# Financial partial correlation network and centrality portfolio

Alexandru Melnic 1692625

## Introduction

In this project the S&P 500 returns will be analysed by inferring their dependency structure through the partial correlation graph. In particular the network will be built by using the GLasso algorithm, that estimates a sparse precision matrix. An analysis of such network will be made both on the single companies and sectors. From the network structure, it will be shown an approach, inspired by [1], to build a portfolio by exploiting the node centralities. This portfolio in the end will be confronted on different time periods against the classical minimum variance and the naive $1/N$ portfolios.

## Data processing

The data are downloaded from [2] and they contain information about all the S&P 500 daily close prices between the year 2013 and 2020 and the respective company sectors. The data is processed in order to get the returns, defined as the daily price variations that, mathematically speaking are:

$$r_i = \frac{p_i - p_{i-1}}{p_{i-1}}$$

where $r_i$ is the return and $p_i$ is the close price at day $i$.

## Graphical models

### Undirected graphs

An undirected graphical model is a graph $G$ where the nodes $V = \{X_1, ..., X_N\}$ are associated to random variables. If there is no edge between two nodes it means that the associated random variables, conditioned to all the others in the network, are independent. This one is called the pairwise Markov property and under the assumption of the nodes following a multivariate Gaussian distribution a graph can be constructed by inspecting the partial correlation matrix $\tilde{R}$. In fact from it one can build a graph by putting an edge whenever the term of the matrix $\tilde{\rho}_{ij} \neq 0$. In reality the partial correlation matrix $\tilde{R}$ is not known and has to be estimated. This process is subject to errors that even if are small it would never lead to a solution where even one term of the estimated matrix is zero. For this reason to build this type of graph it is not easy to estimate the matrix.

### Partial correlation

The partial correlation is the correlation evaluated on the residuals given by the variables and the linear predictions of them by using all the other available variables. In simple terms given an $N$-dimensional random vector $\mathbf{X} = [X_1, ..., X_N]$ the partial correlation $\tilde{\rho}_{ij}$ between $X_i$ and $X_j$ is the correlation between the variables but without the linear effect of all the others. In practical terms this can be done in the following way:

$$\tilde{\rho}_{ij} \propto \mathbb{E}[(X_i - \hat{X}_i(\mathbf{X}/\{X_i, X_j\}))(X_j - \hat{X}_j(\mathbf{X}/\{X_i, X_j\}))]$$

where $\hat{X}_i(\mathbf{X}/\{X_i, X_j\})$ is the linear prediction of $X_i$ by using all the features except $X_i$ and $X_j$.

Under the assumption of $\mathbf{X}$ following a multivariate normal distribution the correlation implies independence and so does the partial correlation and conditional independence. This means that by knowing the partial correlation matrix it is possible to construct the respective Markov network.

### Precision matrix

There exists an important relationship between the precision and the partial correlation matrix. Denoting with $P_{ij}$ and $\tilde{\rho}_{ij}$ the terms of the precision and partial correlation matrices respectively, the elements on the diagonal of the partial correlation matrix are given by:

$$\tilde{\rho}_{ii} = \frac{1}{P_{ii}}$$

Furthermore the off-diagonal elements are:

$$\tilde{\rho}_{ij} = -\frac{P_{ij}}{\sqrt{P_{ii}P_{jj}}}$$

Thus by knowing the precision matrix it is possible to immediately get also the partial correlation matrix. This allows to infer the partial correlation graph by using only the precision matrix since the off diagonal term of it being $P_{ij} \neq 0$ implies that $\hat{\rho}_{ij} \neq 0$ which also implies that $X_i$ and $X_j$ are conditionally independent given all the other nodes.

## Graphical lasso

As anticipated in the previous section the estimation of the partial correlation matrix is not easy, or even impossible with simple methods. In this project the graphical

lasso approach will be used.
Suppose we have $N$ titles which daily returns are represented by the random variable $\mathbf{x}$ which follows a multivariate Gaussian distribution:

