Financial stock returns and complex network analysis

Ixandra Achitouv LIP6

arXiv: 2407.20380 & 2408.11759

Outlines:

Financial market dynamics exhibit properties of complex systems:

- Non-linear behaviour & inter-dependencies of stock prices
- Emergence phenomena: trends, bubbles, adaptative behaviours
- Feedback mechanisms: rising prices →further investors →prices up
- Collective behaviours = market modes = synchronized movements of stock prices or groups of stocks.

Questions:

- How to identify these market modes? PCA, RMT, network science, stochastic field theory
- Can we interpret them? RDW simulations
- How to study dynamical inter-dependencies of stock returns and their collective behaviours?

Causality tests, forecasting

Stock returns correlation & RMT

Data S&P 500 stocks
485 stocks (T=4yrs)
1258 days of closing values for each stock

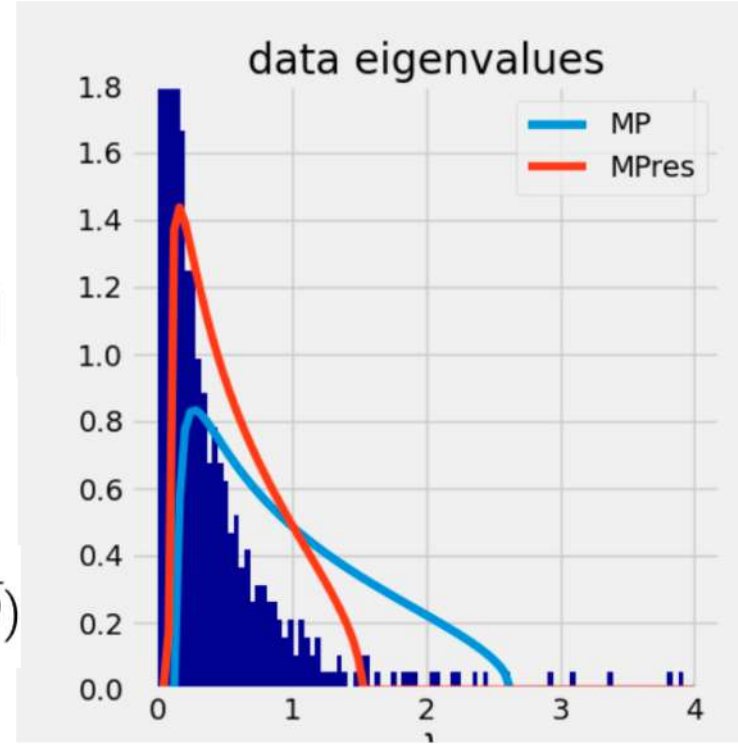
Marchenko-Pastur distribution:
N>>1, T>>1 and
T/N=Q=constant

$$C_{i,j} = \frac{< r_i(t) r_j(t) > - < r_i(t) >< r_j(t) >}{\sigma_i \sigma_j}$$

$$r_i(t) = \log[P_i(t)] - \log[P_i(t-1)]$$

$$\rho(\lambda) = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}}{\lambda}$$

$$\lambda_{min}, \lambda_{max} = \sigma^2(1 + Q \pm 2\sqrt{1/Q})$$



- The maximum of the eigenvalues λ_{market} is approximately an order of magnitude larger than λ_{max}.
- Treating sigma as an effective parameter → matching 94% of continuous spectrum Laloux et al. 1999
- The outliers from the continuum are correlations driven by economic factors.

