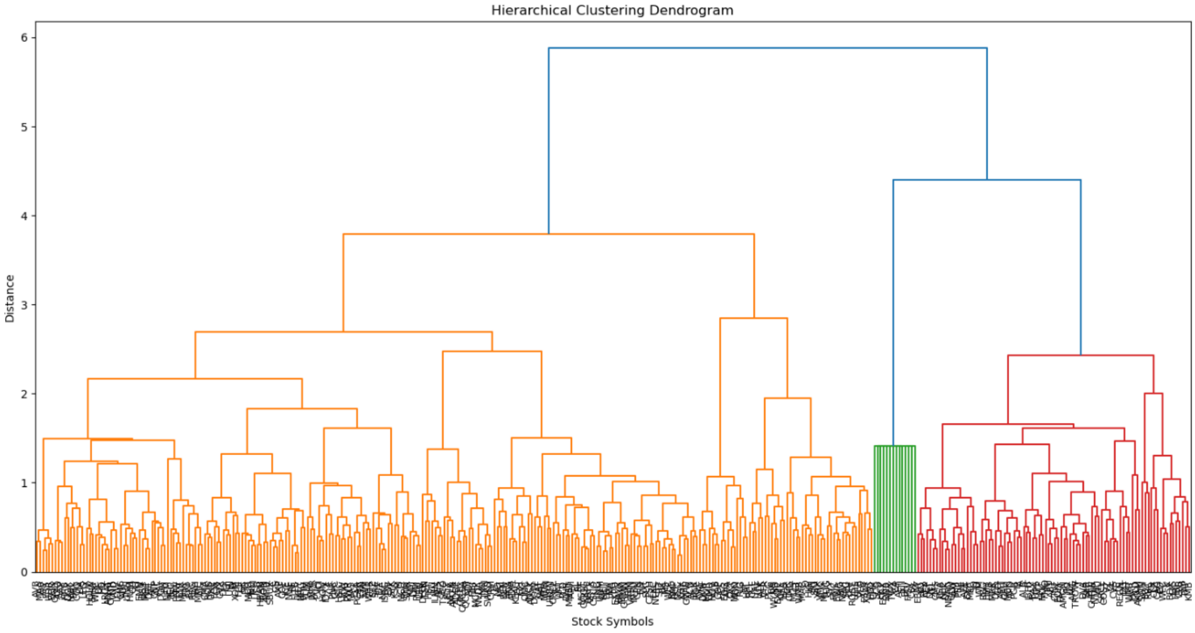
## 2:

Dendrogram for 2008-09-15 to 2008-10-15:



Number of nodes: 371
Number of edges: 24343
Number of clusters: 48
Average clustering coefficient: 0.890

Dendrogram for 2020-03-02 to 2020-04-01:



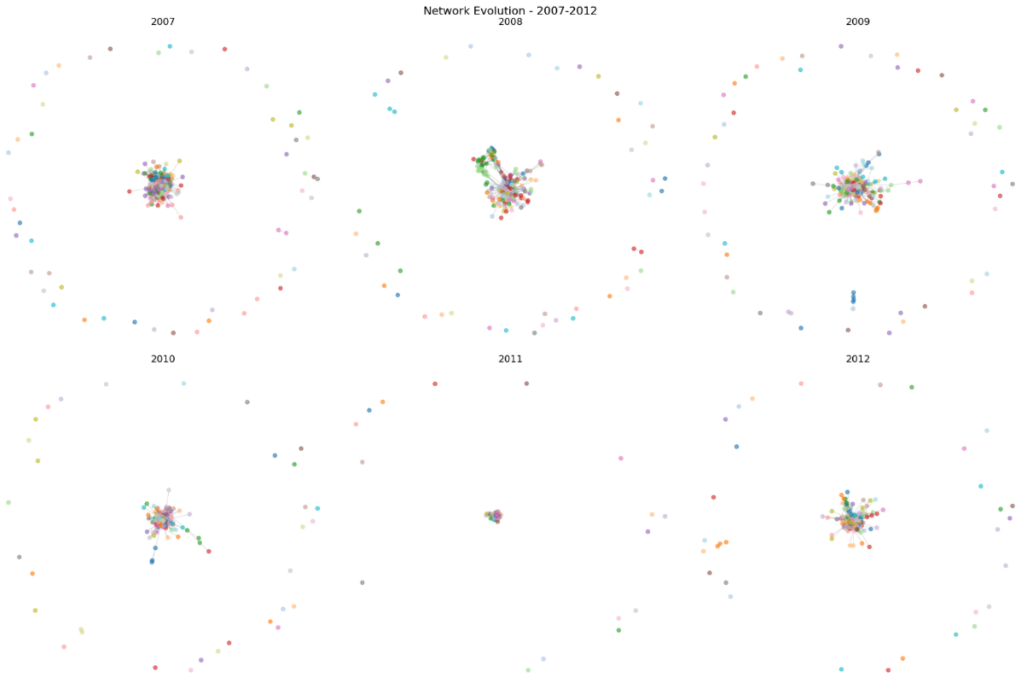
Number of nodes: 371
Number of edges: 54325
Number of clusters: 85
Average clustering coefficient: 0.890

