The Financial Network Of The Italian Stock Exchange

Tommaso Lussetich, Quantitative Finance, 0001078902
Niccolò Marzi, Artificial Intelligence, 0001103537
Elisa Venturoli, Artificial Intelligence, 0001103710

November 23rd, 2023

1 Introduction

Portfolio allocation is one of the largest and most developed strands of literature of the wider field of financial economics. Almost every financial service and intermediary revolves around the correct application of its principles. This practice can be divided into two steps. First of all the assets to be purchased are selected from a broader pool of financial instruments. This phase is known in literature as cardinality-constrained portfolio selection. Afterwards, the relative amounts (the so called weights) are chosen by optimizing a specific function that embeds the preferences of the economic agent.

This work focuses on cardinality-constrained portfolio selection by applying network analysis techniques to the Italian stock exchange. The theoretical basis we rely on is the classical mean-variance portfolio optimization model by Markowitz, 1952[1]. This groundbraking study gave rise to a new set of models that present a very intuitive approach to the investment decision making process by framing it in a risk-return trade-off. Here, the volatility is a proxy for risk, while return is expressed in expected (mean) returns.

Network analysis comes in aid as in these kind of models the relationship among assets is represented by the correlation between the historical returns of two assets. Correlation can be seen as the degree by which assets exchange information. Indeed, as price movements infer information about the general state of a financial market, if two assets have a strong -positive or negative- correlation, it means that they transmit the same (or opposite) signal through the system. As a portfolio is formed by multiple instruments, in asset selection it is useless to evaluate the correlation among assets pairwise. Instead, it is crucial to understand how feedbacks and incentives circulate within the network.

2 Problem and Motivation

This paper aims to provide a comprehensive and updated investigation of the Italian stock exchange by means of the network structure emerging from the correlation between stocks. The results can be exploited in the context of asset selection for the composition of an optimal portfolio that maximises the expected return and while controlling for the risk associated with the assets.

Furthermore, it brings forth the comparison of two alternative strategies for the construction of an optimal mean-variance portfolio. One strategy focuses on stocks with low correlation, while the other on assets with negative and (almost) null correlation.

3 Datasets

Borsa Italiana SpA publicly discloses the list of companies aggregated in the FTSE Italia All Share index. This index contains the whole set of securities traded in the Italian stock market which amounts to 202 assets. Each stock is uniquely identified by an International Securities Identification Number (ISIN) obtained by searching for the list of quoted companies in the AIDA Database.

Through the Refinitiv Datastrem database we obtain the closing prices at a daily frequency, industry codes, annual revenues and monthly trading volumes for the 202 firms. The time window we investigate is 01/01/2021 to 11/07/2023, for a total of 742 daily observations for each stock. The final dataset consists only of daily prices of companies with available information for each trading day of the time horizon considered. Some of them have missing data because absent on the database or because they were not listed yet at the beginning of the observation period. Thus, we are left with 184 firms.
Afterwards, we proceed by computing the daily logarithms of returns as

$$R_i(t) = ln\left(\frac{P_i(t)}{P_i(t-1)}\right) \tag{1}$$

Where $P_i(t)$ and $P_i(t-1)$ stand for the price of the security i at times t and *t-1* respectively. We choose to express it in the logarithmic representation because, even if it is of a less immediate interpretation compared to the percentage form, expressing financial returns in logarithmic terms helps reduce the apparent variability of the data and makes proportional variations more evident.

In order to assess which is the correlation coefficient that best fits the data we resort to the traditional assumptions about returns distribution. A significant strand of financial literature poses that returns are normally distributed. Seminal works about the Efficient Market Hypothesis as Bachelier,1900[2] and Cowels, 1933[3] support this principle, as well as more recent studies by Fama (1970)[4] and Merton (1973)[5]. In this classical setup the idea that prices follow a random walk and that their future value is unpredictable implies that the expected return must be zero. This framework is essential to understand why prices embed all publicly available information and transmit signals through the market. Indeed, the first three moments of the distribution confirm this assumption as they are of the same order of those of the standard normal distribution. The fourth moment, i.e. kurtosis, slightly departs from the theoretical value of three signaling as the distribution is leptokurtotic.

Since the distribution of returns does not significantly depart from the normal distribution, it is safe to estimate the correlation between two securities i and j according to the Pearson coefficient:

$$\rho_{ij} = \frac{Cov(i,j)}{SD(i)SD(j)} \tag{2}$$

This ratio can take values in the interval [-1,1] and expresses the linear relation between the trends of the returns of two securities. This computation is repeated for each couple of stocks in order to obtain a symmetric matrix of dimensions 184×184 and with a diagonal composed by ones. In order for this to be the adjacency matrix of a non-fully connected network we have to choose a threshold to characterize the resulting network according to the kind of relationship we want to analyze, namely positive correlation, low correlation, or negative correlation. The firms that survive this filtering constitute the nodes of the networks.

In order to set the ex-ante thresholds used in section 5 we study the distribution of the correlation coefficients by dividing it into deciles. This result is reported in Table 1.

Quantile	Correlation Value
1	0.0526
2	0.0840
3	0.1130
4	0.1387
5	0.1671
6	0.1988
7	0.2407
8	0.2917
9	0.3592

Table 1: Correlation coefficient values per quantile

4 Validity and Reliability

The validity of the data on which this study is based is ensured by legal obligations that the listed companies and the exchange operator, i.e., Borsa Italiana SpA, are required to meet. Each day, the closing prices are disclosed and reflect the intentions and expectations of the agents involved in trading activities. Nevertheless, the Refinitiv Datastrem database does not present information for one of the 202 listed companies. Despite this and all the other companies that we discard because of missing observations as explained in section 3, we end up covering more than the 90% of the securities traded in the Italian stock market.