Stock returns correlation & RMT

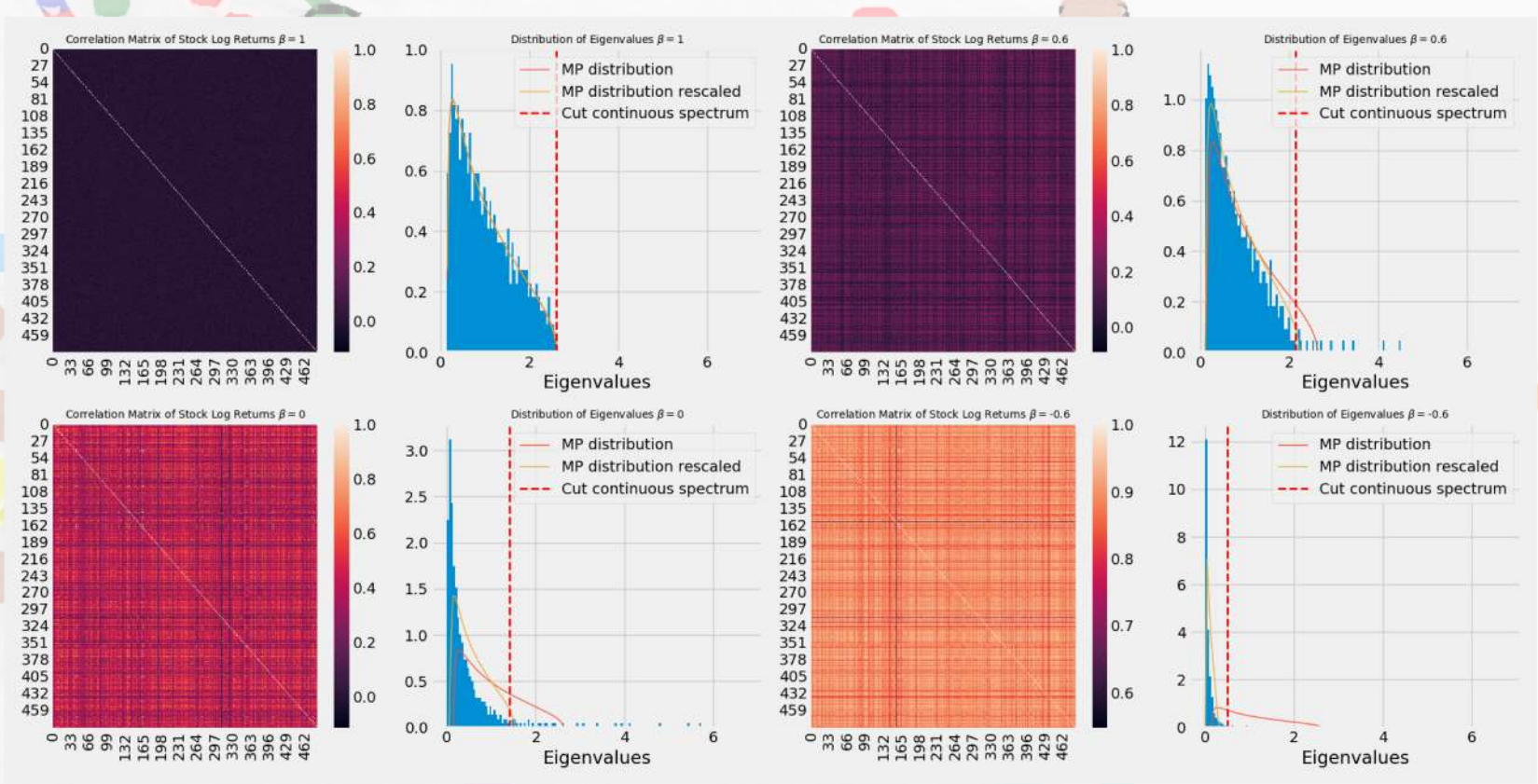
Varying signal-to-noise:

Geometrical Brownian Motion (GBM)

$$S_t^{\mathrm{GBM}} = S_0 \exp\left(\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t\right)$$

Input: expected return and volatility coefficient

Wt=Wiener process random component of stock price change

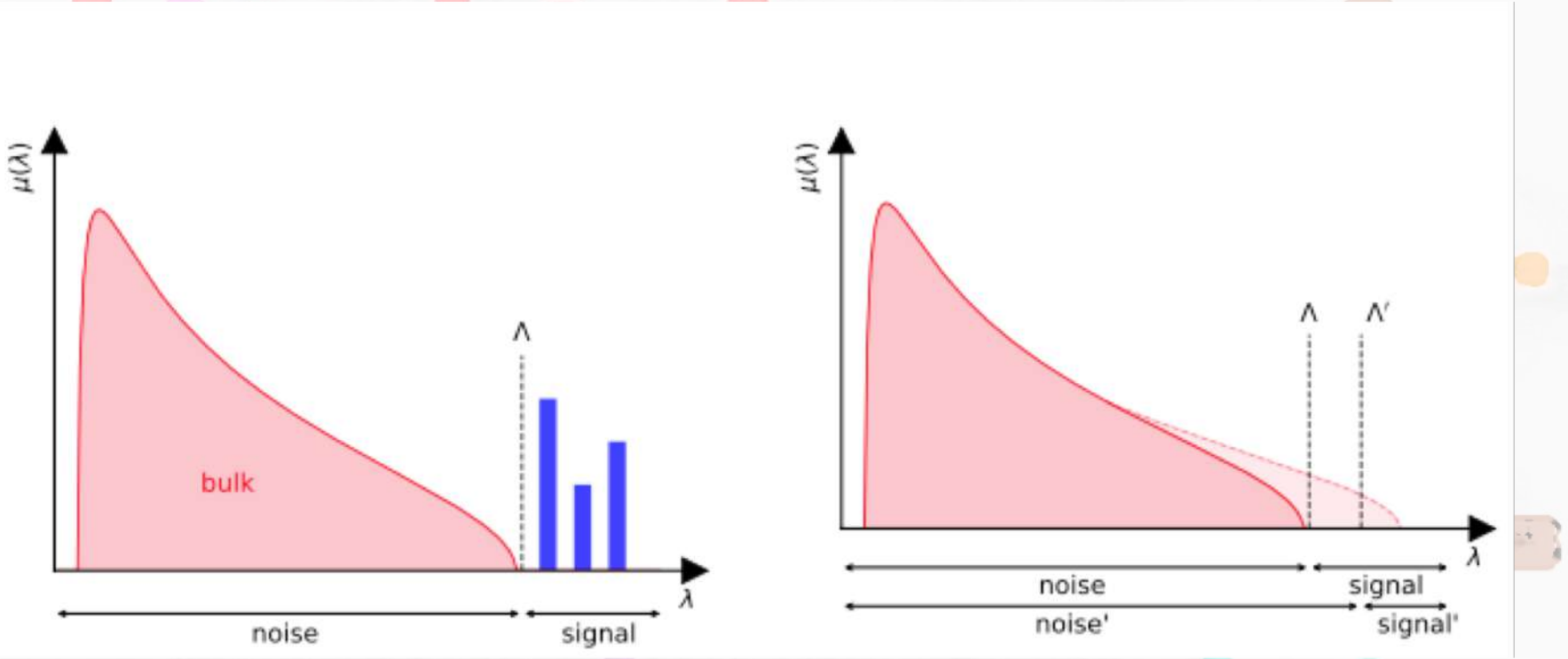


$$S_t^{\mathrm{sim}} = \beta S_t^{\mathrm{GBM}} + (1-\beta)S_t \;\forall \beta \geq 0$$

$$S_t^{\mathrm{sim}} = -\beta S_t^{\mathrm{GBM\ corr}} + (1+\beta)S_t \;\forall \beta < 0$$