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N|\Sigma|}}\exp\{-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}\}$$

where $\Sigma$ is the covariance matrix and for simplicity the vector has mean zero.
If we have $T$ *independent* observations of the daily returns the precision matrix $\Sigma^{-1} = P$ can be estimated through the MLE:

$$p(\mathbf{x}_1, ..., \mathbf{x}_T|P) = \prod_i \frac{1}{\sqrt{(2\pi)^N|\Sigma|}}\exp\{-\frac{1}{2}\mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i\}$$

$$\log p(\mathbf{x}_1, ..., \mathbf{x}_T|P) \propto \frac{T}{2}\log|P| - \frac{1}{2}\Sigma_i\mathbf{x}_i^T P\mathbf{x}_i = \frac{T}{2}\log|P| - \frac{1}{2}\mathrm{Tr}(X^T PX) =$$

$$\frac{T}{2}\log|P| - \frac{1}{2}\mathrm{Tr}(XX^T P) = \frac{T}{2}\log|P| - \frac{T}{2}\mathrm{Tr}(\hat{C}P)$$

where $\hat{C} = \frac{1}{T}XX^T$ is the sample covariance, $X$ the design matrix and in the last part the invariance under cyclic permutations property of the trace was used. From this the MLE for $P$ is given by:

$$\hat{P}_{ML} = \mathrm{argmax}_{P\geq 0}\log|P| - \mathrm{tr}(\hat{C}P) = \mathrm{argmin}_{P\geq 0}\mathrm{tr}(\hat{C}P) - \log|P|$$

The solution is simply given by $\hat{P} = \hat{C}^{-1}$, but if $T < N$ the rank of the matrix $\hat{C}$ is less than $N$, thus it is not invertible.
This is where the graphical lasso algorithm comes in. Instead of solving the previous problem, in the objective it is added a penalization term, obtaining the optimization problem:

$$\mathrm{argmin}_{P\geq 0}\mathrm{tr}(\hat{\Sigma}P) - \log|P| + \lambda||P||_1$$

The last optimization problem is convex (in particular a SDP one) since it is the sum of three convex functions and the domain is the one of symmetric positive semi-definite matrices. In particular the first term is a linear operation in the components of $P$, $\log|P|$ is a concave function, thus $-\log|P|$ is convex and lastly $||P||_1$ is convex since it is a proper norm and the multiplication for a non negative coefficient keeps it convex.

## Graph inference

An approach one can take to infer the dependency structure between a group of random variables is by estimating the precision matrix by inverting the covariance matrix. After that by selecting a threshold $t$ it is possible to set to zero the term $\tilde{\rho}_{ij}$ of the partial correlation matrix if $|\tilde{\rho}_{ij}| < t$.
The partial correlation is evaluated on the daily close price returns of the period

between 2013 and 2018. In the GLasso problem the $L_1$ penalization term introduces sparsity in the solution thus the thresholding is done automatically by selecting the regularization parameter $\lambda$. One drawback of the GLasso is that not for every choice of $\lambda$ the algorithm converges, problem that doesn't exist in the first approach.
In this project tipical regularization values range from 0.00015 to 0.00030, where the first value corresponds to low and the second to high sparsity of $P$.