For what concerns the choice of correlation as the link that expresses a relationship between two stocks we refer to a consolidated strand of financial literature that followed the landmark paper of Markowitz, 1952. This paper introduce the Modern Portfolio theory and is at the basis of the majority of portfolio allocation models (e.g., the capital asset pricing model by Sharpe, 1964[6], the Black-Litterman model, 1990[7], Fama and French three-factor model, 1992[8]). Toghether with this, we also refer to later works in the field of financial networks such as Bogingski et al. 2005[9], Chi et al. 2008[10] and Coletti and Murgia, 2016[11].

Finally, the choices underneath the correlation thresholds have a strong impact on results. We focus on three deciles of the distribution that showcase the extreme links among the stocks. By making this decision ex-ante we intend to exclude possible manipulations that would arbitrarily highlight certain characteristics of the network.

Given due consideration to all of the above remarks, we think that the results that follow are valid and reliable for a twofold reason. First, the publicity of the data we retrieved implies that this study can be easily replicated and tested. Moreover, the choice of the return correlations as a link between stocks is very natural and intuitive, as well as strongly supported both by the classical models of asset allocation theory and by the more recent strand of literature of financial networks.

5 Measures and Results

This section is devoted to the network analysis of three subsamples originated by selecting specific quantiles of the correlation distribution. We end up with three simple one-mode networks, where nodes represent a single set of entities, and neither self-edges nor multi-edges

are allowed. Each network is undirected and weighted with the weights representing the correlation coefficients between nodes.

5.1 The high correlation network

[1]

In first instance, we focus on the network of the most correlated companies. This is generated by selecting the 9th decile of the correlation distribution, i.e., those firms that have correlation equal or grater than 0.3592. This subset results to be formed by 110 stocks.

To start with, we take care of the distribution of the number of edges attached to each asset in the network, that is the degree of the nodes. Our interest lies in determining whether the degree distribution follows a specific trend, since the network displays significant characteristics that are closely linked to this distribution. This distribution is then plotted in a log-log graph and interpolated with a line. The result of such procedure is presented in Figure 1, which exhibits a regression line of slope -0.30.

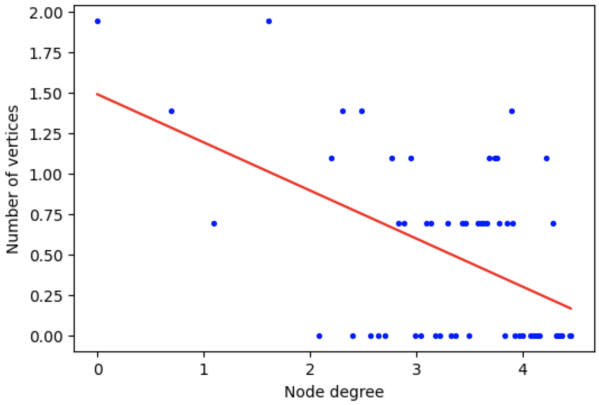

Figure 1: Degree distribution in log-log scale for the high correlation network. We obtain a linear regression with $r - value = -0.51$ and $p - value = 7.3 \times 10^{-5}$.

If the degree distribution of a network follows a power law distribution, then the network is said to be scale free. The power-law distribution is often expressed as $P(k) \propto k^{-\gamma}$, where k is the degree of a node, *P(k)* is the probability of a node having degree k, and γ is the exponent of the power law. The γ exponent is the slope of the fitting line. In the context of scale-free networks, a lower γ value corresponds to a more pronounced heavy tail in the degree distribution. A higher γ value implies a slower decay of the probability of high-degree nodes, resulting in a less pronounced heavy tail. Scale-free networks are robust, meaning that they preserve their characteristics even if some nodes are randomly removed. In our case, this implies that if certain companies are removed from the market, the network continues to exhibit the same features; it would take to remove many of them to observe any relevant change. In spite of the linear regression in Figure 1 not yielding an r-value indicative of a strong correlation, the p-value lets us suppose that the network may be approximated as scale-free.

Then we proceed by computing the edge density as the ratio between the actual connection taking place in the network over the all possible connections. The value that we obtain is equal to 0.28.

[1]The definitions of the measures presented in this section are taken from Newman, 2018[12]

It is interesting to discuss the density we obtain by comparing it to the computation of the average of the local clustering coefficients, which express the concept of transitivity among nodes. Indeed, the primary form of connection between nodes within a network is defined as being 'linked by an edge.' If this 'linked by an edge' relationship were transitive, it would imply that if node u is linked to node v, and v is linked to w, then u is also linked to w. However, unless all components of a network are cliques, perfect transitivity does not occur. Instead it can be useful to compute the clustering coefficient as the proxy for partial transitivity as follows.

$$C_i = \frac{\text{Number of neighbours of i that are connected}}{\text{Number of pairs of neighbours of i}} \tag{3}$$

$$\overline{C}_i = \frac{1}{\text{Number of nodes}} \sum_i C_i \tag{4}$$

The clustering coefficient can range in [0, 1]. Values of C of 0.20 are thought to be high. Since this network is characterized by a $\overline{C}$ equal to 0.74, we expect to witness to many highly interconnected elements. The combination of a relatively low density and a high clustering coefficient suggests that nodes are arranged in densely interconnected groups. More precisely, it is likely for this network to display cliques of large dimensions.