Stock returns correlation & RMT

- Signal is distinct from the bulk→standard PCA to isolate it effectively.
- the continuity of the spectrum renders PCA ineffective at distinguishing between degrees of freedom



Signal can be diluted within the continuous spectrum of the eigenvalue distribution → see ArXiv: 2409.1971 for a field theory approach to infer signal in continuous spectra

Interpretations

- The largest eigenvalue is related to a strongly localized eigenvector that represents the collective evolution of the system = the Market mode.
- Its magnitude can be interpreted as the coupling strength of the system.
- Smaller effective value of Q could account for the existence of volatility correlations

What about the other values of Lambda > continuous spectra: Can we interpret them ? Stock sectors driven?

Splitting Market vs. noise modes

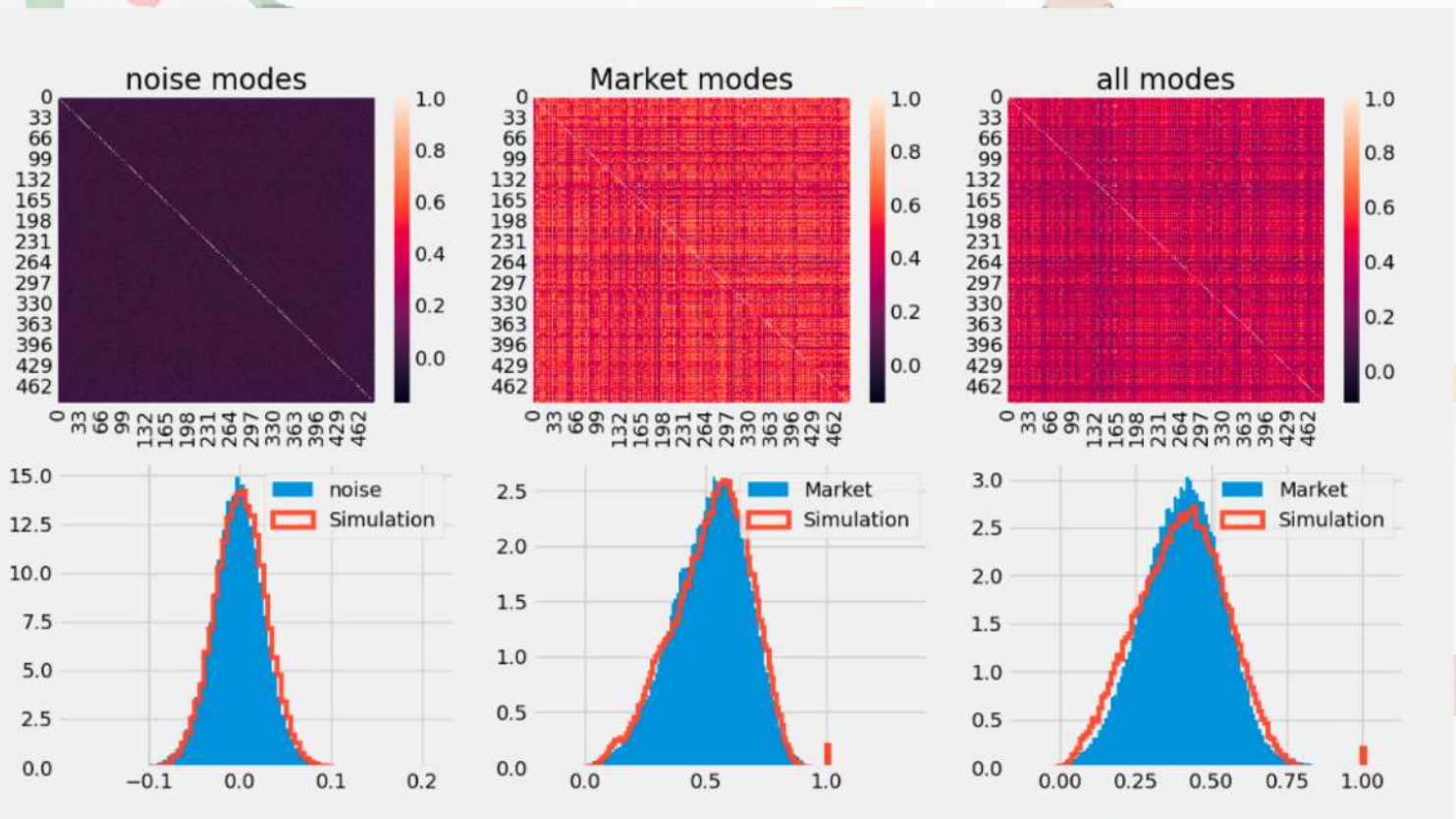
Correlation matrix of the Market modes:

- n_{market} = 12 largest eigenvalues selected

$$M = Q\Lambda Q^{-1}$$

- Q: N_{stocks}*N_{stocks} matrix, where i^{th} column is the eigenvector q_i of C_{ij} Lambda: diag matrix= λ_i if i is a market mode, 0 otherwise

$$C_{i,j}^{Market} = \frac{M_{i,j}}{\sqrt{M_{i,i}M_{j,j}}},$$



Can we infer Market stock returns correlation using network analysis and random walks?

