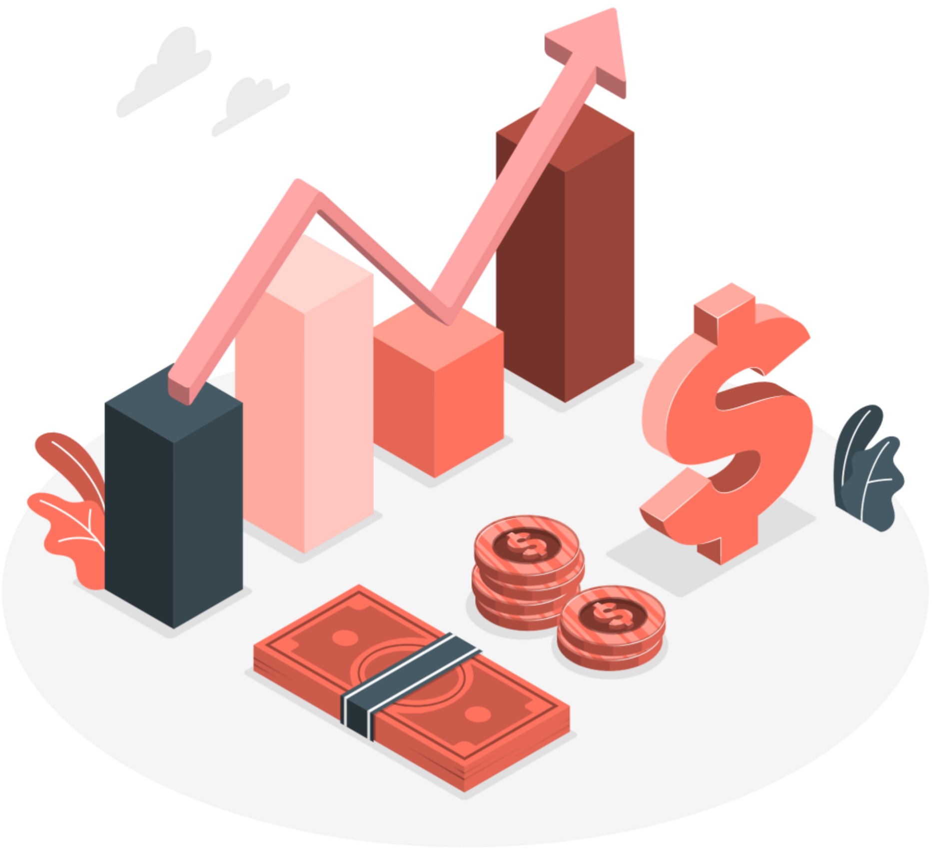


PyData Cluj-Napoca

27th of July, 2024

PREDICTING ROMANIAN STOCK MOVEMENT TRENDS: A COMPLEX NETWORK APPROACH COMBINED WITH MACHINE LEARNING

Author: Orsolya – Dorottya Holgyes

Supervisors: Prof. dr. Camelia Chira and Asist. drd. Bogdan Mursa

Babeș - Bolyai University, Faculty of Mathematics and Computer Science

About me

Orsolya Holgyes

(Orsi)

Bachelor's degree: Business Computer Science @ FSEGA

Master's degree: Data Science for Industry & Society @ CS, UBB

Intern in Financial Audit @ EY
Jan 2021 – Jun 2022

Intern » Junior » Mid Data Engineer @ Endava
finance (stock market) & insurance projects
Jun 2022 – May 2024

Mid Data Engineer @ Vertiv
Device data for Condition Based Maintenance
Jun 2024 – Present

Until now: Learning + Friends & Family
From now on: Friends & Family + Adulthood + Traveling + Cycling & Hiking

From ideas to a dissertation thesis

A desire to have a single topic for all my university projects (including Social Network Analysis), forming in the end my dissertation thesis regarding stock markets.

How to define a stock network?

<u>nodes</u>: different assets on the stock exchanges
(stocks, bonds, etc.)

<u>links</u>: linked if these appear in similar portfolios? if purchased by similar users? If in same index or industry?

available data is only about stock prices and details of the issuers

Research topic #1.1: How to model the Romanian stock market as a network?

Research topic #1.2: analysis of the structure and behavior of the stock network

Source: Janosov (2021)

From ideas to a dissertation thesis

A desire to have a single topic for all my university projects (including Social Network Analysis), forming in the end my dissertation thesis regarding stock markets.

Include somehow machine learning; if potential investors would have a fair assumption on the future market outlook, the invested capital could remain in the Romanian industry.

How to define a stock network?

nodes: different assets on the stock exchanges *(stocks, bonds, etc.)*

links: linked if these appear in similar portfolios? if purchased by similar users? If in same index or industry?

available data is only about stock prices and details of the issuers

Besides the diverse machine learning and deep learning methods applied on the prices or other raw data, lately another research direction is to have as input data properties extracted from the stock network » basically, train algorithms which learn based on network interactions

pl. Milan Janosov (2021). "Network science predicts who dies next in Game of Thrones".

Research topic #1.1: How to model the Romanian stock market as a network?

Research topic #1.2: analysis of the structure and behavior of the stock network

Research topic #2: predict the next day movement of the stock prices based on features extracted from the stock network

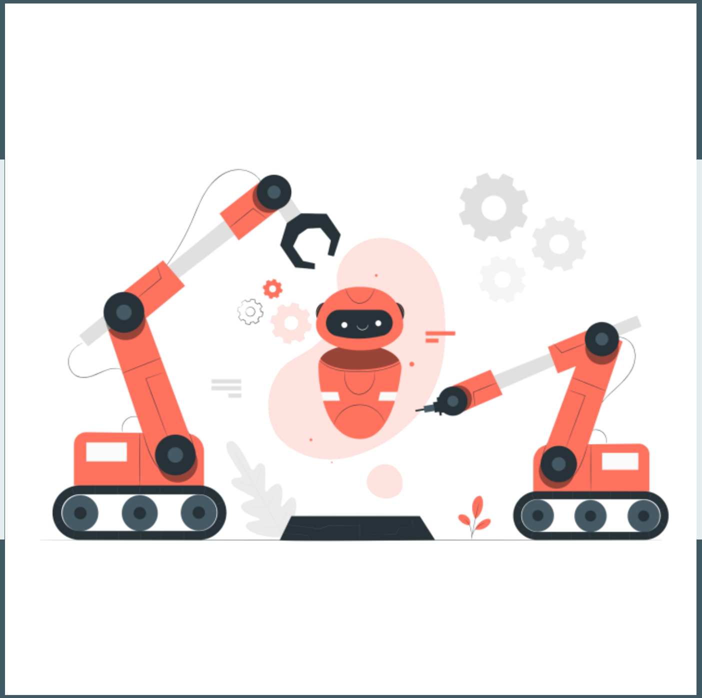
Source: Janosov (2021)

OBJECTIVES & ORIGINAL CONTRIBUTIONS

1. MODELLING THE STOCK NETWORK

2. ANALYZING THE STOCK NETWORK

3. TRAINING A PREDICTION ALGORITHM

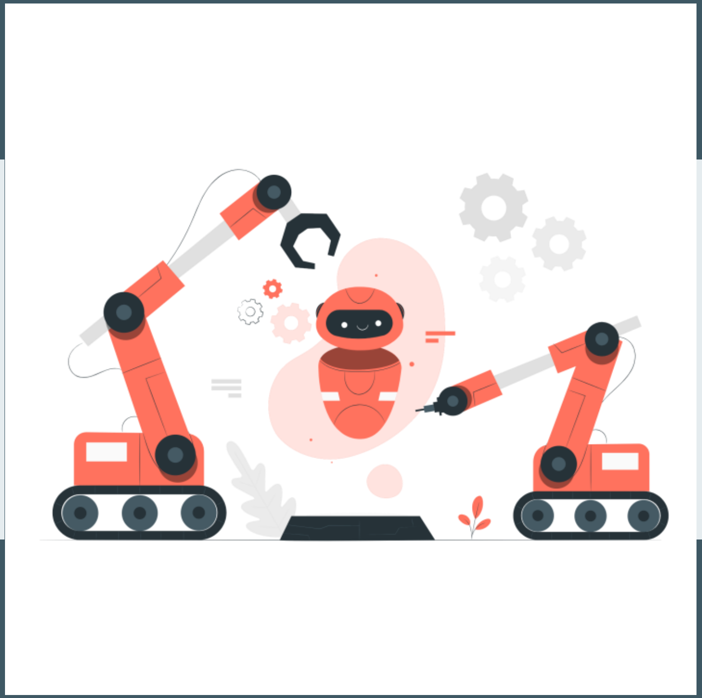


OBJECTIVES & ORIGINAL CONTRIBUTIONS

1. MODELLING THE STOCK NETWORK

2. ANALYZING THE STOCK NETWORK

3. TRAINING A PREDICTION ALGORITHM



How others modelled the stock market

Step 0: choose time range (T) and stocks to include

Step 1: calculate between each stock-pair the Pearson correlation coefficient

$$c_{ij} = \frac{\Sigma_t[(x_i(t) - \overline{x_i}) * (x_j(t) - \overline{x_j})]}{\sqrt{\Sigma_t(x_i(t) - \overline{x_i})^2} * \sqrt{\Sigma_t(x_j(t) - \overline{x_j})^2}}$$

$x_i(t)$ = the close price ($p_{c,i}(t)$) of stock i on day t OR
the trading volumes of stock i ron day t OR
the most frequently: $x_i(t)$ = return = $ln[\frac{p_{c,i}(t)}{p_{c,i}(t-1)}]$