It is insightful to compare the local clustering coefficient with the node degree. By plotting these values on a log-log graph we obtain Figure 2. As one might expect, the local clustering coefficient decreases for nodes with higher degrees, but it still remains at a high value; this suggests the presence of nodes with high degrees even within large-dimensional cliques.

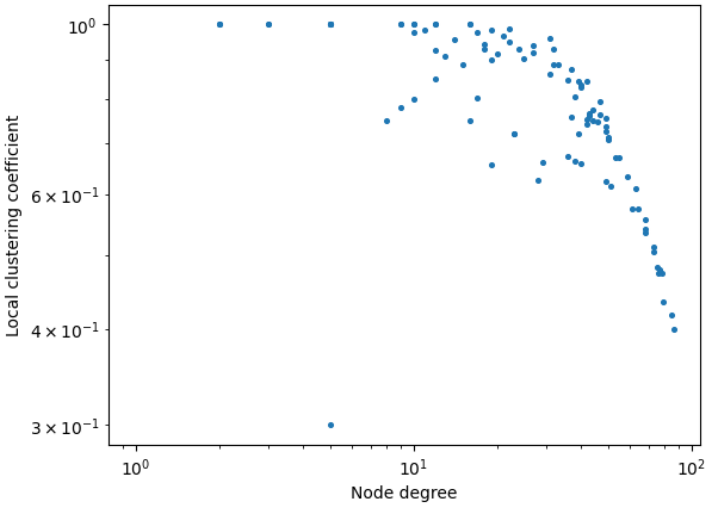

Figure 2: Local clustering coefficient in log-log scale as a function of node degree for the high correlation network.

Next, we are interested in understanding which nodes are the most likely to spread information through the network. Therefore, we compute the betweenness centrality for each node, as in this measure, the concept of importance of a node is translated into its position along as many shortest paths as possible. It appears reasonable to suggest that nodes exhibiting high betweenness centrality may have influence within a network due to their control over the flow of information between other nodes. (in the context of financial markets we talk about signals transmission). If a node with high betweenness is eliminated, all messages that were initially

scheduled to pass through that node must now be redirected through alternative (possibly longer) pathways. In an undirected network betweenness centrality can be computed as:

$$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}} \tag{5}$$

Where n_{st}^i is the number of shortest paths from s to t that pass through i and g_{st} is the total number of shortest paths from s to t. Hereby the ratio assumes the conventional value of zero whenever both n_{st}^i and g_{st} are zero. This is the average intensity of information passing through node i. The five nodes that exhibit the highest values of betweenness centrality are POSTE ITALIANE (0.087), AZIMUT HOLDING (0.061), INTESA SANPAOLO (0.055), ANIMA HOLDING (0.045), INTERPUMP GROUP (0.042). These are the nodes that serve as a conduit for signal transmission in the whole network. It is very likely for these to appear in the cliques of this network, and making a portfolio containing two or more of these assets is highly questionable, as they would exacerbate the performance of other assets fostering volatility.

Before delving into the study of cliques, let's focus for a moment more on the network by calculating the percentage of edges where the two involved nodes operate in the same sector. We obtain a value around 8%, which, on its own, does not strongly indicate a tendency for nodes to be connected with others operating in the same sector. After all, the correlation coefficient does not account for information such as sector or market capitalization. However, this value will become interesting after calculating it for the networks in the next section.

Finally, we analyze the cliques within this network. A clique is a set of nodes that are reciprocally connected one to the other. So, in this step we look for sets of stocks that are all linked one to the other at least by a correlation coefficient of 0.3592. In this network the maximum cliques that can be found are 6, each one formed by 28 stocks. We select one of these in order to see whether there is any pattern in the kind of companies forming the clique. Table 2 reports the names of the stocks as well as their yearly average market capitalization in 2022 (in millions of euros), disclosed revenues for the fiscal year 2022 (in thousands of euros) and industry. The values in columns (2) and (3) are useful indicators of the dimension of the firm considered.

Now that we have identified the clique, let's calculate the Jaccard similarity coefficient for each pair of nodes within it. The Jaccard similarity is defined as the ratio of the intersection size to the union size of two sets of elements. This can be expressed according to the formula

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}\,, \tag{6}$$

where A and B represent, in this context, the respective sets of neighbors associated with the two nodes under consideration. The Jaccard similarity coefficient takes values between 0 and 1: 0 indicates that the two sets have no common elements, while 1 indicates that the two sets are identical. From our clique we obtain a mean value of 0.63, suggesting that, on average, the nodes inside the clique are quite similar. In other words these nodes mostly share their neighbors. This stresses again the fact that nodes inside this network tend to cluster and form groups. From the matrix of Jaccard coefficients in Figure 3 we can notice that for company from the financial services the Jaccard similarities tend to be pretty high.