Network analysis of stock returns

- **The threshold method:**

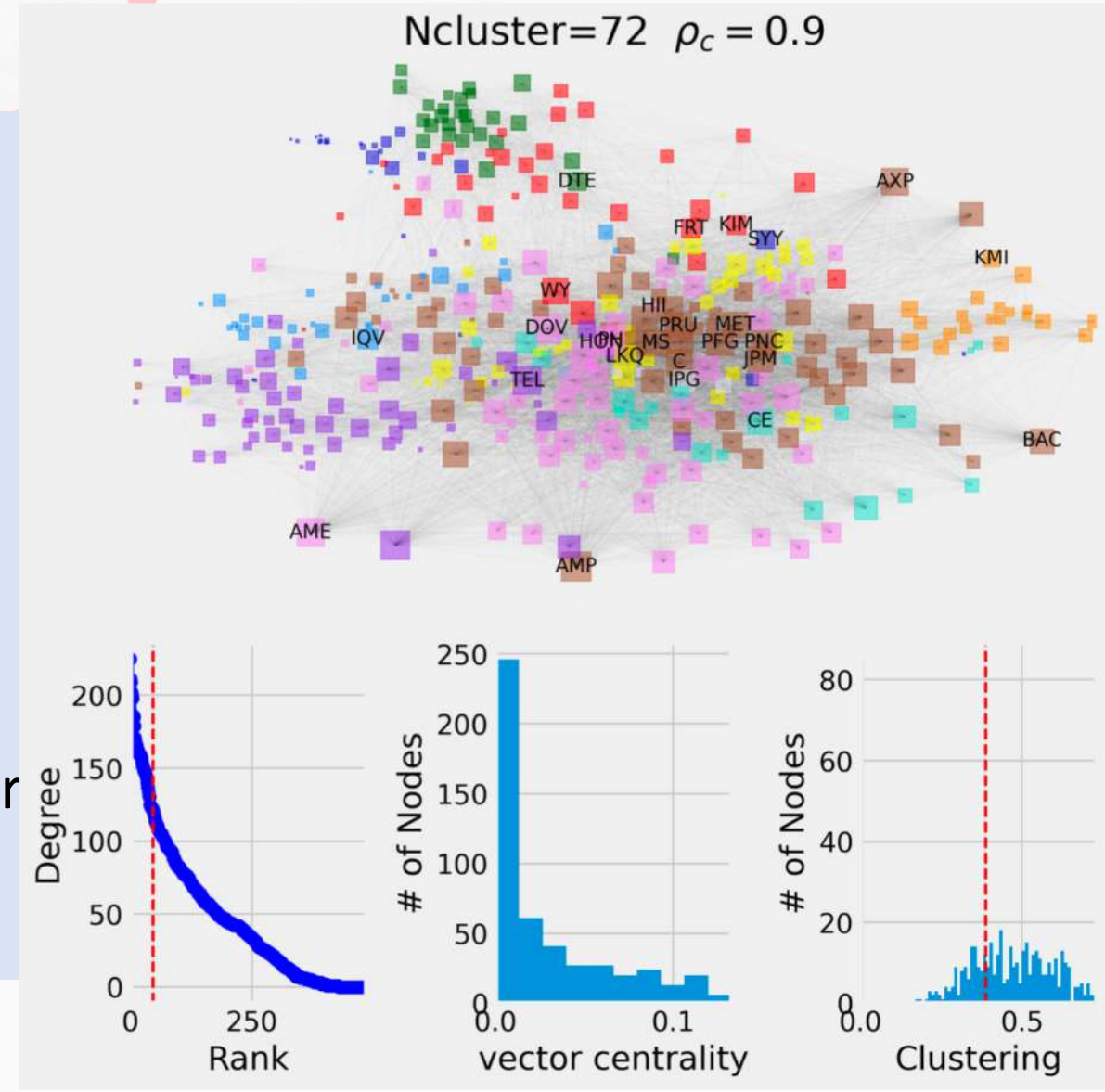
$C_{ij} \rightarrow A_{ij}$ if $|C_{ij}| >$ threshold

Threshold=0 → fully connected network flat degree pdf

- **Network properties:**

Louvain community finder: 72 clusters, 15 have more than 1 stock

12 most influential stocks = eigenvector centralities are in the top 3% of the distribution.



Inferring stock returns correlation matrix with correlated simulations of stock prices

Correlations per Influential stocks:
Fixed seed for each simulated stock price coupled with random seed
Strengh of coupling = largest correlation of the considered stock with one of the 12 most influencial stocks

Correlations per Louvain communities:
seed is given by the label of the community cluster the stock belongs to (72 clusters – 15 with more than 1 stock).
The strength of the correlation is given by the clustering value of the stock in the network

Model: The final simulated stock price is a weighted average of the two correlations type.
Weights are found minimizing the Wasserstein distance btw pdf of Cij_data and Cij_sim (same weights wL, wM for each stocks)

$$S_t = S_0 \exp\left(\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t\right)$$

$$S_t = \frac{w_L S_t^L + w_M S_t^M}{w_L + w_M}$$

Algorithm 1 Simulate GBM Stock Price with Correlation

function SIMULATESTOCKPRICEGBMCORR($S0, mu, sigma, T, coef, seedi$) ▷

Parameters:

S0 (float): Initial stock price.

mu (float): Drift coefficient (expected return).

sigma (float): Volatility (standard deviation of returns).