$\overline{x_i}$ = the average of x_i for the whole analyzed period

Step 2: define a threshold value based on that the links of the stock network can be defined:

if the correlation coefficient c_{ij} between two stocks is greater than the threshold, an undirected link is placed between the two nodes of the stocks

paper	data	modeling approach	link creation	N	L
[29]	daily U.S. stock data between July 2007-2008	cross-correlation on the logarithmic daily price return; no lag considered; undirected network	threshold of 0.6	465	n/m
[6]	daily U.S. stock data between July 2005-August 2007 and June 2007 - May 2009	cross-correlation on the daily logarithmic price return, close price and trading volume; no lag considered; undirected network	multiple thresholds: 0.7, 0.8, 0.85, 0.9, 0.95	19,807	2,359 at thd = 0.9
[9]	daily Tehran stock data between 26 March 2011 and 8 April 2017	cross-correlation on the daily logarithmic price return; no lag considered; undirected network	threshold of 0.1	142	3,457
[10]	daily stock data from Shanghai and Shenzhen between 2003 and 2007	cross-correlation on the logarithmic daily price return; no lag considered; undirected network	experimented with multiple thresholds	1,080	n/m
[18]	daily Tehran stock data between 2014 and 2017	cross-correlation on the logarithmic daily price return; no lag considered; undirected network	threshold of 0.4 selected with an objective function	246	n/m
[11]	daily stock indices data from 70 countries between 2017-2019 and 2020-2022	novel methodology based on open-high-low-close data; no lag considered; undirected network		70	n/m
[3]	S&P500, NASDAQ, DJIA stock data from 2000-2014	price volatility patterns based on cross-correlation; 5-day lag considered; directed and weighted network		47	n/m

Table 3.1: Summary of other researchers' stock network modeling approaches. N refers to the number of nodes and L to the number of edges in the stock network which was modeled in the specific papers.

What value should the threshold take? How can we determine this threshold?

arbitrary value

based on experiments

Source: Chi et al. (2010); Sun et al. (2015); Heshmati and Rahimnezhad (2021) ; Huang et al. (2009, 2024) ; Mogdaham et al. (2019) ; Cao et al. (2019)

The approach in this thesis

Data: ethically sourced close prices of stocks from the Bucharest Stock Exchange between 2022-2023

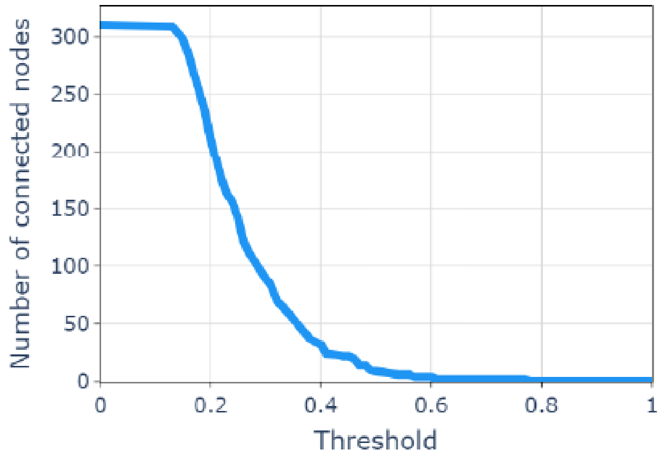
Nodes of the stock network: the 328 stocks on the Romanian stock market, from those 18 had constant close prices and are isolated nodes

Links of the stock network: based on the cross-correlation of the stocks' returns between 2022-2023

Threshold: for each value between 0,00 and 1,00 a different stock network was modelled. The optimal threshold was chosen based on network-level metrics which reflect the structure of the network

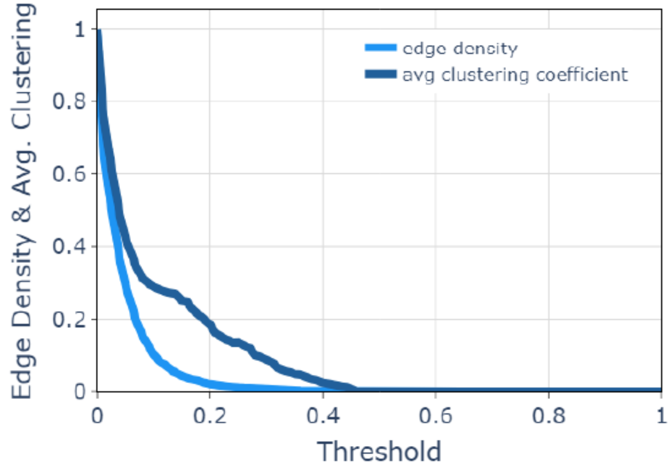
The threshold must be in the interval of [0, 0.20] so that valuable data and insights can be retrieved from the network, based on that algorithms can truly learn the stock interactions.

Figure 4.1



Number of non-isolated nodes based on the threshold » as the threshold increases, the number of non-isolated nodes decreases

Figure 4.2

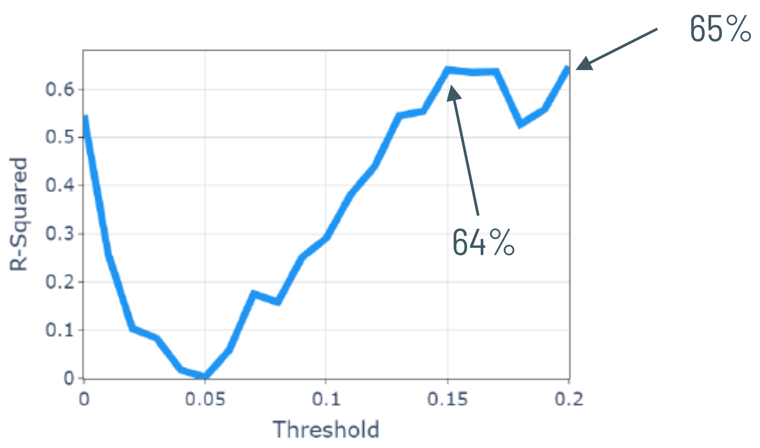


The density and the average clustering coefficient of the network » after the threshold value 0.25, the number of edges are only 1% of all possible edges; the average clustering coefficient is higher than density, therefore the dense relations in the network were expected

The approach in this thesis

Depending on the extent of the scale-free property: for each network built based on thresholds in the [0, 0.20] interval, a linear regression model was fitted on the degree distribution of each network and the power-law distribution, both on log-log scale» we should choose the threshold with the highest R^2

$$P(k) = \alpha * k^{-\gamma} \overset{\log}{\Longleftrightarrow} logP(k) = log\alpha + (-\gamma) * logk$$



Threshold = 0.20

Number of connected nodes = 211
Number of edges = 1005

Threshold = 0.15

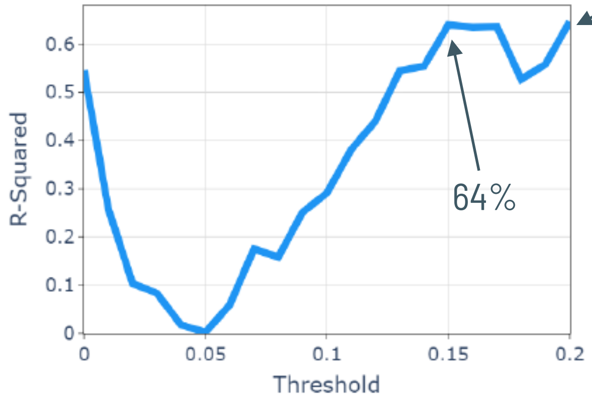
Number of connected nodes = 298
Number of edges = 2089

The approach in this thesis

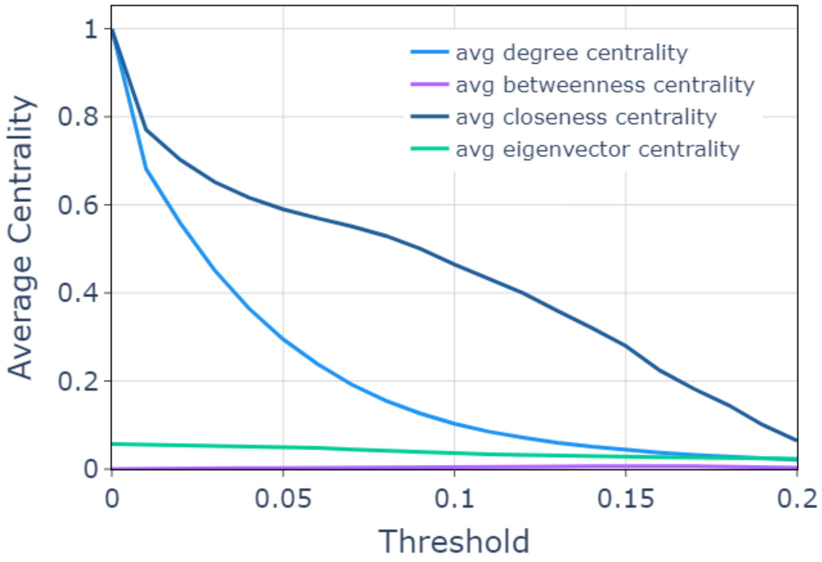
Depending on the extent of the scale-free property: for each network built based on thresholds in the [0, 0.20] interval, a linear regression model was fitted on the degree distribution of each network and the power-law distribution, both on log-log scale» we should choose the threshold with the highest R^2

$$P(k) = \alpha * k^{-\gamma} \overset{\log}{\Longleftrightarrow} logP(k) = log\alpha + (-\gamma) * logk$$

65%



Average the node-level centralities



Threshold = 0.20

Number of connected nodes = 211
Number of edges = 1005

Threshold = 0.15

Number of connected nodes = 298
Number of edges = 2089

Centralities are higher for threshold 0.15, especially the closeness centrality which means that information flows are more efficient in this network.