Name	Market cap	Revenues	Industry
POSTE ITALIANE	13149.46	28584679.46	Courier, Postal, Air Freight & Land-based Logistics
AZIMUT HOLDING	2986.67	1399455.60	Investment Management & Fund Operators
BANCA GENERALI	3782.05	1211866.51	Banks
BANCA MEDIOLANUM	5891.54	4048019.28	Banks
ANIMA HOLDING	1407.88	1119169.05	Investment Management & Fund Operators
INTESA SANPAOLO	42671.48	27859706.33	Banks
MEDIOBANCA BC.FIN	8402.45	3366368.30	Banks
BANCA IFIS	776.03	655495.35	Banks
STELLANTIS	48094.88	164476543.20	Auto & Truck Manufacturers
PIRELLI & C	4730.29	5972355.99	Tires & Rubber Products
FINECOBANK SPA	8536.63	1335052.57	Banks
TAMBURI INV.PARTNERS	1523.09	52204.60	Investment Management & Fund Operators
INTERPUMP GROUP	5125.55	1840654.89	Industrial Machinery & Equipment
ASSICURAZIONI GENERALI	27430.01	88699500.96	Life & Health Insurance
BIESSE	501.37	782235.01	Industrial Machinery & Equipment
OVS SPA	567.78	1419643.22	Apparel & Accessories
MONCLER	14840.85	2323962.15	Apparel & Accessories
ENEL	65126.80	109829028.80	Electric Utilities
CEMENTIR HOLDING	1213.06	1541191.01	Construction Materials
PRYSMIAN	8663.88	14398303.26	Electrical Components & Equipment
CNH INDUSTRIAL	18375.06	19394736.55	Heavy Machinery & Vehicles
DANIELI	906.12	3513456.19	Industrial Machinery & Equipment
WEBUILD	1722.82	6815302.96	Construction & Engineering
BUZZI	3944.07	3720007.20	Construction Materials
UNICREDIT	27291.40	25751681.38	Banks
BANCO BPM	4728.87	5046544.05	Banks
BANCA PPO.DI SONDRIO	1681.50	1206429.36	Banks
BPER BANCA	2871.81	4229680.56	Banks

Table 2: Stocks inside the clique selected from the high correlation network

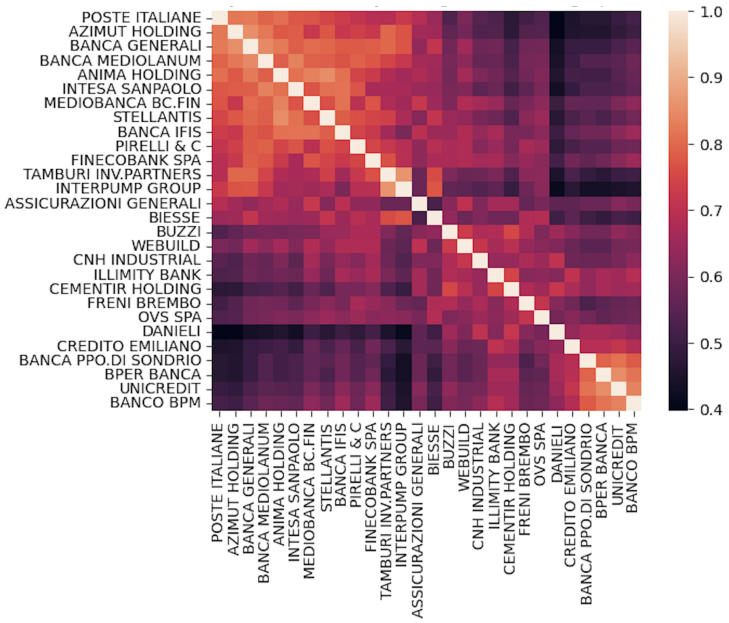

Figure 3: Jaccard similarity matrix for the clique selected from the high correlation network.

5.2 The negative and the low correlation networks

In order to find good candidates for the composition of an optimal portfolio according to the mean-variance theoretical framework, we select two more deciles of the correlation distribution and perform the same analysis exposed in section 5.1.

- The very low-negative correlation network (from now on, NC network[2]) originates taking into account the first decile of the correlation distribution. This means that the upper threshold of this subset is 0.0526.
- The low correlation (LC) network derives from the second decile of the correlation distribution, meaning that we consider stocks connected by a correlation coefficient ranging in]0.0526; 0.0840].

Both networks are composed by 184 nodes, meaning that the above thresholds do not reduce the total number of firms in the sample. They also exhibit the same values of edge density (0.10). It is not surprising that this value is lower than that of the network in Section 5.1; on the other hand, these two networks have the same number of edges as the other but all the nodes.

Figure 4 shows the degree distributions of the NC network (left) and the LC network (right).

[2]From now on this network will be also referred to as negative correlation network for wording reasons, even if it is clear that it includes also non-negative edges

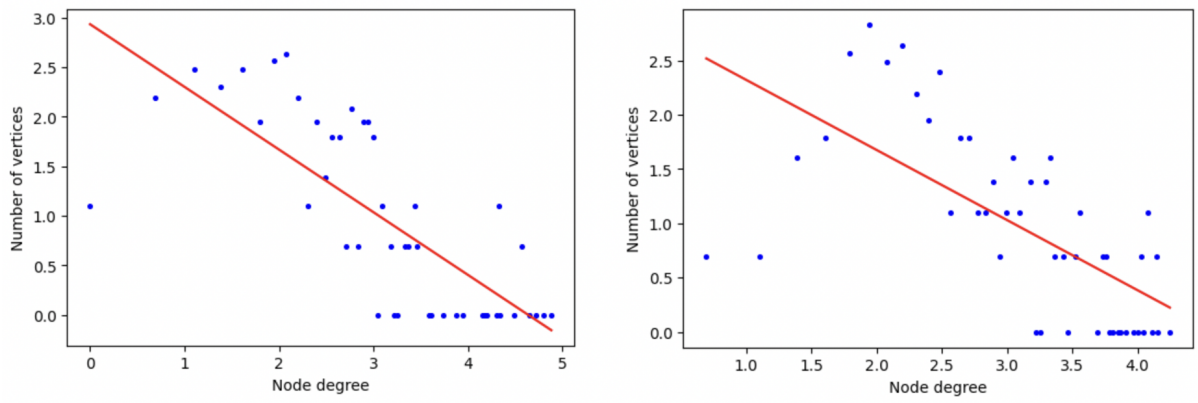

Figure 4: Degree distribution in log-log scale for the NC (left) and LC (right) networks. We obtain a linear regression with $r-value = -0.76$ and $p-value = 6.4 \times 10^{-10}$ for NC and $r-value = -0.68$ and $p-value = 8.6 \times 10^{-8}$ for LC.

