LIUC
Università Cattaneo

Università Cattaneo Working Papers

# Stock markets as a network: from description to inference

**Marcello Esposito**

Numero 10 - 2021

Collana: Università Cattaneo Working Papers
ISSN: 2532-554X
Università Carlo Cattaneo - LIUC
Corso Matteotti 22, 21053 Castellanza (VA), Tel. 0331-572.282, email: biblio@liuc.it

# Stock markets as a network: from description to inference

Marcello Esposito*

**Abstract**
Among the statistical techniques used to describe the behaviour of the financial markets, one of the most promising is based on the network analysis of the stock market. In this framework, the stock market is represented as a graph with nodes (the single stocks), edges (connections between stocks), and attributes (industry classification, volumes …). The application of network analysis to the stock market is not new, but in previous contributions the market graph has been mainly derived from the correlation matrix of the stock prices. This is a limitation, and the risks are to express in different words what traditional financial econometrics has already said about the returns' distribution. Moreover, if we want to use network analysis not only as a descriptive tool but also as an inference instrument, we need other data than the correlation matrix itself. For this reason, we integrated the analysis and built the market graph with new type of data taken from the observation of the information gathering activity performed by retail investors through the Google's search engine. We focussed the attention on financial crises, when a shock hits the economy in such a profound way that almost all the parameters entering the pricing equation of stocks must be reassessed. Those periods are relatively rare and short. They are characterised by extremely high levels of volatility and correlation. In these moments, searching for new information becomes of paramount importance. And then it is in these moments that we expect to observe more neatly the working of the underlying network.

Keywords: Stock Market Graph; Google searches; Cross-correlation.

JEL Codes: C58, D85, G11

## Introduction

Among the statistical techniques used to describe the behaviour of the financial markets, one of the most promising is based on the network analysis of the stock market. In this framework, the stock market is represented as a graph with nodes (the single stocks), edges (connections between stocks), and attributes (industry classification, volumes …). Depending on the structure of the graph, one can describe the price movements inside the market using typical concepts of network analysis, such as "communities", "hubs", "transitivity" …

The application of network analysis to the stock market is not new. Mantegna (1998) and Boginski, Butenko and Pardalos (2005) derived the market graph from the correlation matrix of the stock price returns and their approach inaugurated a stream of research where different markets and time periods were analysed.[1] However, we

* "Cattaneo" University (LIUC), Quantum Financial Analytics. Email: mesposito@liuc.it

[1] Zhuang et al. (2007) and Huang et al (2009) applied network analysis to the Chinese stock markets. Chi et al. (2010) extended the U.S. stock market analysis using also trading volumes and propose a new approach for selecting stocks for inclusion in a stock index. Bonanno et al (2004) analysed d the NY stock exchange in the 12-year period

strongly believe that the derivation of the market graph from the correlation matrix does not allow to display the full potential of network analysis. And, probably, there are more efficient statistical tools than “communities” and “triadic closures” to discriminate among groups of stocks or measure their level of connectedness. Fortunately, today we can integrate the information set with new type of data that was not available until some years ago. These data are obtained from the social networks or the web search engines and allow for the direct observation of some phases of the investment activity, such as the information search, by individual investors.

Investment decisions, as it is well known, depend on a variety factors such as, for example, the risk tolerance of the individuals, their level of knowledge and their cultural background, their wealth and income stream. But whatever the motivation behind an investment act, the decision is preceded by an activity of information gathering that in our times is mostly done using the Google’s search engine. This search activity reveals the underlying portfolio held by the investor, the branching of the investment decisions as the search activity proceeds, and the reaction to good and bad news. Aggregating the individual search activity, one can gather insights into the dynamics both of “normal” and of “abnormal” investment activity. How do investors react to the “monetary”, “real” or “health” nature of the macroeconomic shocks hitting the financial markets? Which factors drive the investors in the selection of specific categories of stocks? Are the investors contrarians or trend-followers? Are they showing one of the many pathologies described in the behavioural finance literature?

We think that Google searches give the opportunity to answer these questions and network analysis is probably the best statistical tool to organize and explore those data. Specifically, we use Google Trends to collect data about the most popular searches about listed stocks (and the related searches) during the pandemics. Representing the search activity’s results as a network is very “natural”, because, as it is well known, the PageRank algorithm of Google is an evolution of the eigenvector centrality metrics. As we will show, the network structure that emerges is richer than the one that we would obtain from the correlation matrix and confirm the necessity to further investigate the use of this new type of data to enrich our understanding of the working of the stock markets.