Source: Pósfai és Barabási (2016)

The modelled network

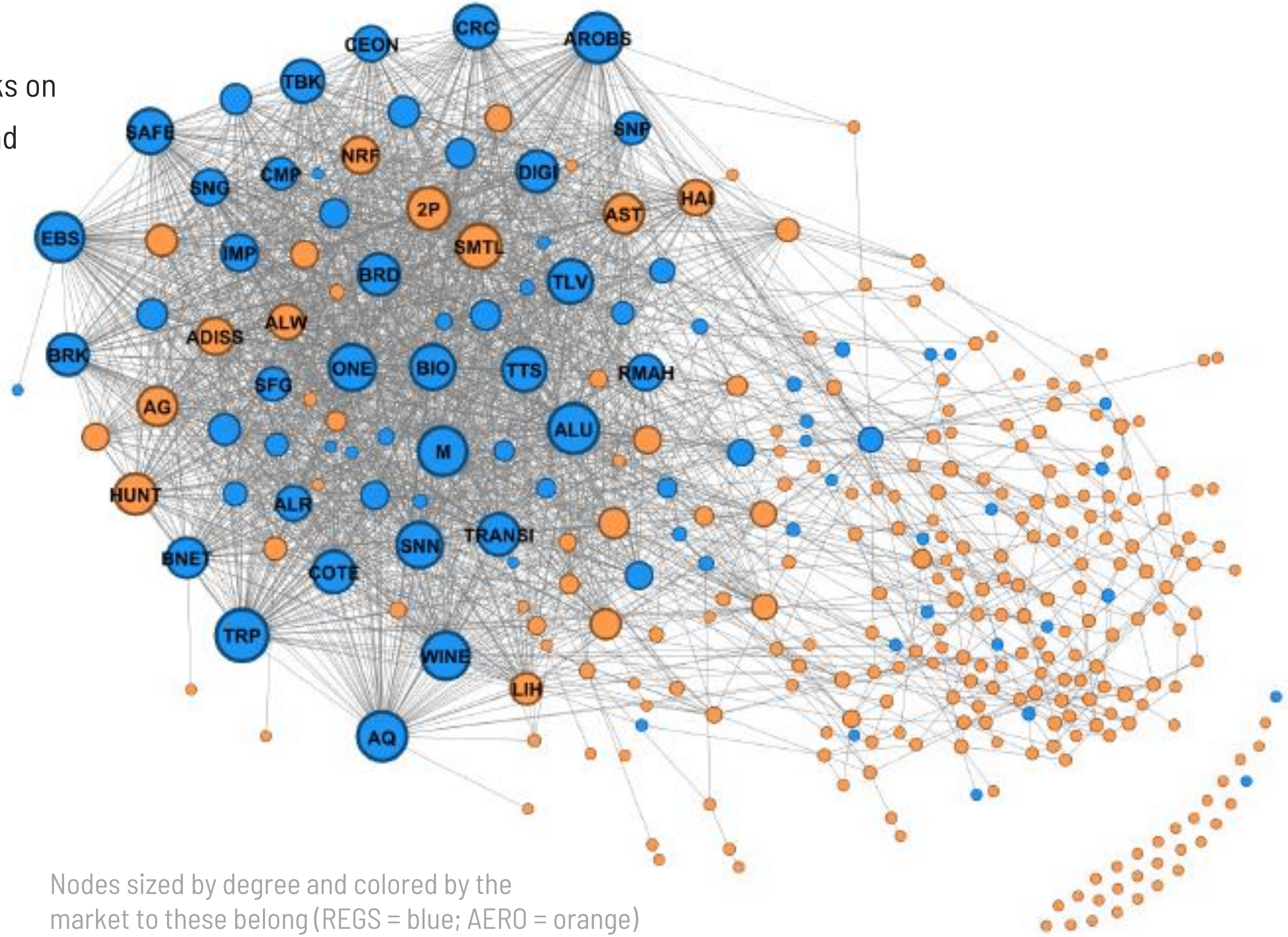
Nodes of the stock network: the 328 stocks on the Romanian stock market, from those 18 had constant close prices and are isolated nodes

Links of the stock network: based on the cross-correlation of the stocks' returns between 2022-2023; if it is higher than 0.15

N = 360

L = 2089

Nr of isolated nodes = 20



Nodes sized by degree and colored by the market to these belong (REGS = blue; AERO = orange)

OBJECTIVES & ORIGINAL CONTRIBUTIONS

1. MODELLING THE STOCK NETWORK

» based on the cross-correlation between the stock returns

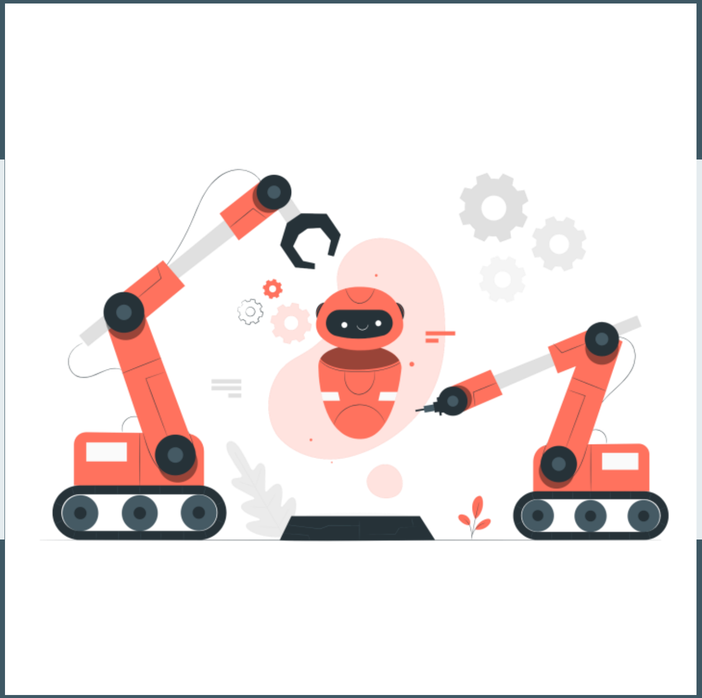
» the nodes are the 328 stocks from the Romanian stock market

» the value of the threshold used at edge construction was defined to be 0.15, based on experiments

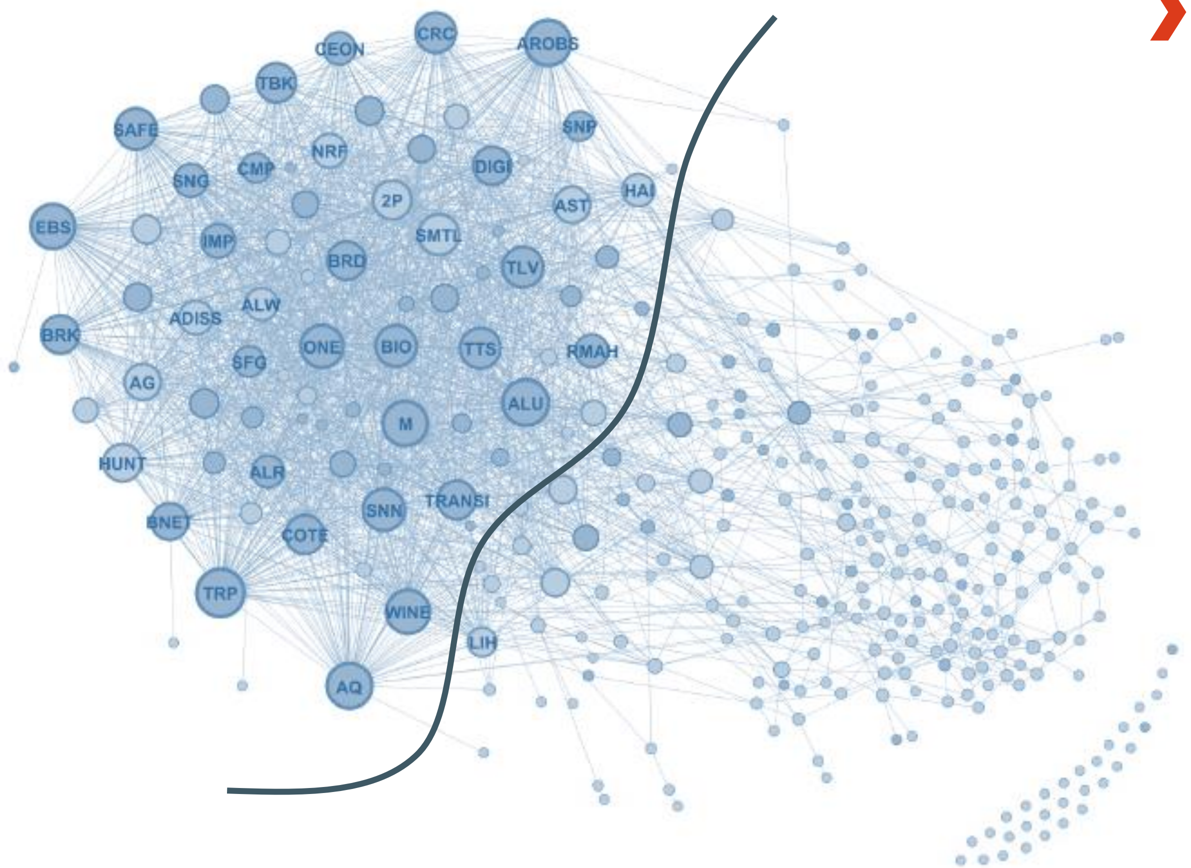
» the threshold significantly influences the structure of the stock network

2. ANALYZING THE STOCK NETWORK

3. TRAINING A PREDICTION ALGORITHM

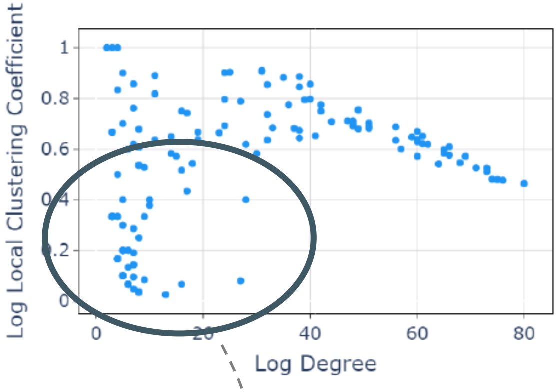


Romanian stock market ~ network perspective



- Two structurally different components: a wider, denser component which tails into a much smaller sparse component