The slope of the fitting line of the NC network is -0.63, while that of the LC network is -0.65. The r-value and p-value of the two linear regressions from Figure 4 make us confident that our networks are approximately scale-free.

The average of the local clustering coefficients are 0.59 for the NC network and 0.22 for the LC network. Despite they have the same edge density, LC has roughly one third of NC's clustering coefficient meaning that its nodes are less prone to forming groups, thus we expect the former to exhibit smaller sized cliques.

Figure 5 displays the graphs of the local clustering coefficient as a function of the nodes degree for the NC network (left) and the LC network (right). As expected, the local clustering coefficient decreases for nodes with higher degrees. In these two cases the values are notably smaller than in the previous network, and this suggests, especially for the LC network, that nodes with high degrees are unlikely to be within large-dimensional cliques.

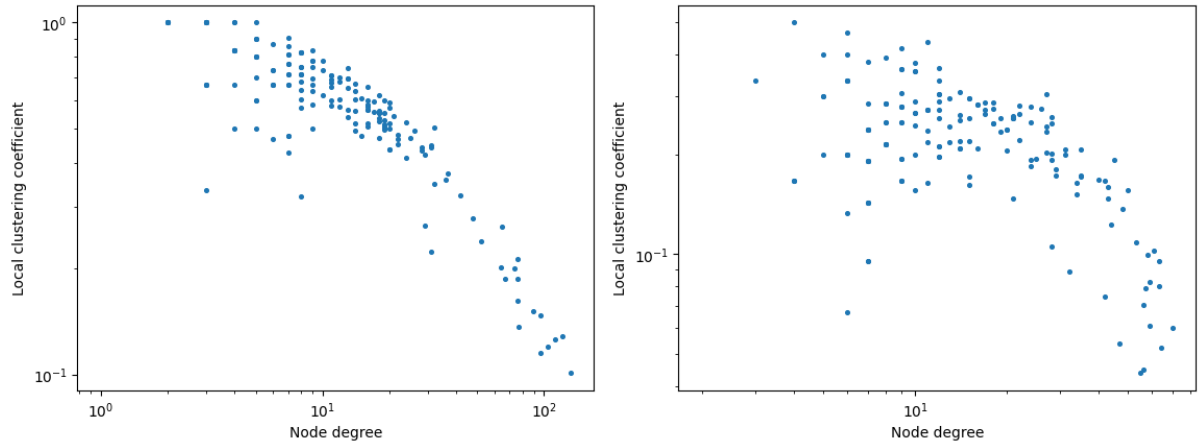

Figure 5: Local clustering coefficient in log-log scale as a function of node degree for the NC (left) and LC (right) networks.

As in Section 5.1, before going through the study of cliques, let's calculate the percentage of edges where the two involved nodes operate in the same sector. We obtain a value around 2% either for NC and LC. Again the value its own does not indicate a tendency for nodes to

be connected with others operating in the same sector. However, what is now interesting is the fact that the value obtained for NC and LC networks is approximately one-fourth of that obtained for the high correlation network. This suggests that even though the sector of the company is not explicitly used in constructing the networks, the information captured by the correlation coefficient appears to be somehow linked to the type of companies. In other words it is more likely for two companies operating in the same sector to have a high correlation coefficient rather than one that is nearly zero or negative. That said, we should expect to find industry-heterogeneous cliques in these two networks.

Finally, we extract the maximum clique from each of the two network. As stated above, given the clustering coefficients, we can expect these two cliques to be much smaller than the one reported in Section 5.1. Indeed the NC network has 30 cliques of dimension 8, whereas the LC network gives rise to 52 cliques of dimension 5. From these data it appears evident the inverse proportionality between the number of maximum cliques and the clique dimension. Again, we randomly select one maximum clique from each set in order to avoid any bias in the choice. Tables 3 and 4 present the names of the companies, their yearly average market capitalization in 2022 (in millions of euros), disclosed revenues for the fiscal year 2022 (in thousands of euros) and industry for the NC and LC networks respectively. From Figure 6 we can also appreciate the different stock daily price trends and price magnitudes of these companies.

Name	Market cap	Revenues	Industry
GAS PLUS	122.55	159938.01	Oil & Gas Exploration and Production
NETWEEK	6.84	23769.13	Consumer Publishing
CIA	6.57	331.41	Real Estate Services
ENERVIT	60.16	70516.41	Food Processing
BASTOGI	84.40	109092.93	Real Estate Rental, Development & Operations
NEWLAT FOOD	259.12	648300.74	Food Processing
AUTOSTRADE MERIDIONALI	112.35	54933.04	Construction & Engineering
TXT E-SOLUTION	157.08	123508.30	Software

Table 3: Stocks inside the clique selected from the NC network.