1987-1998. Caraiani (2012) studied European emerging countries’ stock markets. Lee et al (2007) applied the methodology to the South Korean stock market [33], whereas the Hong Kong and the UK markets were studied by Li et al (2006) and by Chu et al (2017), respectively.

Financial crises and the stock markets are probably the best periods and the best places to investigate. According to the ancient Greek meaning of "turning or separating point", we define a crisis when a shock hits the economy in such a profound way that almost all the parameters entering the pricing equation of stocks must be reassessed. Those periods are relatively rare and short. They are characterised by extremely high levels of volatility and correlation. Differently from other important segments of the financial markets, where the interference of the central banks can distort the working of the pricing mechanism and destroy the quality of the signals that prices send, the stock markets remain the most precise gauge of the "sentiment" of the investors about the future. In these moments, the information search activity becomes of paramount importance. And then it is in these moments that we expect to observe more neatly the working of the underlying network.

## 1. Some aspects of the stocks' returns distribution

During economic crises, stock markets typically exhibit an increase in volatility and in the average level of correlation among single stocks prices. Whereas the increase in volatility is easily understood because of the higher level of uncertainty about the future caused by a financial crisis, the effect on correlation is often misunderstood and interpreted as a sort of structural break with the past. This happened during the 2007-2008 financial crisis, and it happened again during the 2020 crash caused by the COVID-19 pandemic. The difficulty in understanding the spectacular increase in correlation that we observe during economic crises is probably due to a cognitive bias that tends to emphasize the "real" consequences of economic crises at the expense of the "financial" consequences. However, also if the economic shock originates in the "real" side of the economy and it is asymmetric in its consequences across industries and sectors, the quantity of money printed by the central banks is much more relevant than the underlying companies' fundamentals to determine the "nominal" price of a "real" asset.

In Esposito (2015) the behaviour of correlation during the 2008-9 crisis was explained using a quite simple dynamic CAPM model, calibrated over the previous 20 years (see Appendix 1). Despite the extreme simplicity of the model, it fitted very well the observed dynamics of stock markets' volatility and correlation. However, the scope of the analysis was very narrow and limited to the major international stock market indices. Moreover, the new role of central banks in inflating asset bubbles was still in its infancy. Today, we have a new major financial crisis to investigate in a deeper way and

we have more than 10 years of new data to understand how the central banks changed the way in which stock markets behave. So, it is worthwhile to go deeper and analyse the stock markets behaviour at a more granular level, i.e. at the level of single stocks.

## 1.1 The distribution of the returns' correlation

The database we used comprises single stock prices starting in 1963 and ending in July 2020. Taking long time series of stocks' returns can be troublesome because of possible structural breaks in their statistical distribution due to the deep changes occurred in the socio-economic environment and in the financial structure. Between 1963 and 2020 we observed the man walking on the Moon, the fall of the Soviet Union, the surge of East Asia and China as the new World growth engine. As far as the financial structure, the changes have been radical: the end of Bretton Woods, the new European currency, the changing role of central banks, the derivatives revolution, just to name a few. It would then be quite surprising if the statistical properties of the stock market have not been affected by these cataclysmic events.

Returns and their correlations can be measured at different time intervals. Typically, one uses the following time intervals: 1 day, 3 days, 1 week, 1 month. Which interval to choose? There exists a trade-off between the precision of the estimates and the noise that one takes on board by reducing too much the length of the resampling time interval.

In the first part of this section, we want to give a bird's eye view on the correlation distribution and then we divide our sample in 5 decades (1963-1969, 1970-1979, 1980-1989,1990-99, 2000-2009, 2010-today), resample the observations on different non-overlapping time intervals (1Day, 3 Days, 1 week, 1 month) and calculate the log returns and their correlation matrix.

In the second part of this section, instead, we will divide our sample on a quarterly basis and we will analyse the correlation distribution quarter by quarter. This is because we want to focus on periods of financial crises and as we will show those periods are quite short. In this case our sub-samples will be constituted by approximately 65 observations. It is not possible in this type of analysis to resample at a frequency lower than 1 week, because the number of observations would be reduced to just 3 if we consider, for example, monthly returns.