T (float): Total time period.

dt (float): Time step.

coef (float) : strength of the correlation

seedi (integer) : si seed used for correlation

Returns:

np.array: A simulated stock prices

$dt \leftarrow 1$ ▷ Time step

$N \leftarrow \lfloor T/dt \rfloor$ ▷ Number of time steps

$t \leftarrow$ Create a time grid from 0 to T with N points

Generate standard normal random variables $W1$ of size N with random seed

Generate standard normal random variables $W2$ of size N with seed= $seedi$

for i in range 0 to $N - 1$ **do**

$W[i] \leftarrow W1[i] * (1 - coef) + coef * W2[i]$ ▷ Generate correlated random numbers

end for

$W \leftarrow$ Cumulative sum of W multiplied by $\sqrt{dt}$ ▷ Wiener process

$X \leftarrow (mu - 0.5 * sigma^2) * t + sigma * W$

$S \leftarrow S0 * exp(X)$ ▷ Final stock prices

return S

end function

Results & Applications:

- For Market modes:

74% of the returns can be captured by the 12 Market stocks we identify using network properties. The other correlated modes correspond to the community clusters (not sector communities)

- Application: Use the Market simulated walks to improve on portofolio construction, finding the optimal weights using Cij^{Market}

Simulated market GBM to build an optimal portfolio and find that it outperforms standard mean-variance models based on historical prices when we re-balance the weights every dT days, with dT≤ 84

Dynamical analysis of financial stock network variables

Global network variables:

- **Average degree of the top 90% nodes**
- **Mean eigenvector centrality**: A high eigenvector score means that a node is connected to many nodes who themselves have high scores
- **Mean closeness centrality:** Closeness centrality of a node u measures the average length of the shortest path from the node to all other nodes in the network
- **Mean betweeness centrality:** Betweenness centrality of node v ratio btw the total number of shortest paths from node s to node t and the number of those paths that pass through v.

$$C_C(u) = \frac{1}{\sum_{v \in V} d(u,v)}$$

$$C_B(v) = \sum_{v \neq s \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

- high eigenvector values can be problematic for systemic risks as contagion risk is linked to these influential stocks
- could be used as a network systemic risk measure similarly to the eigenvector centrality, when building an optimal portfolio.

Dynamical analysis of financial stocks network variables

- **Mean clustering:**

s_i =sum of the weights of the edges connected to node i, k_i = the degree of node i, w_{ij} =weight of the edge btw nodes i and j, and a_{ij} = 1 if nodes i and j are connected, 0 otherwise

$$C_C(i) = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{(w_{ij} + w_{ih})}{2} a_{ij} a_{ih} a_{jh}$$

- **The largest component:**

$|C_i|$=size of the component C_i

$$L = max_i \mid C_i \mid$$

- **The resilience of the network:**

size of the largest connected component after a certain fraction f of nodes or edges have been removed

$$R(f) = \frac{\mid L(f) \mid}{\mid V \mid}$$

- **The stability:** persistence of community structures over time or under perturbations → remove a fraction of edges and compare community partitions using Louvain

Additional variables:

- Global log-return
- The maximum eigenvalues of stock returns computed from their correlation matrix