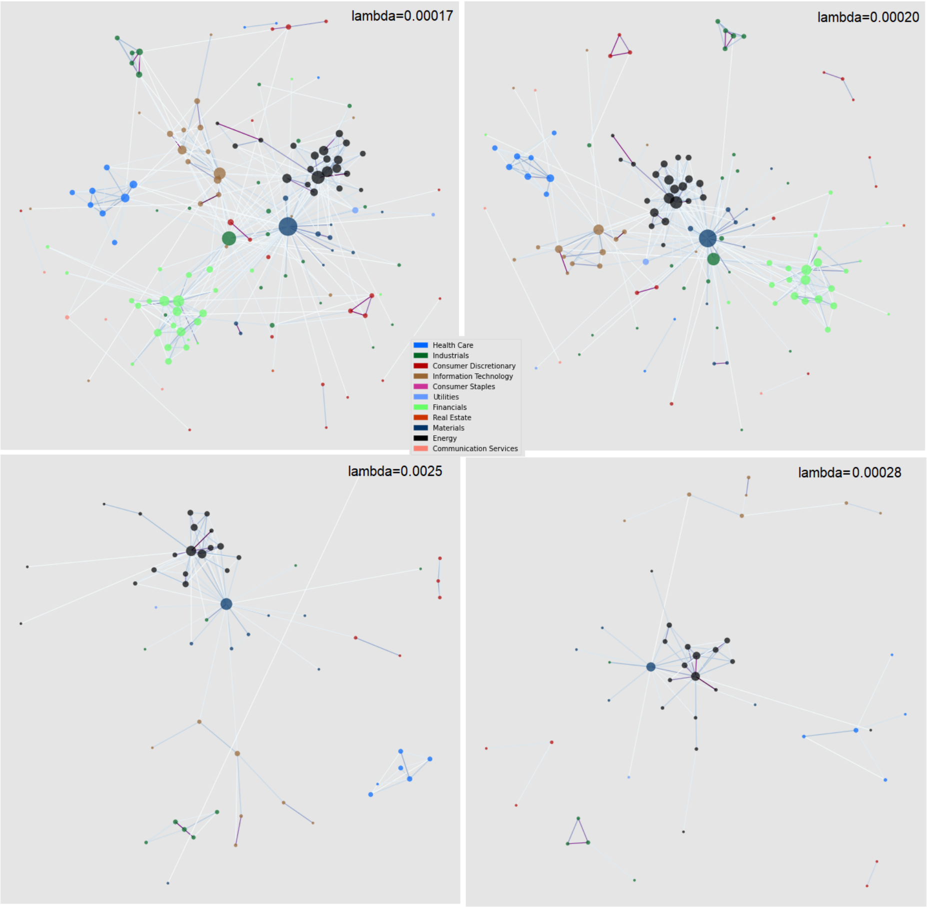

**FIGURE 1.** *Partial correlation networks for different values of $\lambda$. Node color corresponds to sectors, node size to degree and edge color intensity to the partial correlation coefficient. Nodes without an edge are not represented.*

From all the network plots[1] in Figure 1 it is possible to notice that the sectors are clustered, it means the companies within each sector tend to influence mainly each other rather than companies of other sectors. The components that are most connected in the graph are given by the energy, finance, information and health care sectors. The energetic one is the *strongest* component in the graph and in fact those edges are the last ones to disappear by varying the sparsity parameter.

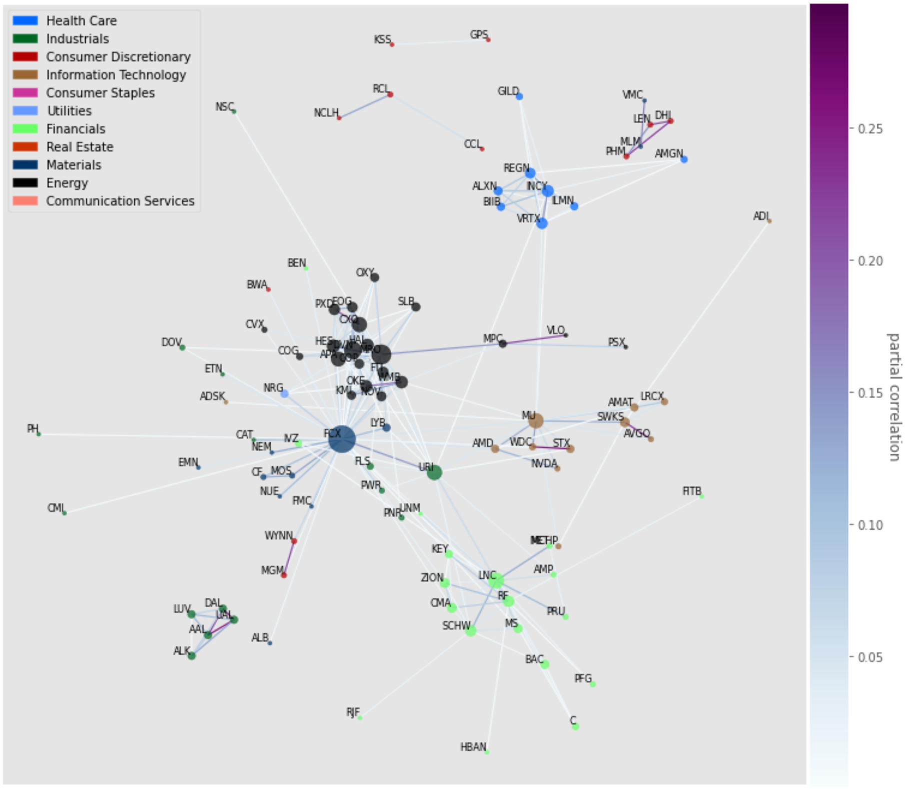

**FIGURE 2.** *Partial correlation graph for $\lambda = 0.00020$ with the respective company tickers.*

In Figure 2 are interesting the disconnected components present in the network, in particular two of them: the airlines component (LUV,AAL,UAL,ALK,DAL) and the buildings component (PHM, DHI, LEN). These two parts stay disconnected from the