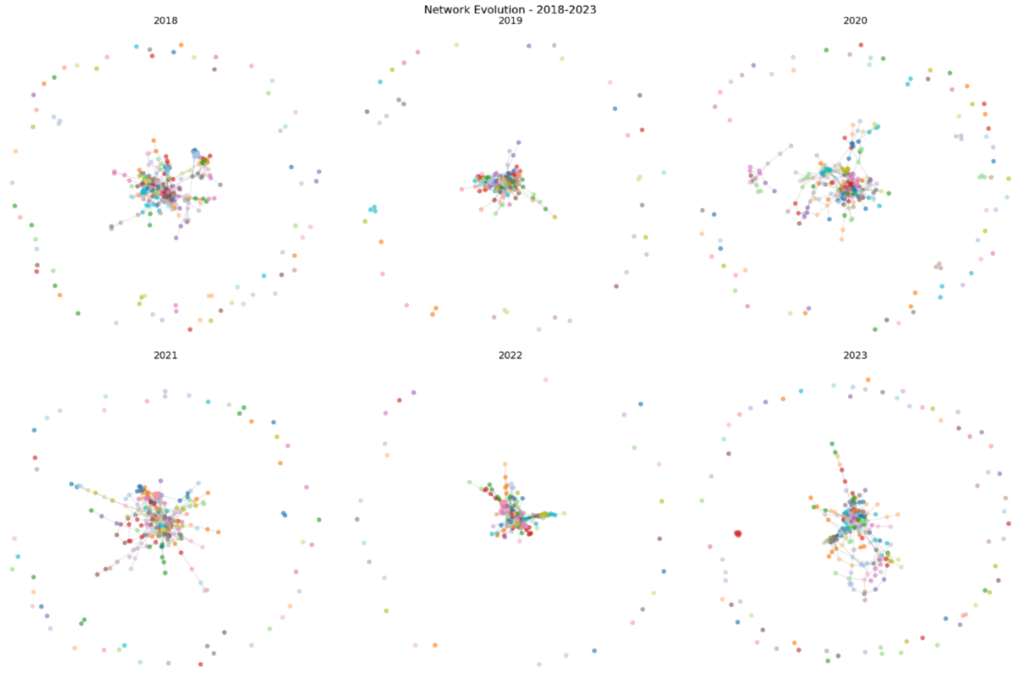
Here, the cluster density is inversely proportional to cluster size.

The dendrograms give a visualization of the communities obtained from the model (AGM).