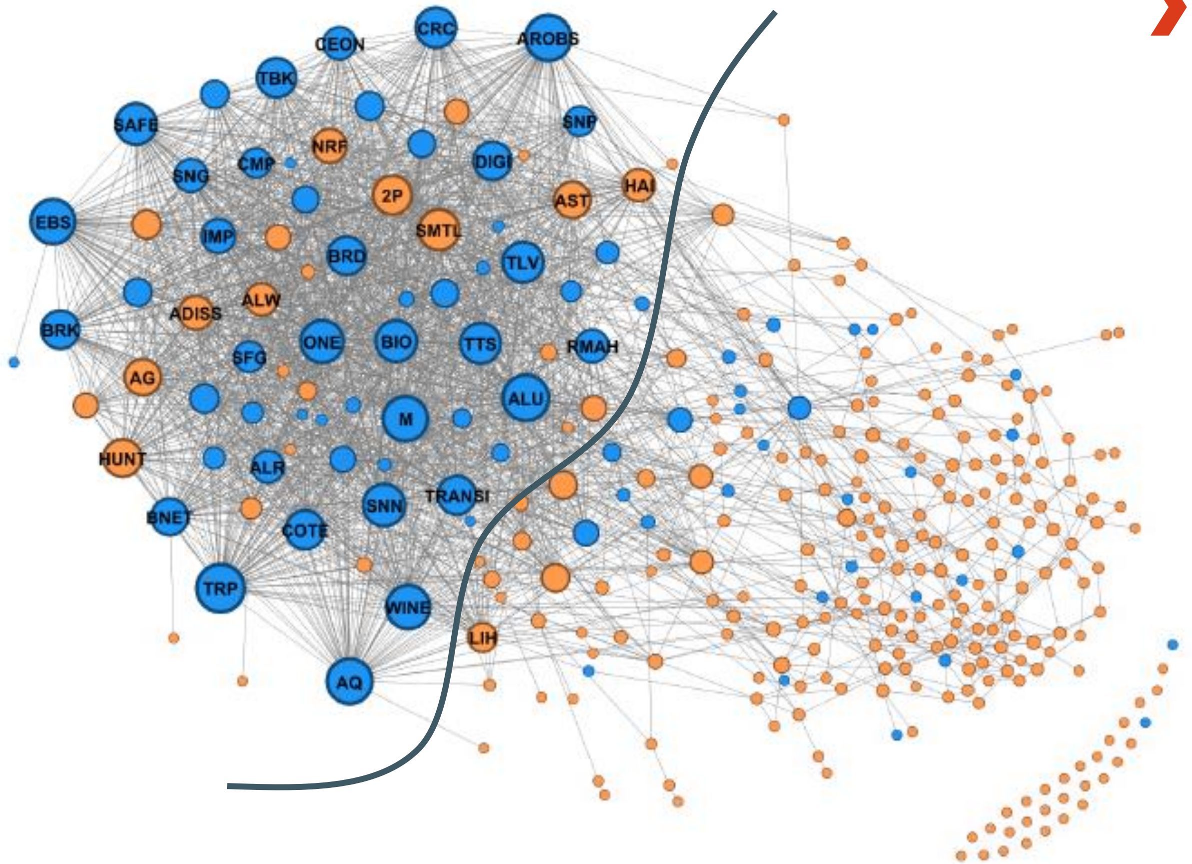
  - Expected, since edge density (0.04) < avg clustering coeff(0.24)
  - Other researchers from the literature had the same result of the average clustering coefficient being higher than the edge density.



Unexpectedly many nodes with small degree and small clustering coefficient » these influence the average value of the clustering coefficient

Source: bvb.ro; Huang et al. (2009)

Romanian stock market ~ network perspective



stocks listed on the main market (REGS)

stocks listed on the secondary market (AERO)

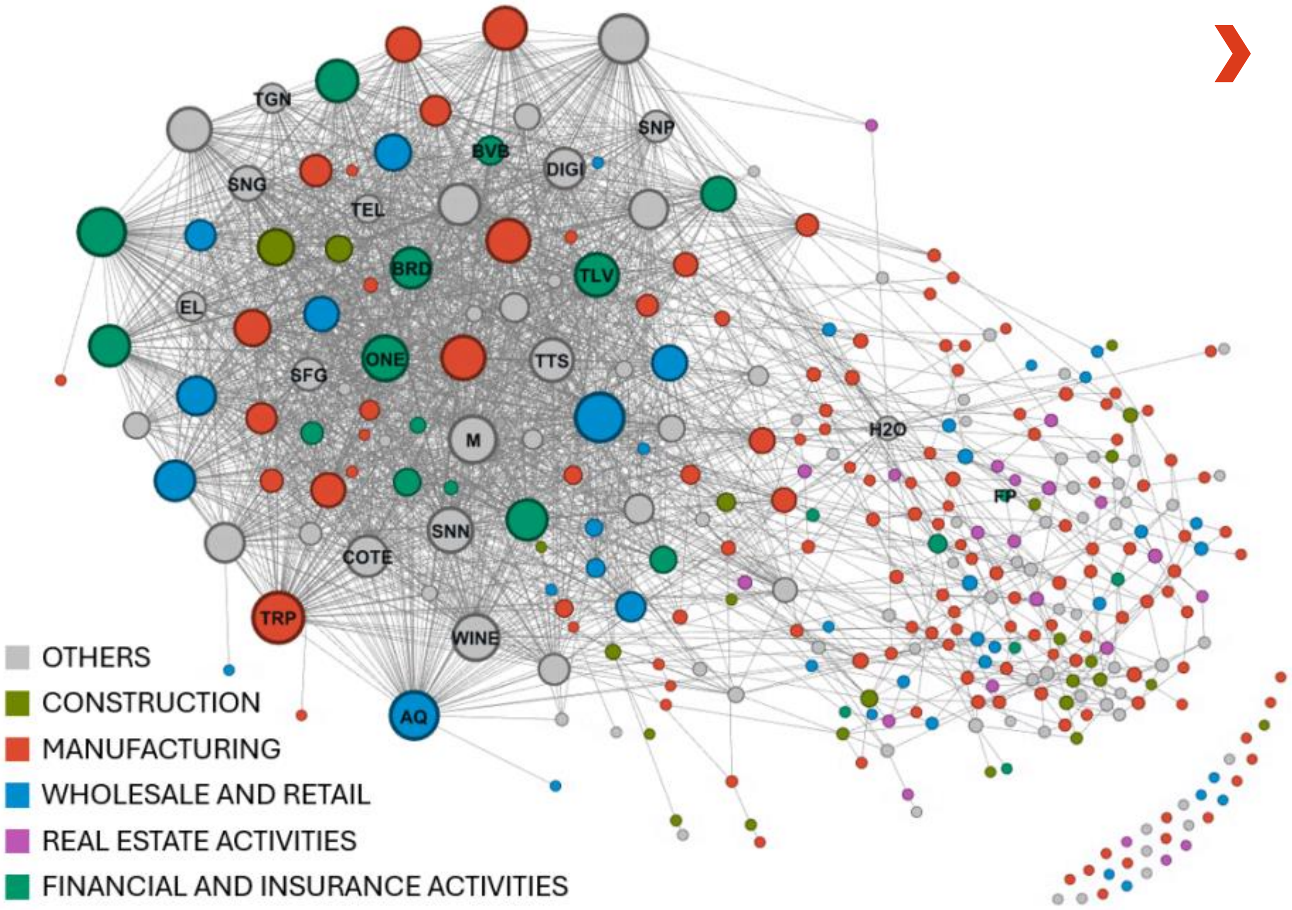
Two structurally different components: visually looked like the dense component mainly consists of stocks listed on REGS and the other component is dominated by AERO stocks

The network is actually homogeneous regarding the distribution of nodes based on market:
» 41% of edges are linking together REGS and AERO nodes
» the cumulative degree of AERO-nodes is 1719 (41%), while the REGS nodes have in total 2459 degrees, even though there are fewer REGS nodes than AERO

Can be explained through the investing behavior on the Romanian market: investors balance out the riskier AERO stocks with the mature and secure REGS stocks in their portfolios. Secondary market driven by the Primary market.

Source: bvb.ro

Romanian stock market ~ network perspective



The nodes of the network are labeled if the stock that these represent belongs to the BET index. There are also low-degree nodes belonging to BET.

The nodes are not clustered by industry, even though this was the expectation: the prices of assets in the same industry tend to move together, hence the correlation should be higher.

The network is homogeneous also based on the clustering on industry, the nodes connecting also with nodes from other industries:
» 40% of stocks activate in the manufacturing industry but this does not represent the industry with the highest cumulative degrees
» the industry with the highest cumulative degrees is finance and insurance services to that belong only 21 nodes

The number of shares belonging to an industry doesn't mean consequently that the industry drives the stock market. This does not depend on the index either, even though indices represent benchmark portfolios, and it was expected that these are popular among investors.