In Fig. 1-2-3 (see the section "Figures" at the end of the paper), we show the estimated density function[2] of the correlation of log returns on a 1 day, 3 days, and 1 week periods in the 6 decades from 1963 to 2020. In practice, we extract from the correlation matrix the upper triangle, eliminate the diagonal elements, and vectorize the resulting triangle matrix.

As one can see, the density is skewed (to the right) and quite asymmetrical. In some decades (the '70s and the last 20 years), it also shows a sort of bimodality that visually tend to disappear at the weekly interval and afterwards. Since the financial crises caused by a macro-factor tend to generate periods of very high correlation, this is probably the legacy of financial regimes prone to extreme instability. In the '70s this turbulence was caused by the end of Bretton Woods, the oil shock and the inflation surge. The last 20 years saw instead the abrupt end of the "goldilocks economy" caused by the subprime meltdown of the financial system in 2008 to which followed the gigantic amount of money pumped through the unorthodox monetary policies engineered by the central banks of the developed World.

The different role of the monetary policy in those two periods can be appreciated observing the location of the highest hump in the estimated density function. Whereas in the '70s the highest hump of the distribution is on the right, in the first and second decade the highest hump is on the left. The '70s were turbulent years but at that time monetary policy was not targeting the stock market level. During the last 20 years, instead, the unwanted effect of the extraordinary monetary policy measures implemented to keep the moribund financial system afloat has been the one of inflating and not pricking the speculative bubbles.

The presence of a pronounced skewness in the distribution of the correlation is not a particularly surprising finding when working with correlation and it is well known in the statistical literature that this phenomenon tends to underestimate the value of the mean if one simply takes the average of the sample values. To correct for this bias, Fisher (1921) envisaged a transformation of the correlation coefficient that produces a symmetric distribution and rapidly converges to Normality as the sample size increases, the so-called z-transform.[3] This is very useful to obtain unbiased estimates of the mean and can be used to build confidence intervals and test hypotheses.

[2] The density curve is estimated using the *Kernel Density Estimation* (KDE) method as implemented in the *Statistics* library of *Python*.

[3] The so-called z-transform consists in:
$z=1/2\ (\ln(1+\rho))/(\ln(1-\rho))=\operatorname{arctanh}(\rho)$
Once the empirical distribution of z is obtained, one needs simply to apply the inverse of the function and calculate

As one can appreciate from Table 1 (see the section “Tables” at the end of the paper), the ‘90s have been a period of relatively low volatility and low correlation. From the point of view of active asset management, this is the best possible environment to produce “alpha” with bottom-up investment strategies based on the analysis of the single stocks. On the contrary, the ‘70s and the first two decades of the new millennium have been the perfect environment for “passive” strategies and top-down macro strategies.

In the ‘90s diversification alone could have reduced the riskiness of relatively concentrated portfolios, while after the dot.com bubble explosion and the Lehman Brothers default only macro strategies and diversification with other asset classes could have succeeded in obtaining a significant reduction in riskiness of portfolios. These facts can help explaining the decline of active mutual funds and the surge in ETF that have changed the shape of the asset management industry in the last 20 years.

## 1.2 The correlation behaviour during financial crises

Financial crises determine spikes in volatility and correlation. Obviously, volatility changes continuously and there is no reason to think that its “real” value has remained constant when the economic environment has changed so much since our time series start, in 1963. However, when we talk about financial crises the spikes consist in a doubling or tripling of the volatility level with respect to the previous period of calm. During the last big crises following the Lehman Brothers default or the Covid pandemics, the volatility reached levels even of 70-80%! It remained at those levels only for a few days but, as a rock hitting the surface of a calm pond, it took many weeks before the ripples created by the impact subsided.

In order to investigate the behaviour of correlation during financial crises we divided the sample in calendar quarters, and we considered non-overlapping 3-days log returns. Fig. 4 shows both the relationship between the volatility of the S&P500 and the average correlation among the stocks.

The top 10 spikes in correlation are illustrated in table 2. As one can see, they occurred during the most turbulent quarters of the recent history of the stock market: the crash of ’87, the eurozone crisis and the before mentioned Lehman Brothers and Covid crises. Just to put the numbers in the right perspective, the average levels of volatility and

the estimate of the mean, the confidence intervals, etc:
ρ=tanh(z)

correlation over the entire period were 11.9% and 13,24% respectively. Except for some quarters at the end of the Bretton Woods era, they were characterised by double digit crashes in the market prices.