## 3:

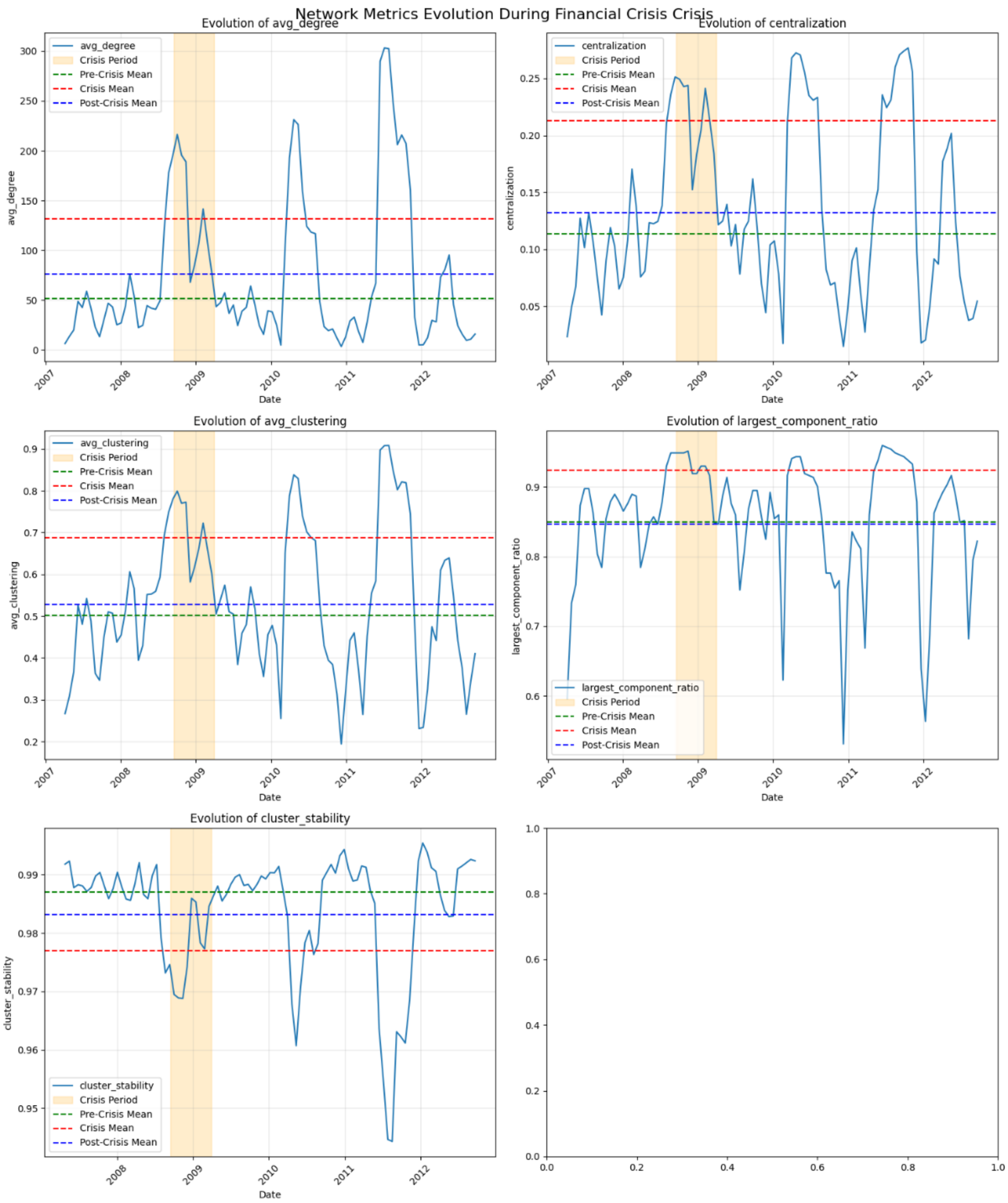
The images below illustrate the evolution of the stock correlation network from 2007 to 2012, and 2018 to 2023, showing snapshots from selected dates. Each node represents a stock, and each colour represents a different cluster of stocks that are more closely correlated with each other.





The changes in the number and spread of clusters reflect how these relationships change, potentially in response to events like economic crises or other market shifts. These results also support our hypothesis that the number of clusters seems to decline whenever there is a financial crisis and slowly increase as the market builds its way back up.

## 4:



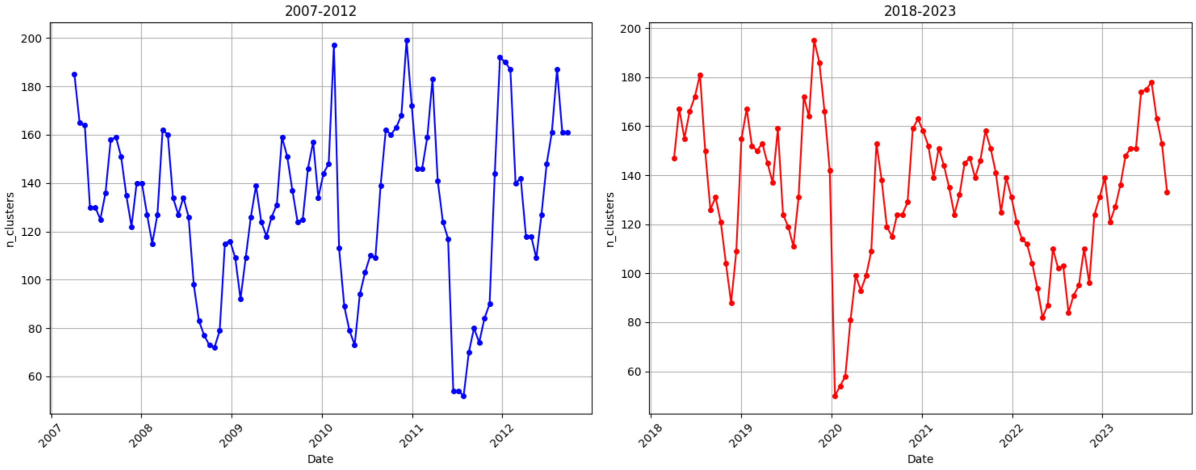
These plots illustrate how the 2008 financial crisis increased correlations among stocks, leading to a more interconnected and centralized network. The crisis disrupted some typical clustering patterns, and while the market stabilized post-crisis, some of these network metrics remained altered compared to pre-crisis levels.