Name	Market cap	Revenues	Industry
SABAF	239.29	258,165.79	Appliances, Tools & Housewares
AUTOSTRADE MERIDIONALI	112.35	54,933.04	Construction & Engineering
MONRIF	14.69	141,298.70	Consumer Publishing
SAES GETTERS	370.33	220,173.85	Commodity Chemicals
SAIPEM	2,149.77	8,428,514.39	Oil Related Services and Equipment

Table 4: Stocks inside the clique selected from the LC network.

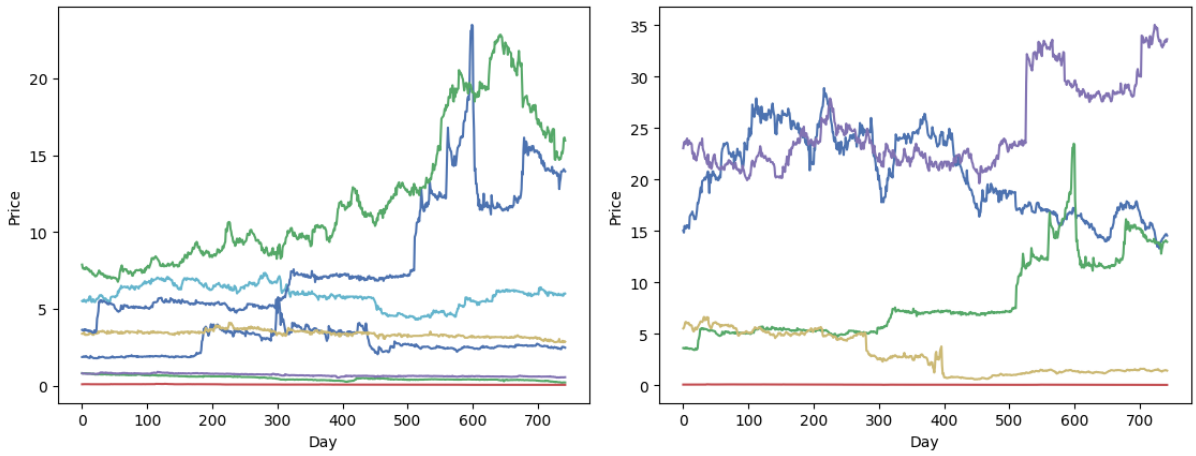

Figure 6: Daily price movements for the cliques selected from the NC (left) and LC (right) networks.

Now that we have identified the two cliques, let's calculate the Jaccard similarity coefficient for each pair of nodes within the two cliques as in Section 5.1. From the sets of similarities we compute the means, obtaining a value of 0.15 for LC clique and 0.28 for NC clique. These two values, on average, indicate low similarity between the nodes in the two cliques, suggesting that these nodes mostly do not share many neighbors outside of those within their respective cliques. See Figure 7 for the similarity matrices from the two cliques.

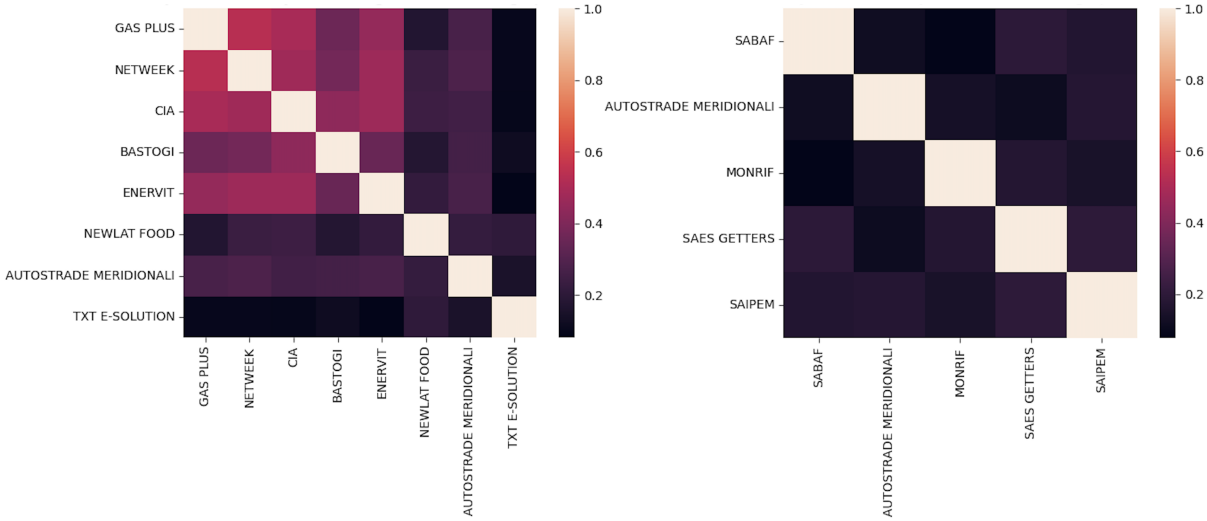

Figure 7: Jaccard similarity matrices for the cliques selected from the NC (left) and LC (right) networks.

5.3 Portfolio optimization application

In this section we propose an estimate of the goodness of asset selection by means of the analysis of cliques of negatively and low correlated network of stocks. We form two portfolios, i.e. two collections of assets, departing from the two cliques analyzed in Section 5.2. These sets are named NC and LC portfolios. In order to assess and compare the performances of these two strategies we opt for a classical mean-variance optimization technique (Markowitz, 1952). Given an initial collection of financial instruments with a specific correlation structure,

this procedure allows to compute the relative weights that maximize (or minimize) a given optimization function. In other words, this methodology aims to tell the portion of wealth to be invested in each asset of a portfolio.