Dynamical analysis of financial stocks network variables

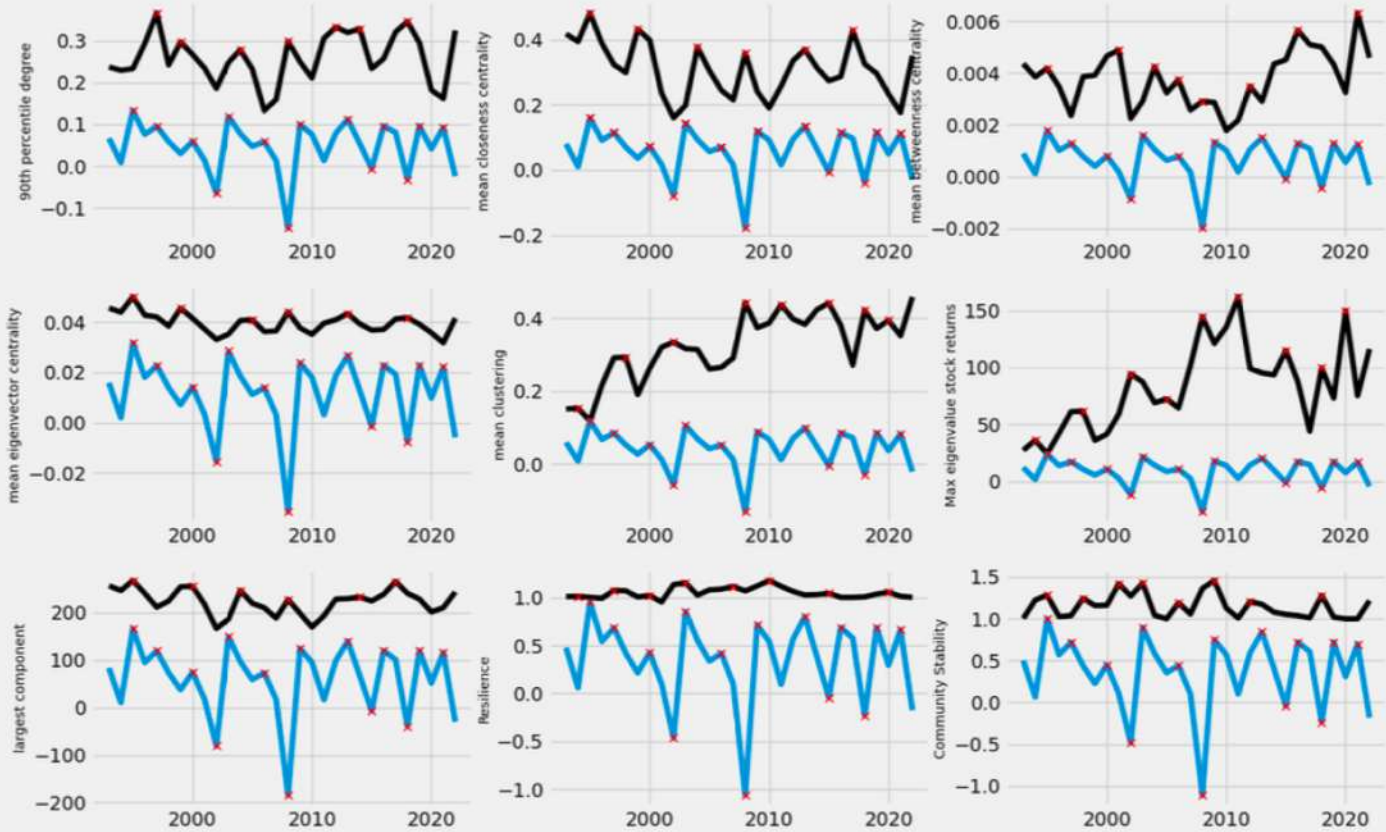
- **Long time period (per year):**

daily stock closing prices from the S&P500 ranging from 1993-01-01 to 2024-01-01 downloaded from Yahoo finance. We clear out stocks that were not present in the entire time range, ending up with 267 stocks and 7805 days of closing values for each stock.

We then split our data into 30 yearly samples covering N_{obs} = 30 years of business days.

- Granger Causality on global log-return:

Mean Closeness Centrality (lag 3,4) & Community Stability (lag 1,2)