Source: bvb.ro; Chi et al. (2010); Sun et al. (2015); Heshmati and Rahimnezhad (2021)

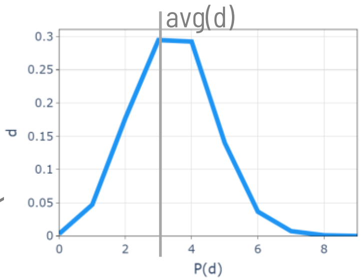
Romanian stock market ~ network perspective

- Fast information-flows: all price changes are quickly incorporated and reflected by all nodes » dynamic system, free economy

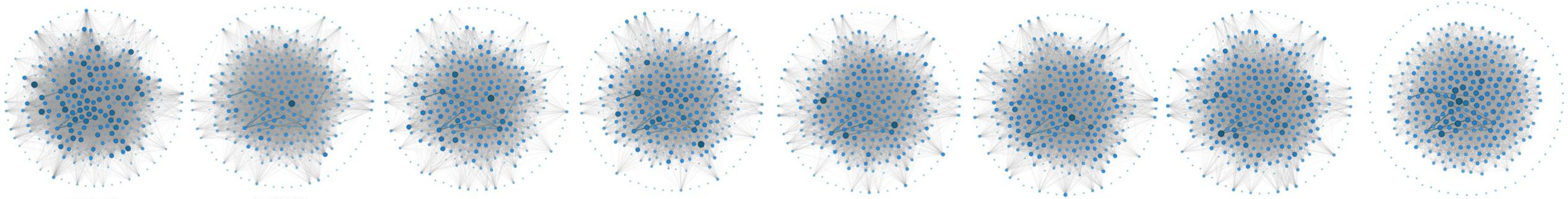
temporal networks: aggregating the data from the 2022-2023 time-period on sub-periods » the smaller the period, the more connections there were » stock prices and returns closer in time have higher correlation

type	metric	N	N_isolated	L	edge_density	avg_clust_coeff	k_min	k_max	k_avg
year	mean	328.0	28.500000	3321.500000	0.061936	0.214279	0.0	73.000000	20.253049
halfyear	mean	328.0	43.500000	5264.250000	0.098162	0.220671	0.0	83.000000	32.099085
trimester	mean	328.0	56.000000	6645.000000	0.123909	0.246695	0.0	105.166667	40.518293
quarter	mean	328.0	67.500000	7508.750000	0.140015	0.267713	0.0	104.250000	45.785061
month	mean	328.0	110.166667	9917.958333	0.184940	0.353211	0.0	140.750000	60.475356

diameter = 9
average shortest path = 3.44
60% of nodes achieve in 4-5 steps (i.e. connections) other nodes
small-world property



The network dynamics as captured by quarterly aggregated temporal networks



(a) 2022Q1 (b) 2022Q2 (c) 2022Q3 (d) 2022Q4 (a) 2023Q1 (b) 2023Q2 (c) 2023Q3 (d) 2023Q4

36 constant connections in all 8 networks

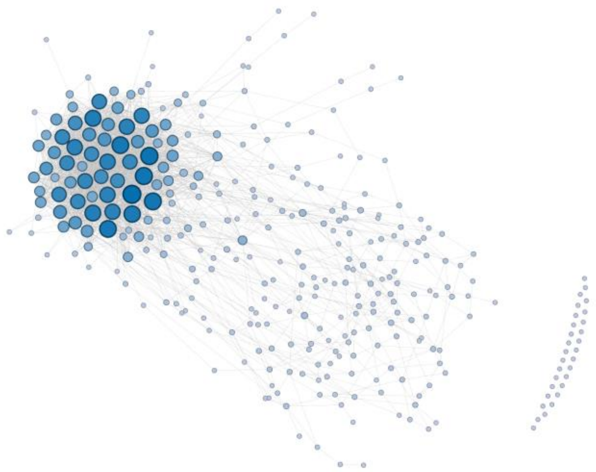
Source: bvb.ro; Chi et al. (2010); Sun et al. (2015); Heshmati and Rahimnezhad (2021); Huang et al. (2009)

Romanian stock market ~ node perspective

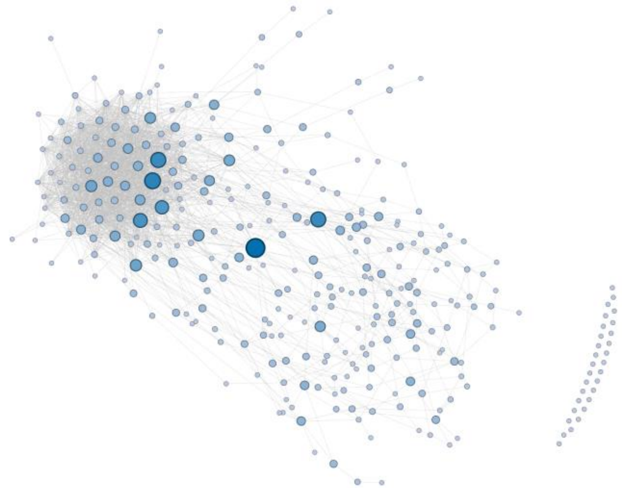
- The most important nodes: based on the four centralities. The price of the most stocks are linked to TRP (TERAPLAST SA; AERO), ALU (ALUMIL ROM INDUSTRY S.A.; REGS), AROBS (AROBS TRANSILVANIA SOFTWARE)

rank	degree centrality	betweenness centrality	closeness centrality	eigenvector centrality
1	TRP (0.24)	H2O (0.06)	TRP (0.4)	TRP (0.17)
2	ALU (0.23)	AROBS (0.05)	ALU (0.39)	EBS (0.16)
3	AROBS (0.23)	ALU (0.05)	M (0.39)	ALU (0.16)
4	AQ (0.23)	MIB (0.05)	AQ (0.39)	M (0.16)
5	EBS (0.22)	WINE (0.04)	AROBS (0.39)	ONE (0.16)
6	M (0.22)	AQ (0.04)	WINE (0.39)	AROBS (0.16)
7	WINE (0.22)	TBM (0.03)	EBS (0.38)	WINE (0.16)
8	ONE (0.21)	TBK (0.03)	ONE (0.38)	SAFE (0.16)
9	SNN (0.21)	COKJ (0.03)	CRC (0.38)	AQ (0.16)
10	SAFE (0.2)	MEOR (0.03)	SMTL (0.38)	SNN (0.15)

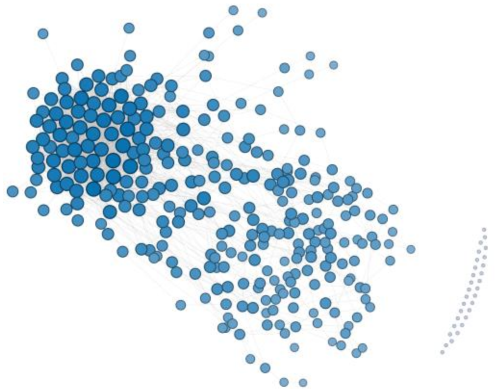
Degree Centrality



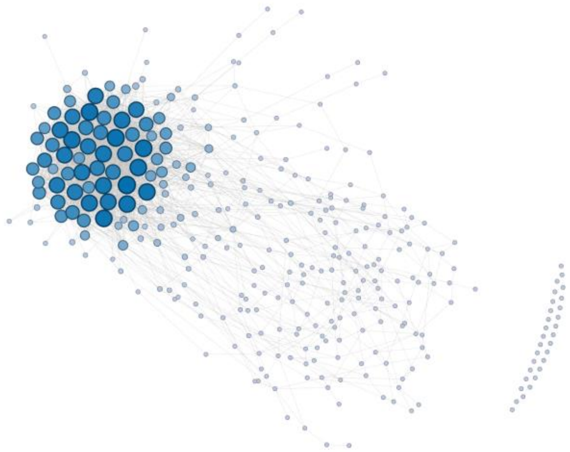
Betweenness Centrality



Closeness Centrality



Eigenvector Centrality



Source: bvb.ro; Chi et al. (2010); Sun et al. (2015); Heshmati and Rahimnezhad (2021)

OBJECTIVES & ORIGINAL CONTRIBUTIONS

1. MODELLING THE STOCK NETWORK

» based on the cross-correlation between the stock returns

» the nodes are the 328 stocks from the Romanian stock market

» the value of the threshold used at edge construction was defined to be 0.15 based on experiments

» the threshold significantly influences the structure of the stock network

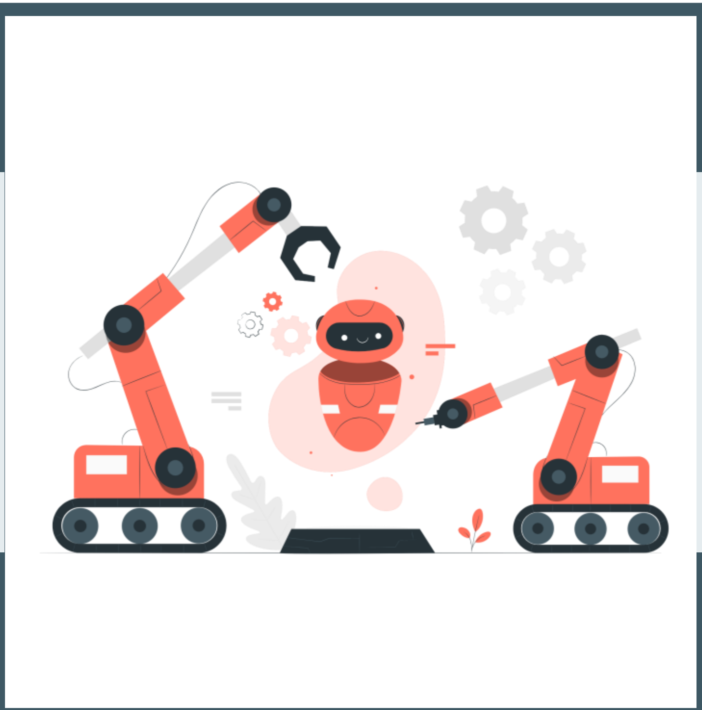
2. ANALYZING THE STOCK NETWORK

» scale-free network

» visually there are two components outlined, but the network is homogeneous by market, index or industry

» the information flows are quick and efficient, therefore the Romanian stock market is a dynamic system, like a free-flow economy

3. TRAINING A PREDICTION ALGORITHM



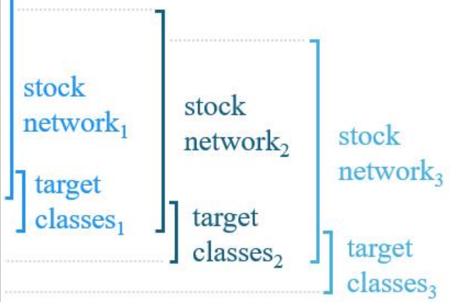
From close prices to prediction

Methodology: quarterly close prices » networks »
5 extracted properties = features of the prediction algorithm