[1] In all the network plots the nodes DISCK and DISCA were excluded since they have a much higher correlation than any other link (shown in Table 2) and in this way they make appear all the edge colors white.

rest of the graph for many values of $\lambda$ and their partial correlations are between the highest in the graph, as shown in Table 1, suggesting that these companies are not much correlated from other companies but mainly from themselves.

| | triangle | average partial correlation | sector |
|---|---|---|---|
| 1 | AAL DAL UAL | 0.24 | Industrials |
| 2 | DHI LEN PHM | 0.19 | Consumer Discretionary |
| 3 | COP DVN MRO | 0.17 | Energy |
| 4 | COP HES MRO | 0.16 | Energy |
| 5 | AAL LUV UAL | 0.16 | Industrials |
| 6 | AAL ALK UAL | 0.15 | Industrials |
| 7 | CXO EOG PXD | 0.15 | Energy |
| 8 | MU STX WDC | 0.15 | Information Technology |
| 9 | ALK DAL UAL | 0.15 | Industrials |
| 10 | DAL LUV UAL | 0.14 | Industrials |

**TABLE 1.** *Top 10 triangles in the graph with* $\lambda = 0.00020$.

In Table 2 are present the nodes that have between them the highest partial correlation for the network with $\lambda = 0.00020$, apart from the nodes that also make the triangles in Table 1 there are other interesting dependencies. DISCA and DISCK are stocks that belong both to Discovery Inc (it has sense that their returns are highly correlated since the economic decisions of one company affects both stocks). Furthermore there are also STX (Seagate) and WDC (Western Digital) that are both manufacturers of hard disks, MGM and WYNN (casinos and gaming) and AVGO and SWKS (semiconductors). Not only these companies are in the same sector but also in the same subsector or even direct rivals.

| | node 1 | node 2 | partial correlation |
|---|---|---|---|
| 1 | DISCA | DISCK | 0.70 |
| 2 | COP | MRO | 0.29 |
| 3 | STX | WDC | 0.26 |
| 4 | AAL | UAL | 0.25 |
| 5 | DAL | UAL | 0.23 |
| 6 | SWKS | AVGO | 0.22 |
| 7 | DHI | PHM | 0.22 |
| 8 | AAL | DAL | 0.22 |
| 9 | MGM | WYNN | 0.21 |
| 10 | DHI | LEN | 0.21 |

**TABLE 2.** *Top edges for* $\lambda = 0.00020$

## Sectors analysis

As anticipated at the beginning of the section, and as it is possible to notice in the previous figures, the sectors appear clustered. The plots in Figure 3 are generated from aggregate information of their respective company networks. In particular the nodes are associated to the partial correlation within a sector and the edges are associated to the partial correlation with the companies of other sectors. The scale is the same for both nodes and edges and the metric used is the sum of partial correlations. This metric was chosen to keep into account not only the weight of the edges but also the number of vertices contributing to the score. In this way the score is high both for cases where there are few companies with strong connections and in the case where there are many companies with weak connections, effect that would lead to different results if the mean would have been considered as a metric instead.