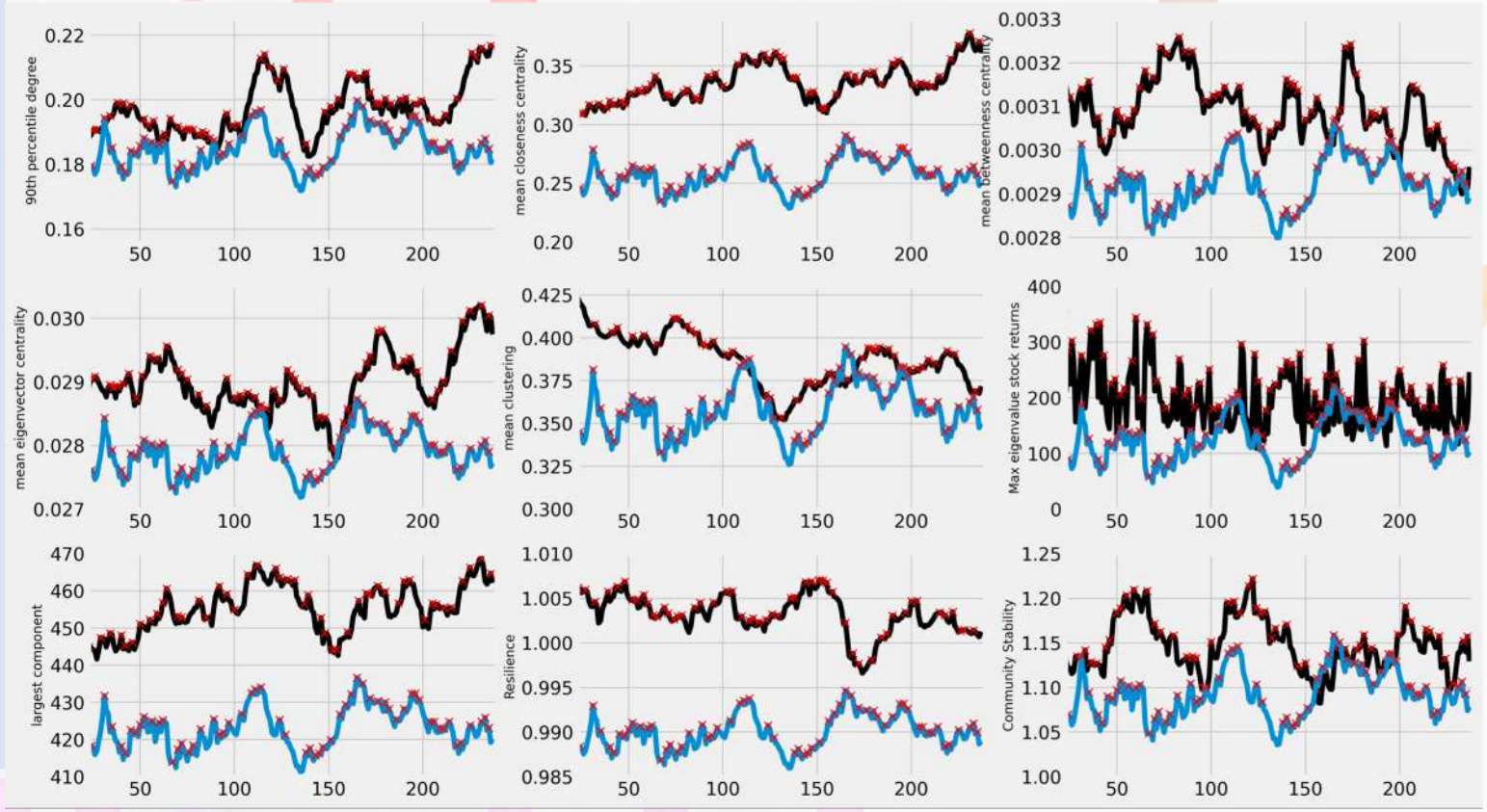


Dynamical analysis of financial stocks network variables

- **Short time period (2days):**

hourly stock prices from the S&P500, from 2022-08-31 to 2024-07-31
downloaded from Yahoo finance. We clear out stocks that were not present in the entire time range → 488 stocks and 3346 hours of recorded price for each stock.

We then split our data in time intervals of 14 hours and we end up with N_{obs} = 238 measures



Dynamical analysis of financial stocks network variables

- **Short time period (2days):**

Variable	Best Lag	SSR F-test p-value
90th Percentile Degree	2	2.16×10^{-10}
Mean Closeness Centrality	2	0.00939
Mean Betweenness Centrality	3	8.67×10^{-8}
Mean Eigenvector Centrality	6	0.00370
Mean Clustering	3	0.00039
Max Eigenvalue Stock Returns	1	4.09×10^{-8}
Community Stability	1	4.50×10^{-5}
Largest Component	2	5.00×10^{-7}
Resilience	3	1.51×10^{-11}

Granger Causality on the global log-return

Can you improve forecasting of individual stock returns using network variables ?

Input Variables:
1-Global variables (averaged over all network e.g. clustering,…)
2-Individual variables (e.g. clustering of each stock within the network, centrality measures…)

Predictions with ML algo:
We use the 15% of stocks ($N_{test} = N_{stocks} * 0.15$) not used for training and predict their log return on the entire time scale: $N_{obs} - N_{lag}$.

For each stock prediction we compute the F2 score and the mean absolute error to build their distribution for each regression model.

The regression models.

1- a Gradient Boosting Regressor model using all selected variables (GBR),

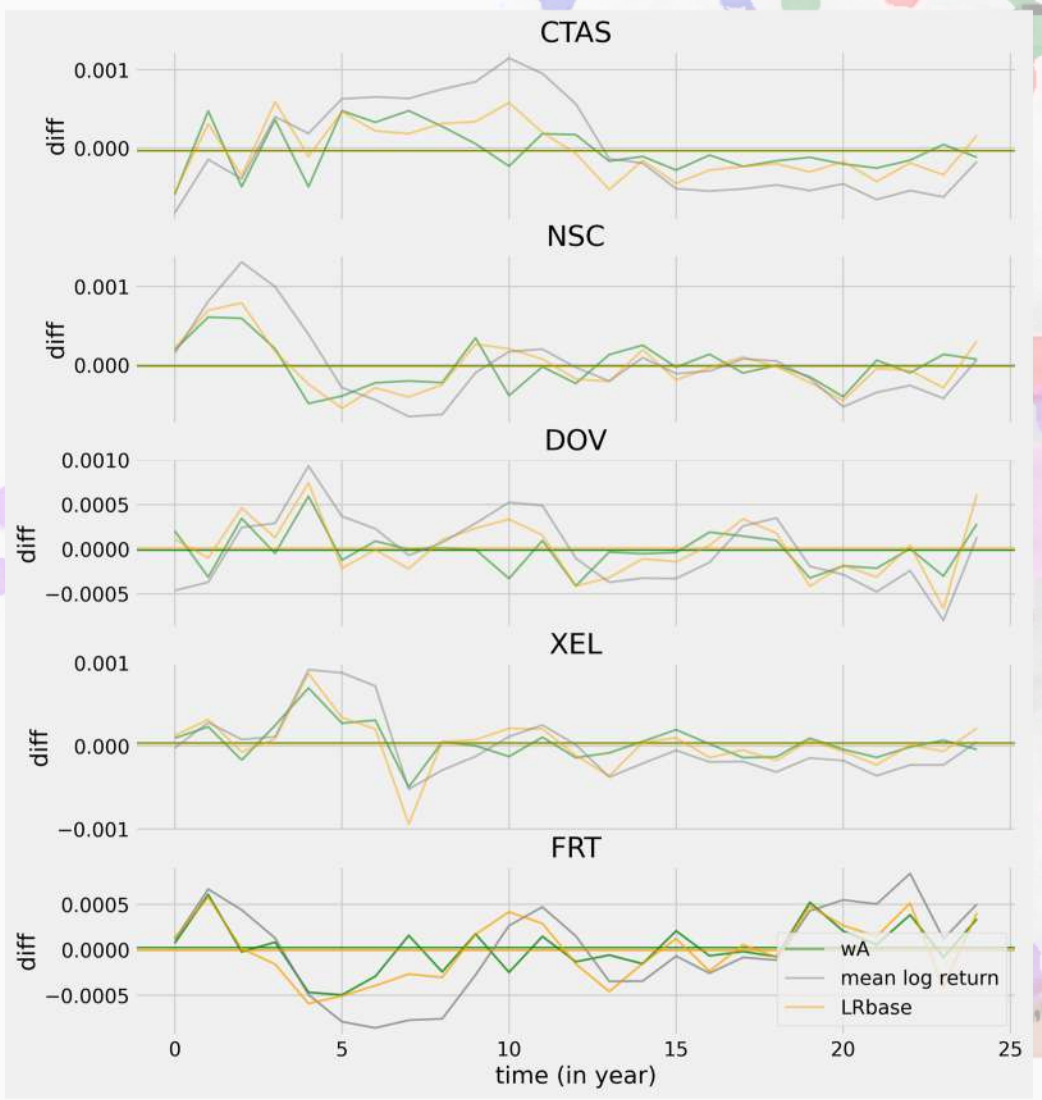
2- a Random Forest Regressor using all selected variables (RFR)