Extract of raw close prices

	dt	AAG	ALT	AQ	AROBS	...	PRDI	TRVC	SIRJ	AUXI
0	2022-01-03	3.60	0.0539	0.9233	0.975	...	0.125	6.0	1.38	22.0
1	2022-01-04	3.56	0.0539	0.9233	1.010	...	0.125	6.0	1.38	22.0
2	2022-01-05	3.58	0.0539	0.9200	1.040	...	0.125	6.0	1.38	22.0
3	2022-01-06	3.54	0.0579	0.9167	1.040	...	0.125	6.0	1.38	22.0
4	2022-01-07	3.58	0.0539	0.9167	1.090	...	0.125	6.0	1.38	22.0
...	...	...	...	...	...	...	...	...	...	...
59	2022-03-25	3.52	0.0480	0.8700	0.992	...	0.125	6.0	1.38	22.0
60	2022-03-28	3.56	0.0480	0.8580	0.993	...	0.125	6.0	1.38	22.0
61	2022-03-29	3.50	0.0480	0.8500	1.025	...	0.125	6.0	1.38	22.0
62	2022-03-30	3.50	0.0525	0.8500	1.010	...	0.125	6.0	1.38	22.0
63	2022-03-31	3.46	0.0545	0.8500	1.015	...	0.125	6.0	1.38	22.0

Stock networks built based on data from the period of the sliding window, containing T observations



Extract from each network five metrics on node-level and associate each observation-set with the target class of the node

	clustering_coeff	degree_centr	closeness_centr	betweenness_centr	eigenvector_centr	target_class	stock
0	0.595426	0.492248	0.663239	0.004778	0.078745	increase	AAG
1	0.519380	0.500000	0.666667	0.008708	0.074586	stagnate	ALT
2	0.684020	0.577519	0.702997	0.002290	0.099315	increase	AQ
3	0.616003	0.600775	0.714681	0.005485	0.098044	decrease	AROBS
4	0.575313	0.426357	0.635468	0.005369	0.066308	increase	ARTE
...	...	...	...	...	...	...	...
2592	0.670513	0.154440	0.541841	0.000499	0.025700	stagnate	IPRO
2593	0.596703	0.494208	0.664103	0.004947	0.080254	decrease	BIOW
2594	0.422222	0.138996	0.537344	0.000722	0.016172	stagnate	NORD
2595	0.753737	0.386100	0.619617	0.001080	0.070019	stagnate	EPN
2596	0.642715	0.444015	0.642680	0.004415	0.074595	stagnate	TSLA

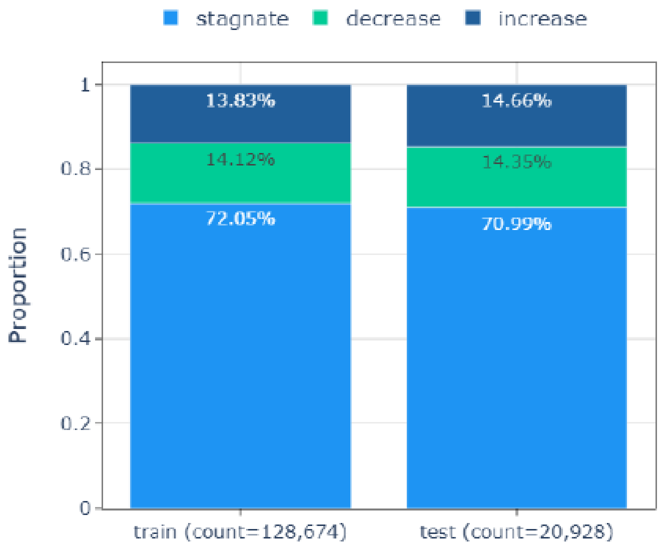
input variables of the prediction model

target variable of the prediction model

T = 60 trading days ~ a quarter

460 stock networks
149,602 observations
Target class distribution: stagnate = 72%, increase = 14%, decrease = 14%

train dataset ≤ 2023Q4 (86%)
test dataset = 2023Q4 (14%)



multi-class classification problem; unbalanced classes, but this is the natural distribution. 30% of the stagnate class because of isolated nodes.

Source: Pósfai és Barabási (2016)

The multi-class classification problem

Goal

Classify in increase, decrease or stagnate classes the movement of the next-day close prices of the stocks on the Bucharest Stock Exchange

Models

Multinomial logistic regression » for each class there is a two-class logistic regression built: one class vs. all the others. The results are combined based on the SoftMax function.

```
LogisticRegression(multi_class='ovr')
```

Decision Tree » at each level there is a split made based on the feature that best divides the data into two different classes.

```
DecisionTreeClassifier()
```

SVM » searching for the optimal hyperplane which best divides the data on the basis of one class vs others.

```
LinearSVC(dual=False)
```

Performance metrics:

accuracy » the ratio of the correctly predicted classes from all observations

unbalanced classes, accuracy can be biased, therefore precision, recall and F1 score are equally important

precision » how many predictions were correct from all the cases when a certain class was predicted

recall » measures how many of the observations in a certain class were predicted correctly

F1 score » the harmonic mean of precision and recall

Source: Pedregosa et al. (2011)

The multi-class classification problem

Goal

Classify in increase, decrease or stagnate classes the movement of the next-day close prices of the stocks on the Bucharest Stock Exchange

Models

Multinomial logistic regression » for each class there is a two-class logistic regression built: one class vs. all the others. The results are combined based on the SoftMax function.

```
LogisticRegression(multi_class='ovr')
```

Decision Tree » at each level there is a split made based on the feature that best divides the data into two different classes.

```
DecisionTreeClassifier()
```

SVM » searching for the optimal hyperplane which best divides the data on the basis of one class vs others.

```
LinearSVC(dual=False)
```

Performance metrics:

accuracy » the ratio of the correctly predicted classes from all observations

unbalanced classes, accuracy can be biased, therefore precision, recall and F1 score are equally important

precision » how many predictions were correct from all the cases when a certain class was predicted

recall » measures how many of the observations in a certain class were predicted correctly

F1 score » the harmonic mean of precision and recall

metrics on class level

model-level metrics = the simple mean of the class-level metrics

Source: Pedregosa et al. (2011)

The learning algorithms

- Baseline model: predict for each observation the *stagnate* class

$$f_{baseline}(x) = \text{"}stagnate\text{"}$$

dataset	indicator	increase	decrease	stagnate	model
train	accuracy	0.7205			
	precision	0	0	0.7205	0.2402
	recall	0	0	1.000	0.3333
	F1 score	0	0	0.8375	0.2792
test	accuracy	0.7099			
	precision	0	0	0.7099	0.2366
	recall	0	0	1.000	0.3333
	F1 score	0	0	0.8303	0.2767

- Machine learning models: from each algorithm-family a single model was trained using the sklearn library

dataset	indicator	LogisticRegression	DecisionTree	LinearSVC
train	accuracy	0.7317	0.9996	0.7310
	precision	0.5009	0.9992	0.4800
	recall	0.4151	0.9993	0.3984
	F1-score	0.4053	0.9992	0.3788
test	accuracy	0.7127	0.6236	0.7149
	precision	0.4835	0.4414	0.4430
	recall	0.4257	0.4519	0.3945
	F1-score	0.4251	0.4450	0.3719

» even though the Decision Tree had the smallest accuracy, even lower than the baseline model (because of the overfit on train dataset), this had the highest recall and F1-score.

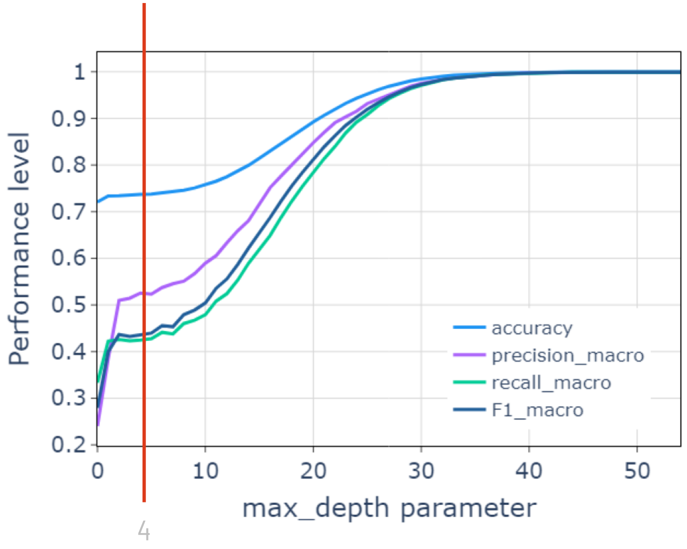
- fine tune this model to reduce the overfitting

Forrás: Pedregosa et al. (2011)

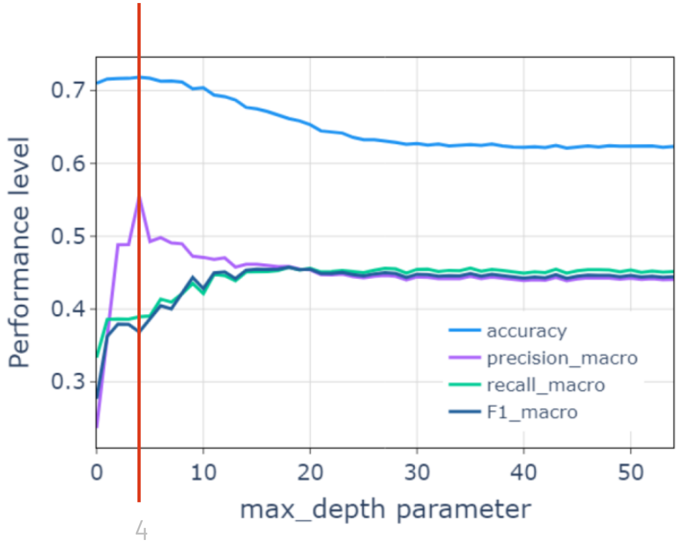
Choosing the best decision tree

The decision tree had a depth of 55: training decision trees with different depths from 1 up to 55, searching for a balance among the 4 performance metrics» decision tree with depth 4

indicator	Baseline	$DecisionTree_{55}$	$DecisionTree_4$
accuracy	0.7205	0.6236	0.7184
precision	0.2402	0.4414	0.3686
recall	0.3333	0.4519	0.5544
F1-score	0.2792	0.4450	0.3894



(a) train dataset



(b) test dataset

» after 10 splits the overfit is evidenced: increasing performance on the train dataset while these decrease on the test dataset
» the precision is the highest at the decision tree with depth of 4 splits

» tried to increase the performance by taking the different combination of features, but it did not help.