The change in the correlation distribution is evident if we compare the empirical distribution estimated during one of these quarters with the empirical distribution estimated in "normal" times. How to choose the "normal" time to compare with the "crisis" time? Since financial crisis are caused by a "shock" and "shocks" cannot be anticipated, we will use the quarters before the crisis unfolded to see the impact of the shock. To see the ripple effect, we will see also the impact on the quarter after the shock impacted.

Whereas all the crises are quite similar in their immediate impact on volatility and correlation, the major crises cause long lasting ripple effects on the volatility and correlation structure. This difference is clearly visible in the case of the first crisis depicted in Figure 5 (the '87 crash) and the other two (the Lehman Brothers default and the Covid crises). While in the first case, the correlation distribution came back in just a quarter to the previous shape and location, in the last two cases the distribution after one quarter is still quite different from the "normal" one.

In the end this is not particularly surprising also if "theoretical" models presuppose the capability by economic actors to immediately calculate the consequences of a shock hitting the economy. However, there is a big difference between a shock caused by a presumed change in monetary policy and a shock caused by a completely new and unexpected event, such as a pandemic burst.

## 2. The stock market as a network

The fact that certain groups of stocks should be more intertwined than others is an old idea, that affected both the practice and the theory of investing. Probably, the most famous and ubiquitous is the one based on the "industry" classification of stocks. If a company produces hats and hats go out-of-fashion, that company can be the most efficient hats' producer, but it will disappear as the other peer companies unless it diversifies away from hats. Different classification methodologies are based on the distinction between "Value" and "Growth" stocks, "Small Caps" and "Big Caps", "Domestic" and "Multinational". More recently, in the so-called ESG approach, one uses also classifications based on the attention paid by companies to the well-being of the stakeholders (workers, communities, …) and the environment.

The presence of different factors in the systematic component of the stock returns generated the literature on multi-factor CAPM models. Statistical tools like principal components analysis of the correlation matrix have been extensively employed to test the relevance of the existing classification criteria to discriminate among stocks behaviour and to build new techniques for managing portfolios and their riskiness. Statistical analysis helped to understand that the classifications based on the industrial sectors or on certain aspects of the balance sheet of a company are useful only in specific occasions to forecast the stock price co-movements. With some notable exceptions for quasi-monopolies induced by strict State's regulation or tolerated because of high barriers to entry, rigid classification criteria are neither suitable to understand the way in which stocks comove nor particularly useful in improving the portfolios' diversification. Eventually, it is the community of the investors that, based on the ever-changing competitive landscape and on the continuous process of technical innovation, can see the most sensitive relationships between the different listed company and understand how and if a certain news or shock impact them singularly or as a group.

The representation of a system as a network requires the definition of "nodes" and "edges". Nodes can be different agents (banks, hedge funds …) or instruments (stocks, bonds …) in the financial markets, whereas the "edges" are the relationship between such agents or instruments. If we want to analyse the connectedness between instruments that are listed on a public market, Pearson's correlation coefficient is probably the most common and easiest way to measure the degree of connectedness, but obviously, depending on the nature of the nodes, one can use other metrics such as Kendall-tau or Spearman rank-correlation.

To clarify some network's concepts that we will use extensively, let us make an example. If we want to represent a company as a network of individual workers, we define each worker as a "node". The age, the gender, the duties, the position inside the organigram … represent the "attributes" of the node. We can investigate the relationship between the workers using the emails exchanged in a specific time interval. The sender is the "source" node and the recipients are the "target" nodes. The set of the target nodes defines the "edges" of the "source" node and its number is called the "degree" of the source node. In graphical terms the nodes will be represented as the vertices of the Network Graph and the edges as the lines connecting one node to the other. Once all the company is mapped in this way, one can start to analyse the characteristic of the network. For example, one can see how

many “communities” exist, if there are workers that are more “central” to the network than others, and more generally the degree of connectedness of the structure. Then, one can use these results “operationally” to understand if, for example, these relationships changed overtime or if they conform to the organigram and the expectation of the management.

The representation of the stock market as a Network was investigated firstly by Boginsky et al. (2005). Since the covariance between stock prices is the epiphany of the underlying network, the authors of these early investigations employed Network Analysis as a different point of view for a statistical analysis of the properties of the correlation matrix. However, a statistical analysis of the stock market, which is only based on the prices’ data, risks to express in a different language the same concepts already investigated by more traditional techniques such as Principal Component Analysis.