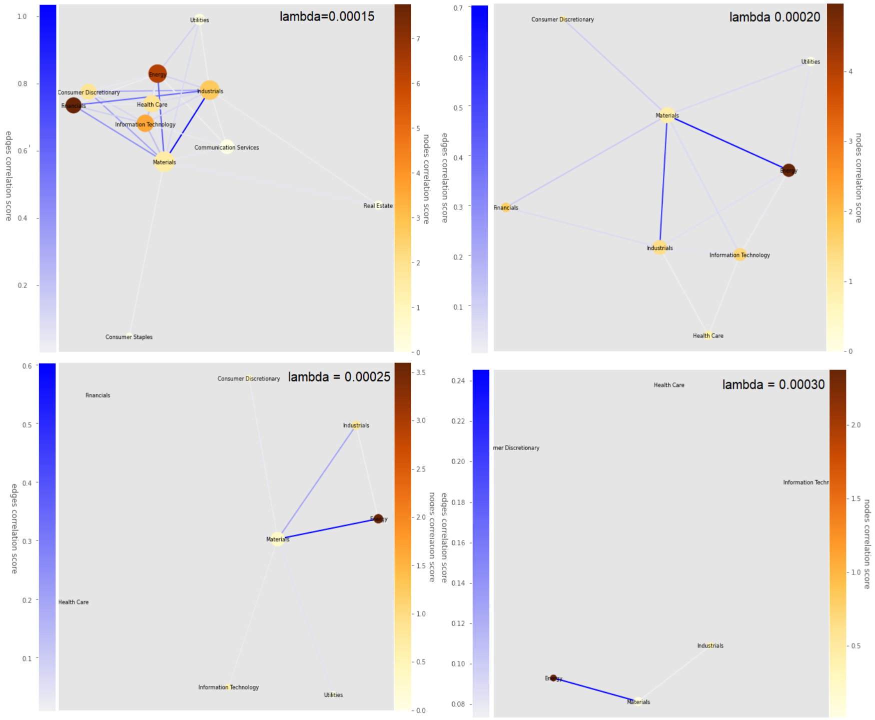

**FIGURE 3.** *Network for the sectors derived from the company graphs for different values of $\lambda$. If some nodes are not shown is because in the relative company graph that sector doesn't have any edge with other companies. The size of the node is associated to the node degree.*

In all cases it appears that the companies in the same sector are more dependent from each other rather than from companies of other sectors, in fact the weights of the nodes are almost one order of magnitude bigger than the weights of the edges. The companies of the Energy sector appears almost always as the most dependent ones as also shown in the previous Figures. Also their connections, in particular with the Materials and Industrials sectors, represent the strongest inter-sector connections within all edges.

## Portfolio optimization

The main idea of this section is to build a portfolio that comes from the structure of the partial correlation network. There exist many types of portfolio models and the simplest is the Markowitz one. The main idea of this model is to find some weights $\{w_1, ..., w_N\}$ that indicate how much of the total budget should be allocated on the different assets $\{1, ..., N\}$. The optimal weights $\mathbf{w}$ can be computed by the following optimization problem:

$$\min_{\mathbf{w}} \mathbf{w}^T \Sigma \mathbf{w}$$

$$s.t. \;\; \mathbf{w}^T \mathbf{1} = 1$$

$$w_i \geq 0$$

where $\Sigma$ is the covariance matrix of the returns.The objective of the previous optimization problem is the variance of the total portfolio return which is defined as $R = \sum w_i r_i$, where $r_i$ is the return of the asset $i$. This optimization problem gives the weights of the portfolio with the lowest risk, and this is usually called as the Minimum Variance (MV) portfolio.

In this project it will be tested an approach involving the use of assets with low centrality in the partial correlation graph. The centrality measure is the eigenvector one, that is an extension of the simple degree centrality with the difference that if one node is connected to a node with many links than that connection weights more than one with a node with fewer links. The main idea behind this portfolio is justified by the Figure 4.

From Figure 4 one can see that the assets with higher average returns and lower volatility have also lower centrality. From here the main idea of the portfolio:

1. Estimate the partial correlation graph through the GLasso optimization problem with a certain $\lambda$.
2. Get the $K$ nodes with the lowest centrality.
3. Assign to all these companies equal weight $1/K$ and zero to all the others.

In this way it is possible to select nodes with *low* volatility and *high* return.

### Results

The evaluation of the method was conducted first by tuning the parameters $\lambda$ and $K$ (number of nodes to select in the centrality portfolio) on the period between 2013 and 2018 (train set), by using a *proto-cross-validation* and then tested on the period