The first step we followed was that of creating the efficient frontier of each portfolio (the blue lines in Figure 5). These are the combinations of assets that guarantee the best, i.e., the fairest, risk-return proportion. The set of weights that compose these frontiers are called "non-dominated" as they provide the highest expected return for a given level of volatility. Then, we choose the maximization of the portfolios' Sharpe ratio as optimization function. The Sharpe ratio is a synthetic index to express the performance of an investment strategy and is defined as

$$SR = \frac{r_p - r_f}{\sigma_p} \tag{7}$$

Where r_p is the expected return of the portfolio, r_f is the risk free rate (we choose the Italian 10-year BTP return of 4.67%) and σ_p is the portfolio volatility. Clearly, a higher Sharpe ratio indicates an investment with better risk-return profile.

The resulting portfolios are marked with a red star in Figure 8.

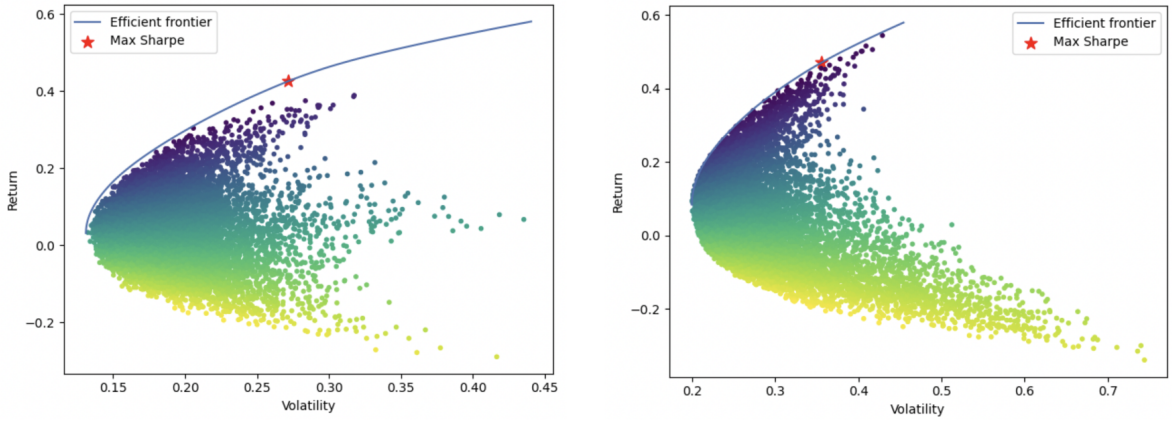

Figure 8: Efficient frontiers for the NC portfolios (left) and LC portfolios (right). Each dot in the graph is a portfolio with different weights (the Sharpe ratio is represented using a color scale). The red stars represent our two portfolios found by maximising the Sharpe ratio.

The Sharpe ratio of the NC portfolio is 1.49, while that of the LC portfolio is 1.27. This means that, on average, the NC portfolio performs better than the LC one.

6 Conclusion

In this section the results obtained for each network are compared and placed inside the theoretical framework we work on.

Comparing the result obtained from the analysis of the three network, it is worth highlighting some findings. To begin with, it is not surprising to see that the NC and the LC networks share the same number of nodes and the same edge density. Indeed, by selecting quantiles of the correlation distribution of the same dimension we allow for the same maximum possible number of edges to take place. Nevertheless, by looking to the ninth decile we witness to a striking change. The number of nodes decrease by 40% while the edge density is almost three times higher. This means that highly correlated stocks are fewer, yet more connected.

This result is confirmed by the γ coefficients of the degree distribution. The high correlation network has a γ coefficient that is half that of the LC and NC ones. As stated above this means that is way more likely to witness to extreme values in nodes degrees. From the linear regression performed on the node degree distribution, it was observed that, in particular, the LC and NC networks can be considered scale-free. This is crucial for our objective of finding cliques because it implies that, even if I remove certain nodes, I can still find cliques since my network maintains the same characteristics. This indicates the absence of decisive nodes that could potentially invalidate the results on the portfolios obtained from cliques.

Moving towards the primary aim of this work, which focuses on portfolio allocation via an exploration of networks' cliques, is important to highlight the significance of the findings related to local clustering. Typically, in networks, the value of a node's clustering coefficient decreases with its degree; it is unlikely that nodes with a high degree have their neighbors connected pairwise, and this is what we observe in our networks as well. However, in the high correlation network, not only is the average clustering coefficient high, but so are those of individual nodes with the highest degree. It is important to note that most nodes with the highest degree in this network correspond to companies operating in the financial sector, so it is normal to expect that their neighbors are also connected, as they operate in a strategical sector. These companies are also the ones that have the highest betweenness centrality, once again underlining their high centrality in that network. As highlighted, the high clustering coefficient and modest density express the tendency of nodes to form cliques or clusters of nodes, even of big dimension. On the other hand, the study of the clustering coefficient in the NC and LC networks indicates a lower tendency of nodes to group together, and thus, nodes with the highest degree are unlikely to be found in the largest cliques but rather in smaller ones. It is interesting to note that in these two networks, companies that were central in the other one are no longer so; in fact, they are not included in the largest cliques. This implies that financial sector companies tend to have few low or negative correlations and, therefore, are almost exclusively highly correlated with the majority of others. The reason behind this is that banks or financial holdings often hold stakes in other companies, and thus, the performance of their stock returns is also influenced by them.