At the beginning of the millennium there were no social networks to scrape and the concept of “open data” was almost absent. Today, things have radically changed. Just to make some example, Google allows for downloading trending searches in the different geographical locations and some trading platforms such as Robinhood make public the positions held by their clients almost in real time or at least with a frequency useful for interpreting stock market movements.

The availability of these new type of data makes the Network Analysis of the stock market much more interesting. We can in fact test hypotheses and models about the way in which information is collected by economic actors and propagated across the network. In other words, the network analysis tools can then be applied not only to the epiphany of the network (the price movements) but also to the underlying activities executed by the participants (research, communication, investment) gaining a deeper understanding of the functioning of the financial system.

## 3. Using the correlation matrix to identify the network

In the representation of the stock market as a network, the single listed stocks are the nodes, and the edges are the stocks that have a relationship with the node. Since the property of listed stocks that we, as investors, are mostly interested in is their price movements, it is quite natural to look at the correlation matrix of the stocks’ returns to

find the pair of stocks with the higher coefficients and add to the list of the edges of a node the stocks that show a correlation coefficient higher than a certain threshold.[4]

To find the correct value of the threshold, Boginski et al (2005) suggest using the fact that complex graphs such as the one employed to describe a social network, or the world wide web exhibit a "power law" in the Degree distribution. The degree of a node is its number of edges, i.e., the number of connections with other nodes, and the degree distribution is the probability that a node has a certain number of connections.

## 3.1 Power law distribution

Power law models (or Pareto distributions) have been extensively used in many fields of economics and social sciences, well before network analysis.[5] Pareto (1896) was probably the first scientist to apply the power law and he did it by describing the upper tail of the income distribution. He found that the number of people with an income or wealth S greater than (a large) X is proportional to $(1/X)^{\gamma}$, as in eq. (1).

$$\text{Prob}(S>X) = (m / X)\gamma = k\, X^{-\gamma} \quad (1)$$

where:

$m>0, \quad X>m, \gamma>0$

The eq.(1) describes the form of the counter-cumulative distribution function.[6] The density function corresponding to this counter-cumulative distribution is equal to:

$$f(X) = k\, \gamma\, X^{-(1+\gamma)} \quad (2)$$

The parameter γ is called the power law exponent (PL exponent or the Pareto exponent)[7], whereas *m* is a threshold level (for example, a minimum income level in the original Pareto analysis) above which the tail's distribution shows the power law behaviour. Empirically, the value of *m* might be determined as the top 1% or 5% percentile of the distribution of the underlying variable or, alternatively, one could use a

[4] The statistical analysis behind the identification of the nodes and the edges of a financial network can be much more sophisticated than the one of applying an high-pass filter to the correlation matrix. Billio et al (2012) for example used extensively principal component analysis and Granger causality to determine the connections between the nodes of a financial network.
[5] Gabaix (2016)
[6] Since power laws are used to model tail behaviour, it is natural to express them in this way, also if traditionally statistical distributions are shown in the cumulative form: Prob (S≤X).
[7] It has to be said that some authors call (1+ γ) the power law exponent.

searching algorithm that stops as soon as the tail's distribution shows the power law.[8] As we will see, in the financial literature the latter approach has been adopted.

The lower is γ, the higher is the probability of finding "outliers" in the distribution. In statistical terms, we could say that the tails of the distribution become fatter. If the exponent γ is equal to 1 we obtain the so-called Zipf law (1949), that was used to describe the occurrences of the most frequent words in a language. More recently, distribution based on the power law have been applied to describe the city sizes or, as we said before, the functioning of the World Wide Web.

It must be noted that the PL exponent is "scale-free", i.e. it does not depend on the unit of measure of S, so that the proportion of large occurrences with respect to smaller ones remains always the same. For example, if we are describing the size of large cities, its value does not change if we consider medieval cities (where the population was of the order of the thousands) or modern cities (where the population is of the order of the millions).

An interesting property of the power law concerning the so-called "rank-size rule", which states that for large values of *n*:

$$i/n = (m / X)^{\gamma} \qquad (3)$$

This rule establishes a relationship between the rank, *i*, of an observation, and its size, *X*.[9] If we take the log of both sides of equation (3), we obtain a linear relationship:

$$y = \alpha - \gamma x + \varepsilon \qquad (4)$$

where:

$y = \ln(i)$

$x = \ln(S_i)$

In the special case of γ =1, we obtain the Zipf law.