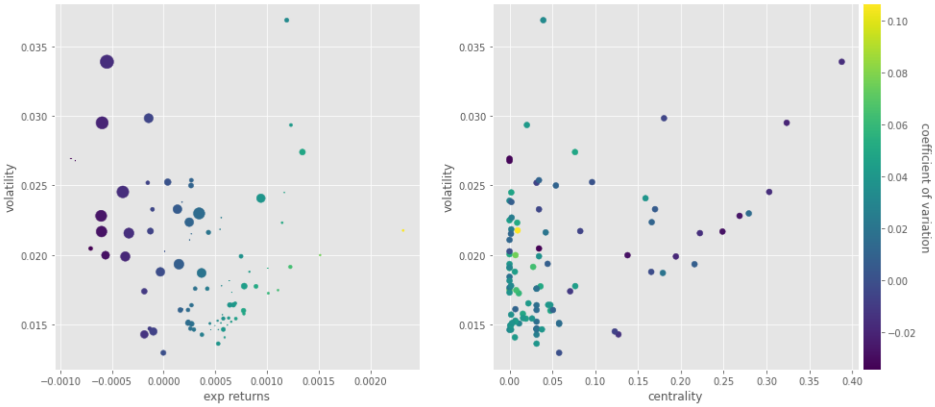

**FIGURE 4.** *On the left the scatter plot of the average returns for all companies and their standard deviation (volatility) where the size is associated to the centrality measure of the partial correlation graph with* $\lambda = 0.00020$*. On the right the scatter plot of centrality and volatility.*

from 2018 to the end of 2020 (test set). The centrality portfolio is then confronted with the minimum variance and the naive $1/N$ portfolios[2]. The parameter tuning was performed by splitting the train set in 5 folds of 250 days each (approximately one year) and then each fold in one part of 150 days for training and the other part in 100 days for evaluation. The evaluation for each fold (and on the test set) consisted in the computation of the wealth for each of the 100 days, that for the day $t$ is given by $wealth_t = \prod_{i=1}^{t}(1 + \overline{R_i})$, where $\overline{R_i}$ is the realization of the portfolio return at day $i$. The tuning was performed to achieve the highest wealth and the stability of the GLasso algorithm, since this one does not always converge. The final portfolio weights, for both the minimum variance and the centrality portfolios were evaluated by computing the average of the weights on all the 5 folds. With this tuning cycle the chosen parameters are $\lambda = 0.0025$ and $K = 20$.

From Figure 5 and 6 it is possible to notice that the centrality portfolio behaves well sometimes, but not enough to substitute the other two. From the results it looks like this portfolio behaves better than the other ones in periods of general market growth and worse in periods of market drop. This is also present in the test performance, since in the first 150 days of general market growth the centrality portfolio performs better then the others, but after that, after around 200 days there is a general market drop, where the portfolio starts to perform worse and is not able to recover from it. Same argument for the drop after around 500 days that corresponds to the March of 2020, beginning of the Covid-19 pandemic.

[2] It is the portfolio where on all the assets it is allocated an equal weight.

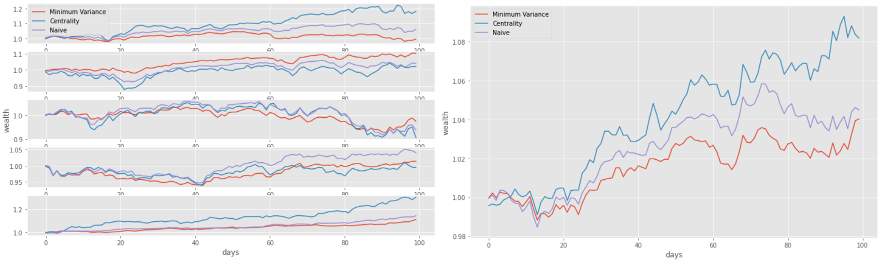

**FIGURE 5.** *On the left the evaluation on the 5 folds and on the right the average performance in the folds.*

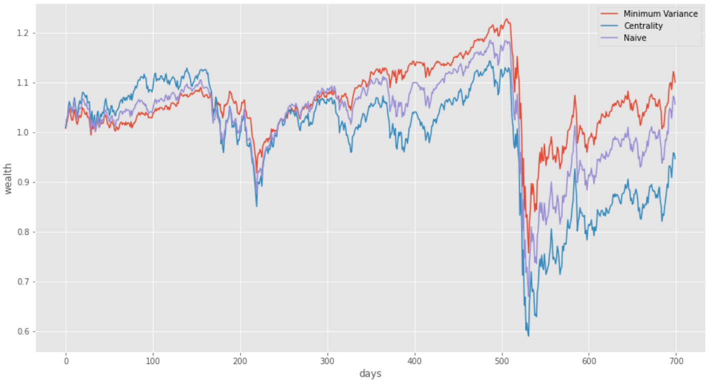

**FIGURE 6.** *Results for the portfolios in the test set.*

In Figure 7 it is taken the same idea of Figure 4 with also the plot of the nodes of the portfolios in red. There is a big difference between the two approaches. The minimum variance portfolio has all low volatility assets meanwhile the centrality portfolio has all higher volatility and higher return assets. This also explains the performance on the test set and can be seen in Table 3: the assets in the centrality portfolio have higher return and volatility in average than the minimum variance ones. Another fact is that the minimum variance portfolio selected nodes do not appear in the scatter plot of Figure 7, that means that their dependence with other nodes is low.