The model corresponds to the goal of this thesis

Validation on data from 2024Q1

1308 observations
85% stagnate, 8% increase, 7% decrease

accuracy = 0.8257
precision = 0.3917
recall = 0.3628
F1 score = 0.3584

CONCLUSIONS

1. MODELLING THE STOCK NETWORK

» based on the cross-correlation between the stock returns

» the nodes are the 328 stocks from the Romanian stock market

» the value of the threshold used at edge construction was defined to be 0.15 based on experiments

» the threshold significantly influences the structure of the stock network

2. ANALYZING THE STOCK NETWORK

» scale-free network

» visually there are two components outlined, but the network is homogeneous by market, index or industry

» the information flows are quick and efficient, therefore the Romanian stock market is a dynamic system, like a free-flow economy

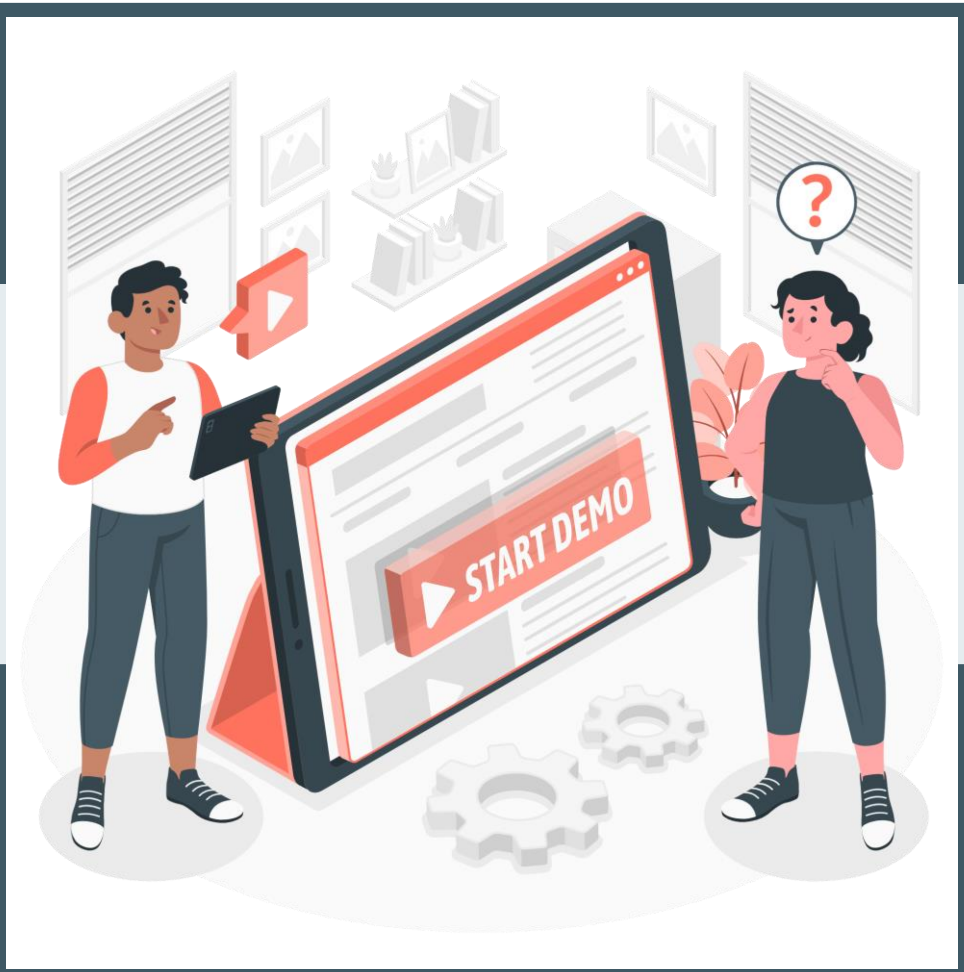
3. TRAINING A PREDICTION ALGORITHM

» multi-class supervised classification problem

» the features are 5 metrics extracted from 460 stock network

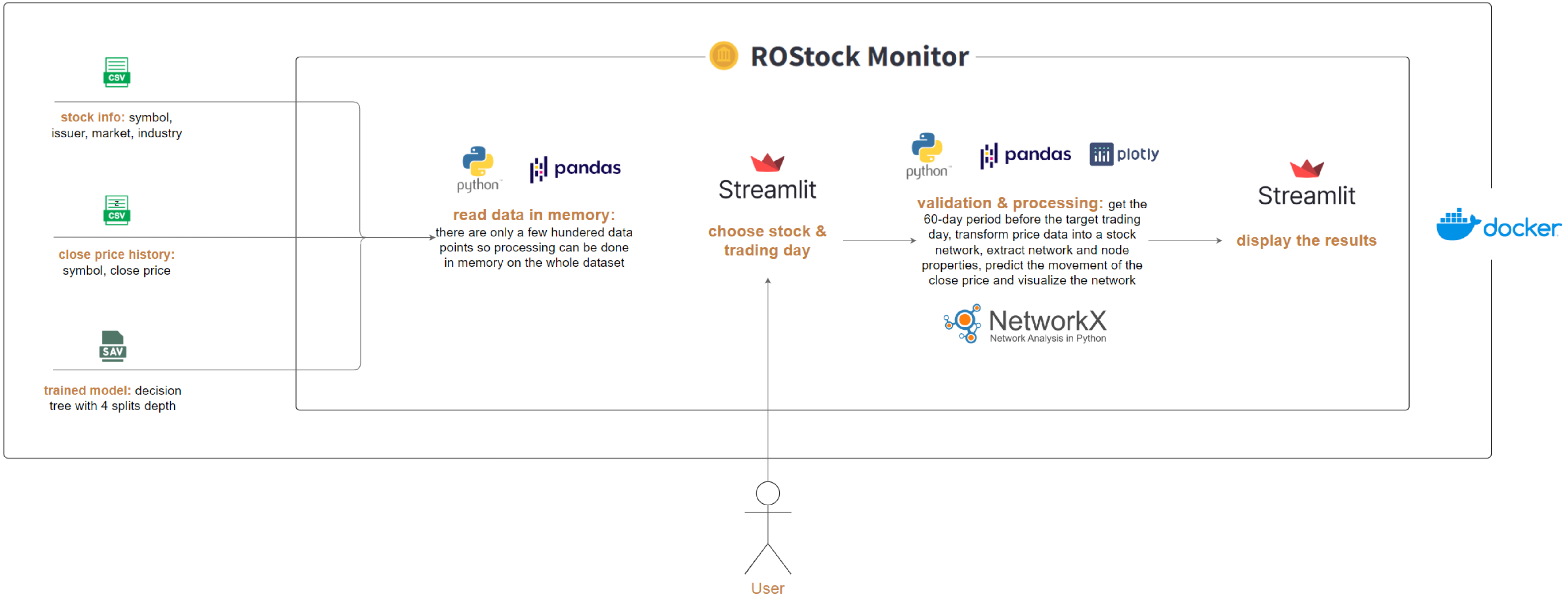
» unbalanced classes

» logistic regression, decision trees, SVM. Final model: decision tree with 4 splits depth, having accuracy of 72% and F1 score of 39%.