Equation (4) can be easily estimated via OLS, also if Gabaix and Ibragimov (2008) have shown that the ranking procedure creates positive autocorrelation in the

[8] Beirland et al. (2004)

[9] Suppose that we draw a random sample of n elements of a certain variable (income, city size, national GDP …) and we order the n observations in descending order of magnitude $\{S_1, S_2, S_3, \ldots, S_i, \ldots S_n\}$, where the subscript *i* denotes the rank of the observation so that $S_i > S_j$ if i < j. Whatever the underlying probability distribution that generated the n sample, if a certain observation, S, is the i-th largest, i.e. $S = S_i$, there are just *(i-1)* observation (out of *n*) larger than S and *(n-i+1)/n* smaller or equal than S. This means that *(n – i+1)/n* and *(i-1)/n* are the cumulative and counter-cumulative empirical distribution of the sorted observations. Now, let's consider the underlying distribution and see if something interesting emerges. In fact, using eq. (1), we know that Prob $(S > S_i) = (m / S_i)^{\gamma}$ .

residuals. They suggest to correct the asymptotic standard error of the slope estimated via OLS multiplying it by $(n/2)^{-0.5}$ and to insert a shift term in the log of the rank in order to correct the small sample bias.[10]

## 3.2 Finding a PL in the correlation distribution.

The idea of Boginski et al (2005) is to try different values of the threshold value, θ, calculate the degrees of each node and check if the upper tail of the degrees' distribution is linear in log. The higher the threshold, the higher the probability of linearizing the log-distribution curve. They found that a value of θ equal to 0.6 determines a linear relationship in the degrees distribution and the exponent coefficient is almost equal to 1, i.e. the degrees of the most popular nodes follow a Zipf law.

We should note that their data sample was based on 500 trading days in 2000-2002 and comprises 6546 financial instruments traded in the US stock markets. These instruments include ETF or other collective investment vehicles, and this represents a mistake in our opinion because it alters the correlation matrix with a sort of double counting and then bias the shape of the underlying network structure. It is then useful to replicate their analysis with a cleaner database and with different threshold levels and let those levels depend on the time-varying average correlation that we have already illustrated.

For this reason, we divided our sample in quarters and in each subsample, and we tested different value of θ, starting from the 0.5 and 0.6 values identified in Boginski et al (2005). Then, we added two other values, based on the changing "average" correlation over time. In particular, we first calculated the average correlation of each stock with the S&P500, an index representing the market factor. Then we divided the distance between the average correlation and its maximum value (i.e. 1) by four and we added this value 2 and 3 times to the average correlation to obtain another couple of dynamic thresholds. Summing up, we test four values of θ in each subsample:

| | |
|---|---|
| $\theta = 0.5$<br>$\theta = 0.6$<br>$\theta_L = \rho_{mkt} + (1- \rho_{mkt})2/4$<br>$\theta_H = \rho_{mkt} + (1- \rho_{mkt})3/4$ | (5) |

[10] Ln(i-s) instead of Ln(i), where the optimal value of s is ½, according to Gabaix-Ibragimov (2008)

To test for the existence of a power law, we do not limit ourselves to a "visual" inspection of the linearity of the log-rank-log-size distribution or to the value of the $R^2$ of the regression, as it is usually done in the previous financial literature. We use the Gabaix-Ibragimov (2008) test, which is based on the following procedure.

Define $S^* \equiv \text{cov}((\ln S_j)^2, \ln S_j)/ 2\,\text{var}(\ln S_j)$, and estimate by OLS the parameters of the following equation:

$$y = \alpha - \gamma x + \mu z + \varepsilon \qquad (6)$$

where:

$y = \ln(i - 0.5)$

$x = \ln(S_i)$

$z = (\ln(S_i) - S^*)^2$

As one can see, eq (6) differs from eq (4) because of the quadratic term, represented by the new *z* variable, that should not be present if the PL holds true. The Gabaix-Ibragimov (2008) states that the null hypothesis of an exact PL is rejected at 5% level if:

$$\left|\frac{\bar{\mu}}{\bar{\gamma}^2}\right| > 1.95(2n)^{-0.5} \qquad (7)$$

As one can see, the best parametrization for the threshold (in terms of number of times that it generates a PL not rejected by the Gabaix-Ibragimov test) is based on the $\theta_L$. However, we must note that there are many quarters (36) where it is not possible to run the test, because the threshold is too high and select too few correlations to make the analysis meaningful. Moreover, since the value of the average correlation with the market index is around 20% most of the times, the value of $\theta_L$ is very near to the 60% threshold most of the times: its average value is 63.5% when the PL is accepted and 63.8% when the PL is rejected. This implies that the value of 60%, used by Boginski et al (2005) appears appropriate, in general.