In figure 8 are shown the top assets that are present in the minimum variance and

| | average asset return | average asset volatility |
|---|---|---|
| Minimum Variance | 3 x10$^{-}$4 | 1 x10$^{-}$2 |
| Centrality | 6 x10$^{-}$4 | 2 x10$^{-}$2 |

**TABLE 3.** *Average asset return and volatility for the minimum variance and centrality portfolios.*

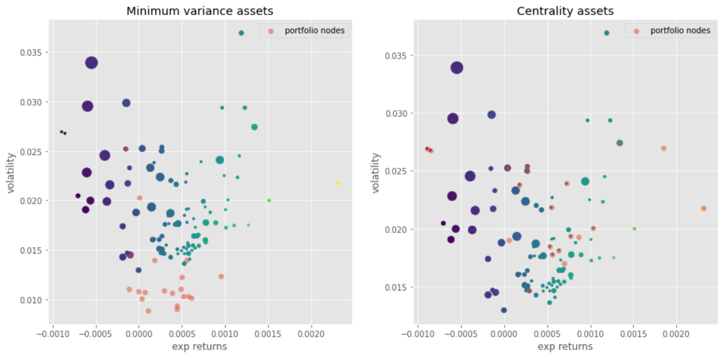

**FIGURE 7.** *Average daily return vs volatility for the portfolio's assets. The size of the points corresponds to the node centrality in the network with* $\lambda = 0.00020$*. In red the nodes of the top 20 assets in the portfolios.*

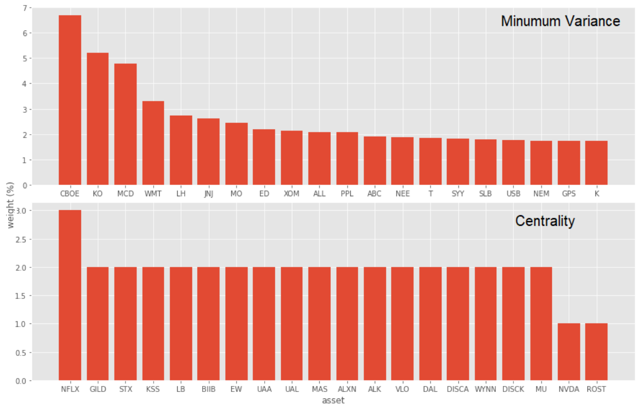

**FIGURE 8.** *Top 20 assets according to their weight for the minimum variance and centrality portfolios.*

centrality portfolios. For the latter with the tuned parameter $\lambda = 0.00025$ only few nodes in the partial correlation network have some links. In fact by also looking at Figure 1 and Figure 2 it is possible to understand which nodes are selected in this portfolio. These assets are mainly those nodes that have *high* partial correlation with other nodes and are not part of the main large connected component of the graph, that is reasonable since they are chosen as the nodes with lowest centrality. Among all the nodes in the centrality portfolio are present the companies of the airlines component (described in the Graph Inference section).

## Conclusion

In the section of Graph Inference some interesting insights, like the disconnected components, or the sector dependencies were found. By making some considerations on the relationship of centrality, expected return and volatility it was possible to build a portfolio that only depends on the structure of the partial correlation graph. In the end this did not outclass the other simpler portfolios. As seen in the last section the nodes selected by the centrality portfolio are not the ones corresponding to the lowest variance and highest return and the main reason for it is the choice of $\lambda$. This tuning parameter depends on the data, and the range of values that leads to particular solutions changes with the data, reason for which it is not easy to tune it when performing the cross validation. Furthermore the graphical lasso algorithm does not converge for all values of $\lambda$ making the whole process harder.

## References

[1] A network approach to portfolio selection, Gustavo Peralta - Albalfazl Zaarei
[2] Train set, Test set