The main inference from all the analysis is that a single hard blow to the pillars of economy results in results steep dip bur recovers fast enough too. But if there are hits not fundamental to economy but is continuous, no such steepness is observed. In crisis all the clusters tend to get closer, but the frequency and intensity of blows remain random or indeterministic (at least from we can observe from the data we analyzed).

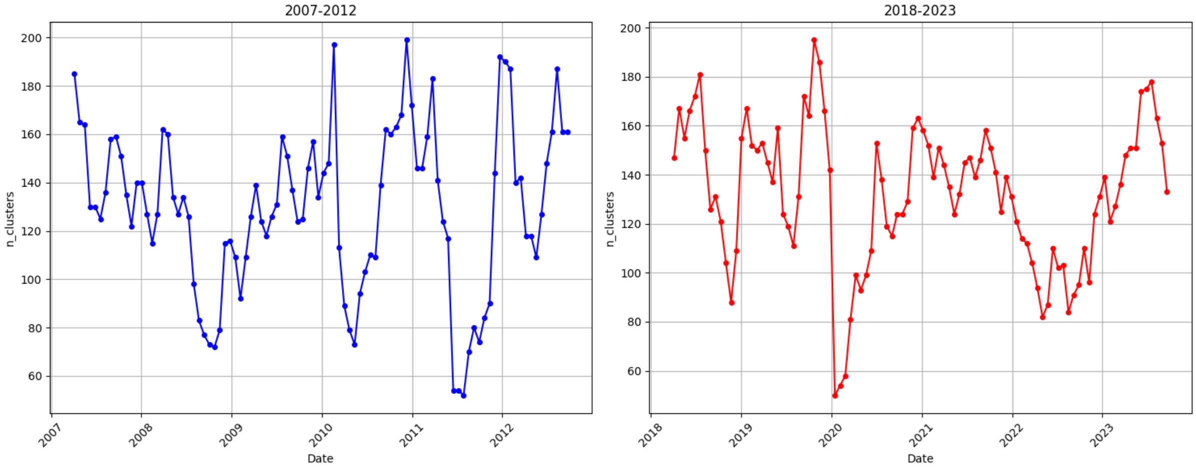
# 5:

Further metrics we have extracted from the stock market data for the given time frames are as follows:

### Cluster size with respect to year



### Plotting of Density Evolution



We were not able to extract any significant patterns or inferences from the above plots.

# Conclusion

During periods of crisis, stocks tend to cluster more closely, leading to a reduction in the overall number of clusters within the network. This pattern suggests a simultaneous decline across all sectors, which contrasts sharply with the more dispersed and sector-diverse clustering seen during stable, non-crisis times.

We arrived at this conclusion through our experiments as  an unexpected pattern emerged in 2007-2008 and 2019-2020, before the The Global Financial Crisis and the COVID-19 crisis respectively, where we observed a reduced number of clusters. Investigating further, we found that an economic slowdown and certain shutdowns had impacted the market during this period, causing stocks to exhibit crisis-like clustering behaviour.

From these observations, we hypothesize that economic prosperity fosters diversification, leading to a "declustering" effect in stock networks. In contrast, economic downturns cause stocks to converge, resulting in synchronous declines across sectors.

Overall, we have concluded that whenever there is a market instability or crash, stocks tend to cluster closer together and have stronger correlations with each other based on our experiments done with clustering