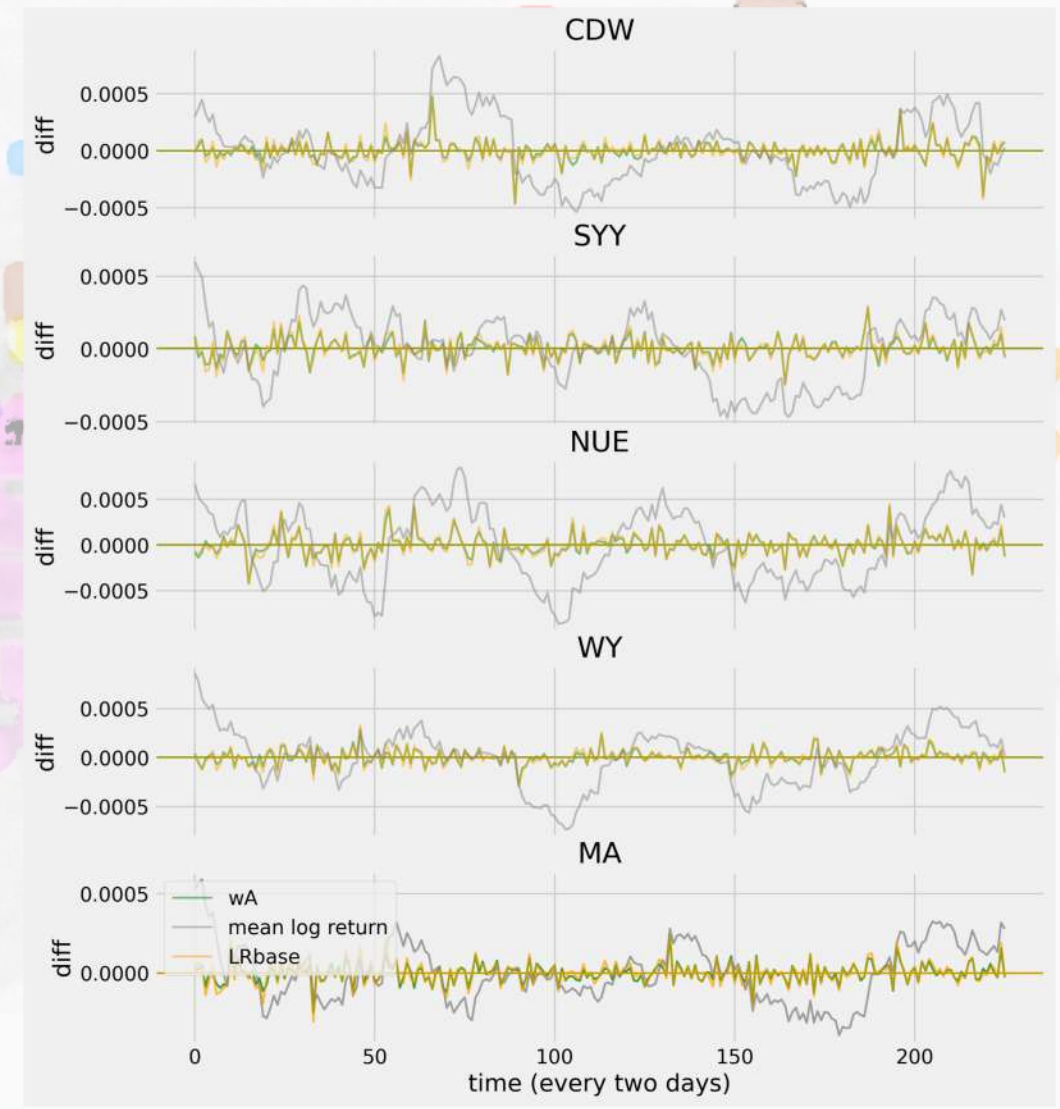
3- a linear regression model using solely the lag 1 of the log return as an input variable (LRbase)

4- a Random Forest Regressor using as input variables of the selected lag of the log return (RFRbase)

5- a weighted average of the first 2 models (wA) where the weight is given by the r2 score.

Results:


CTAS
NSC
DOV
XEL
FRT
diff
time (in year)
wA
mean log return
LRbase


CDW
SYY
NUE
WY
MA
diff
time (every two days)
wA
mean log return
LRbase

Results & Conclusions:

- 2-50% improvement compared to baseline scenario depending on time scale (2 days – 1 year)
- It turns out that global variables are more correlated to the future stock return compared to the network properties of the stock
- Many applications: Forecasting, Risk managment, portfolio optimizations…Can also be applied to any TSA.