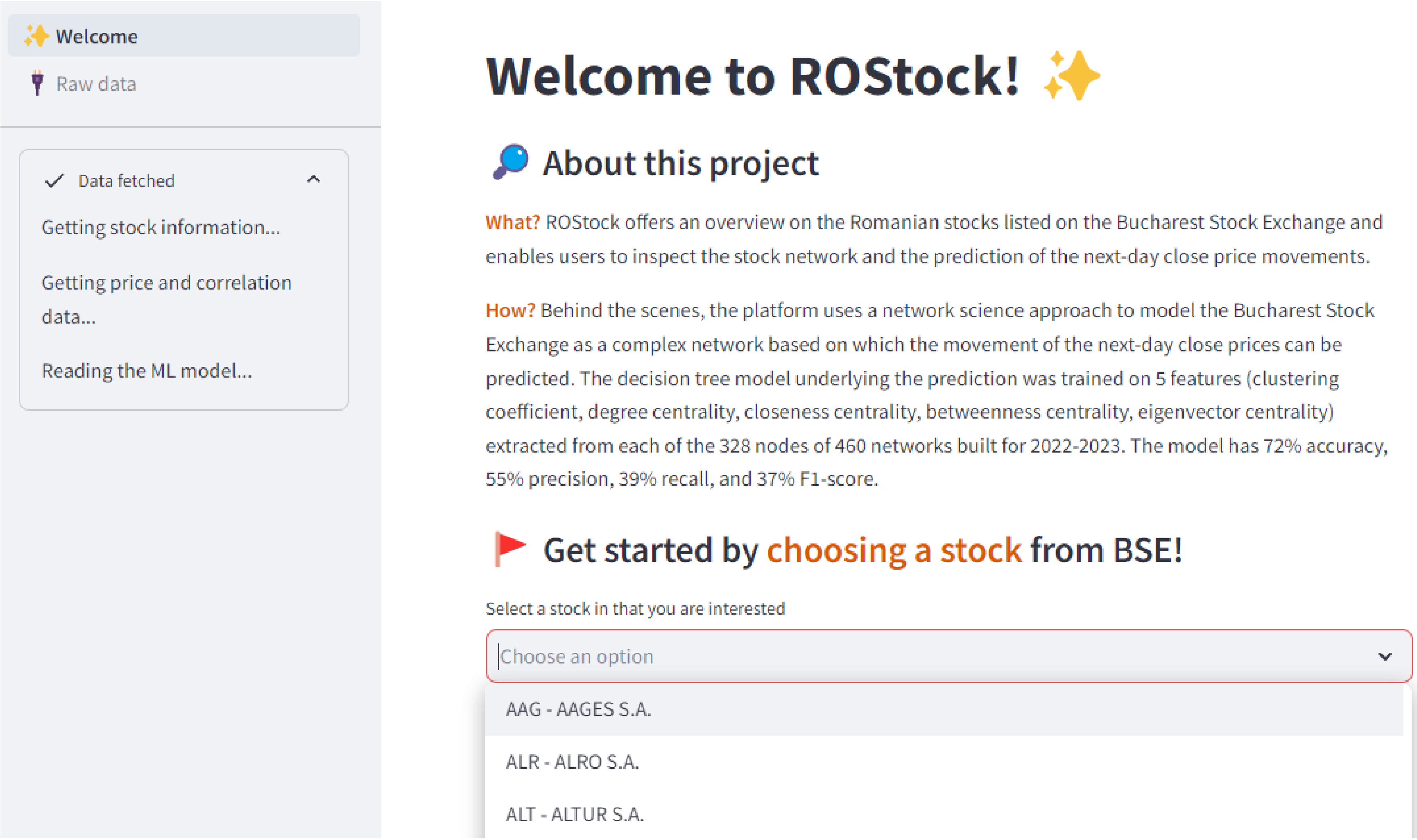


APPLICATION: ROSTOCK MONITOR

Architecture & Technologies used



User Interface – landing page



User Interface – prediction and network

✨ Welcome

Raw data

✓ Data fetched

🚩 Get started by choosing a stock from BSE!

Select a stock in that you are interested

ALT - ALTUR S.A.

📋 About ALT

Issuer company: ALTUR S.A.

Market: REGS **Sector**: MANUFACTURING **Part of BET**: No

Prediction

Select the day for that you want the prediction

2023/12/29

🛠 The network to predict the stock movements on 2023-12-29 is based on a default 60 trading days range, from 2023-10-06 to 2023-12-28

🔑 Input data: Clustering coefficient: 0.2581 || Degree centrality: 0.2080 || Closeness centrality: 0.4559 || Betweenness centrality: 0.0027 || Eigenvector centrality: 0.0710 ||

💡 Predicted movement for the close price of ALT on 2023-12-29: stagnate

✅ Real close price movement was predicted correctly

✨ Welcome

Raw data

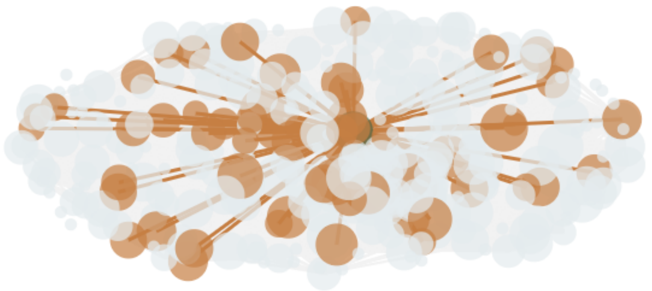
✓ Data fetched

The network underlying the prediction

Network properties: Number of nodes (N): 328.00 || Number of isolated nodes: 68.00 || Number of edges (L): 7408.00 || Edge density: 0.14 || Average Clustering Coefficient: 0.24 || Minimum degree: 0.00 || Maximum degree: 94.00 || Average degree: 45.17 || Average degree centrality: 0.14 || Average betweenness centrality: 0.00 || Average closeness centrality: 0.35 || Average eigenvector centrality: 0.05 ||

Please expand the visualization for better visibility.

Legend: Nodes are sized by degree. The node in green is ALT and the nodes highlighted with brown are the stocks which had a correlation higher than 0.15 with ALT.



User Interface – data behind the app

✨ Welcome

Raw data

✓ Fetching data...

Getting stock information...

Getting price and correlation data...

The data behind ROStock! ⚙️

Information about the stocks:

symbol	share_name	company_name	market	sector
AAG	AAGES S.A.	S.C AAGES S.A.	REGS	MANUFACTURING
ALR	ALRO S.A.	ALRO S.A.	REGS	MANUFACTURING
ALT	ALTUR S.A.	ALTUR S.A.	REGS	MANUFACTURING
ALU	ALUMIL ROM INDUSTRY S.A.	ALUMIL ROM INDUSTRY S.A.	REGS	WHOLESALE AND RETAIL T
AQ	AQUILA PART PROD COM	AQUILA PART PROD COM	REGS	WHOLESALE AND RETAIL T
ARM	ARMATURA S.A.	ARMATURA S.A.	REGS	MANUFACTURING
AROBS	AROBS TRANSILVANIA SOFTWARE	AROBS TRANSILVANIA SOFTWARE	REGS	INFORMATION AND COMM
ARS	AEROSTAR S.A.	AEROSTAR S.A.	REGS	MANUFACTURING
ARTE	ARTEGO SA	ARTEGO SA	REGS	MANUFACTURING
ATB	ANTIBIOTICE S.A.	ANTIBIOTICE S.A.	REGS	MANUFACTURING

Historical stock prices:

	dt	AAG	ALT	AQ	AROBS	ARTE	BCM	BNET	BRK	BVB	CBC	CMC
0	2022-01-03	3.6	0.0539	0.9233	0.975	11.9	0.098	0.3177	0.283	25.4	17.1	0.2
1	2022-01-04	3.56	0.0539	0.9233	1.01	11.5	0.095	0.3255	0.292	25.2	17.1	0.2
2	2022-01-05	3.58	0.0539	0.92	1.04	11.5	0.095	0.3445	0.3	25.4	19.2	0.2
3	2022-01-06	3.54	0.0579	0.9167	1.04	11.5	0.095	0.3491	0.293	25.2	22	0.2

Future work

THE INPUT DATA EXTRACTION

Different periods could be used for the sliding window to optimize the prediction. When modelling the stock networks, different thresholds can be used for each network. It could be also an option to start from a base-network and each day add the new information to this.

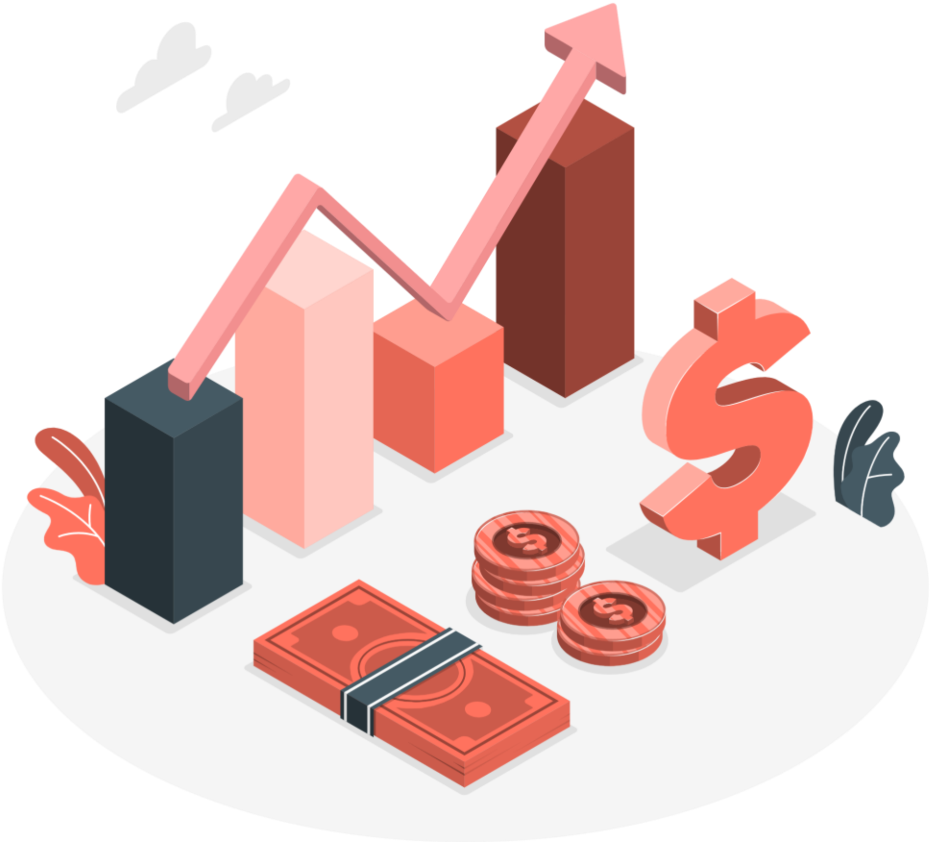
THE PREDICTION ALGORITHMS

Extract more input features from the network, even on the macro- and meso-level. Compare different algorithms and parameters. Try ensemble models, for example stagnate vs. others layer and increase vs decrease layer.

BUCHAREST STOCK EXCHANGE

THE ANALYSIS OF STOCK NETWORKS

Comparing networks modelled based on different thresholds to uncover more insights e.g. constant connections independent of the threshold



PyData Cluj-Napoca

27th of July, 2024

PREDICTING ROMANIAN STOCK MOVEMENT TRENDS: A COMPLEX NETWORK APPROACH COMBINED WITH MACHINE LEARNING

Author: Orsolya – Dorottya Holgyes

Supervisors: Prof. dr. Camelia Chira and Asist. drd. Bogdan Mursa

Babeș - Bolyai University, Faculty of Mathematics and Computer Science