The analysis of the clique composition gives important insights. Indeed the two cliques with higher absolute values of correlation (the HC and NC) exhibit lower heterogeneity than the LC. We can aggregate the results in tables 2, 3 and 4 in order to evaluate the heterogeneity among firms in the dimensions of market capitalization and industry. By estimating the coefficient of variation of the market cap values for the three sets, we obtain the highest value for the LC clique (154.02 against 139.57 and 82.28 for the HC and NC cliques respectively). This means that the LC clique is composed by companies of very different dimensions. The NC clique shows the lowest heterogeneity in firms dimensions, more specifically it is composed by smaller firms. On the contrary, the HC clique is formed by some of the biggest firms in the whole sample to the extent that its smallest company has a market capitalization that doubles that of the biggest company of the NC clique. These results are confirmed when looking at the diversity in industries. In the LC clique each company operates in a different sector, showing the highest degree of heterogeneity. The NC clique presents a good diversification as well since only two stocks share the same industry. The HC clique is instead quite homogeneous, as the firms operating in the financial services macro area, i.e., the industries of banks, Investment Management & Fund Operators and Life & Health Insurance, account for half of the whole clique. This pattern is consistent with the results arising from the betweenness centrality computation. This feature is probably due to the special nature of financial firms. By holding other listed stocks as assets in their balance sheets, their performances seem to be highly connected to those of many other listed companies. These results also confirm the intuition that industry and dimension create a

good conduit for information to spread through the Italian stock exchange and that correlation is a good way to capture it.

The examination of Jaccard similarity enables us to focus on the hypotheses that arise from the correlation between stocks. In the first network, we obtained a high average similarity value from the clique, indicating that nodes with high correlation also have high similarity. In fact, highly correlated companies, such as those in the financial sector, tend to be closely interconnected and share the same neighbors. On the other hand LC and NC cliques exhibit low average similarity values, which is crucial for portfolio construction, as it suggests that the involved companies exhibit diverse characteristics, captured by a low or even negative correlation coefficient.

Finally, the portfolio optimization application provides interesting results in support of the classical portfolio optimization theory. The two values of Sharpe ratio lie below the 2.0 benchmark that usually indicate a portfolio in which it is worth to invest in. Nevertheless, this is a surprising result for such small portfolios. In point of fact, the standard practice in investments is to compose a portfolio made up of at least 10% of the whole targeted market. This means that the ideal portfolio strategy in the FTSE Italia All Share index should have been formed by about 20 stocks. With only 8 stocks we devise a strategy that has an acceptable risk-return profile. Moreover, the comparison between the two strategies is coherent with the mean-variance framework, as the NC portfolio performs better than the LC one. This means that, indeed, when investors agglomerate stocks which prices move in opposite directions, they obtain higher returns at lower risk.

7 Critique

The outcomes achieved from the analysis of the three networks do indeed correspond with the structure of the Italian stock market. Furthermore, the portfolios obtained from the cliques of LC and NC networks show a promising Sharpe ratio. The aforementioned considerations regarding the result suggest that network analysis actually is a valuable tool for portfolio allocation.

Different results could be obtained by allowing for a shorter time window. Indeed, by opting for a longer observation period we safeguard the stability of the correlation structure among firms. Nevertheless, by cutting the time series by approximately 500 trading days, it would have been possible to add four more firms, yet preserving a good correlation stability. This could have yield to slightly different results.

It is worth noting that one of the major caveats of this work is that it heavily relies on the mean-variance theoretical framework. Despite its wide implementation in the financial literature and industry, it has been criticized as too simplistic. More sophisticated results could be obtained by adopting a Bayesian approach and giving more relevance to investors' expectations about the future distribution of returns. Furthermore, by using the Sharpe ratio as an indicator for portfolio performances we could give rise to an endogeneity problem, as it makes use of the same statistics at the basis of the mean-variance model. Thus, by using a more refined proxy for risk (such as conditional value at risk) we could have obtained more robust results.

References

[1] H. Markowitz, "Portfolio selection," *The Journal of Finance*, 1952.

[2] L. Bachelier, "Theorie de la speculation," *Annales scientifiques de l'école normale supérieure*, 1900.

[3] A. Cowels, "Can stock market forecasters forecast?," *Econometrica: Journal of the Econometric Society*, 1933.

[4] E. F. Fama, "The journal of finance," *The Journal of Finance*, 1970.

[5] R. Merton, "An intertemporal capital asset pricing model," *Econometrica*, 1973.

[6] W. Sharpe, "Capital asset prices: A theory of market equilibrium under conditions of risk," *The Journal of Finance*, 1964.

[7] F. Black and R. Litterman, "Global portfolio optimization," *Financial Analysts Journal*, 1990.

[8] E. F. Fama and K. R. French, "The cross-section of expected stock returns," *The Journal of Finance*, 1992.

[9] V. Boginski, S. Butenko, and P. Pardalos, "Statistical analysis of financial networks," *Computational Statistics & Data Analysis*, 2005.

[10] K. T. Chi, L. Jing, and C. Francis, "A network perspective of the stock market," *Journal of Empirical Finance*, 2010.

[11] P. Coletti and M. Murgia, "The network of the italian stock market during the 2008–2011 financial crises," *Algorithmic Finance*, 2016.

[12] N. M., *Networks*. Oxford University Press, 2018.