Using our time-varying calibration of the threshold levels, we observe a sort of structural break in the network structure of the stock market around the Lehman Brothers default. Before the Lehman Brothers default, the network's structure could be identified with relatively low levels of the threshold, as in Boginsky et al (2005), as one can see from Fig. 6. Afterwards, the identification deteriorates for low levels of the threshold parameter and this happens especially during crises periods.

After the Lehman Brother default, we observe that, when the market experiences a crisis, a higher threshold level performs better than a lower one. This is very well represented in Fig.7 where we depicted the log distribution taking as an example a "normal" quarter such as the 2019Q2 and a much more turbulent quarter such as the 2020Q2 that followed the beginning of the COVID pandemics.

The cause of this structural break is in our opinion caused by the regime shift in monetary policy in response to the financial meltdown of the subprime crisis. The huge amount of liquidity pumped in the economic system changed the behaviour of the stock market increasing structurally the correlation among stocks.

## 3.3 The topology of the network

Once we have defined the nodes (i.e., the single stocks) and the edges of each node (i.e., the list of stocks that showed a correlation higher than the threshold), we can analyse the network structure of the stock market. As the word "network" implies, the analysis will be dedicated to investigating the connections and the transmission of shocks between nodes.[11] In other words, we will investigate the connective shape or topology of the stock market.

It must be noted that the type of network that we derived from the correlation matrix is by construction "un-directed". The correlation coefficient, in fact, does not tell anything about the direction of the shock propagating from a node to the other. In more traditional econometric terminology, we say that correlation between two variables cannot be interpreted as a proof of causality from one variable to the other. In an undirected network, connections between nodes are always symmetric.

A way of representing the network is by the so-called Adjacency matrix, $A$, whose element $A_{i,j}$ is equal to 1 if nodes $i$ and $j$ are connected and it is equal to 0 if they are not. The diagonal elements are set equal to 0, i.e. $A_{i,j}=0$. The matrix is square and symmetric since the network is undirected. This implies that its eigenvalues ($\lambda_i$ for $i=1,2 \ldots n$, where $n$ is the number of nodes and the dimension of the matrix) are all real numbers and its eigenvectors ($v_i$ for $i=1,2 \ldots n$) are orthogonal among themselves. We will use the convention to sort the eigenvalues (and their associated eigenvectors) in descending order, i.e., $\lambda_1 > \lambda_2 > \ldots. > \lambda_n$.

[11] We use the algorithms of the NetworkX library for Python.

### 3.3.1 Some basic measures of the network shape

Let us start from a recent crises period, the second quarter of 2020, following the Covid pandemics. First, we want to know how many nodes are in the network (i.e. how many stocks have at least one correlation coefficient with another stock higher than the "threshold" level), how many edges (i.e. what is the number of connections of the nodes), and the average number of edges per node. The number of connections of a node is called "degree". In terms of the Adjacency matrix, the degree of node *j* is the sum of the values of the *j*-th row (or of the *j*-th column).

Now, we can enter the most interesting part of the investigation and analyse the network as a connective shape. We can then see if nodes cluster together or are if they are relatively spread out. We can ask how "long" it takes for a shock to propagate across the network. If there are "sub-communities" inside the market and how populous they are. What are the most central nodes inside a community (the so-called "hubs") and what are the nodes that connect one community to the other.

The first interesting metrics to calculate is the density of the network, i.e. the relationship between the effective number and the maximum theoretical number of edges. In theory every node could be connected to another node by an "edge", as it could happen in a small social network such as the family. In practice, in large networks the number of edges is much smaller than the maximum possible. In the case of 2020Q2, the stock market exhibited a network density of 2%. This number might seem "small" but is 50 times bigger than the density coefficient observed by Boginski et al (2005) with a 0.6 threshold for the 2000-02 period.

We can expect that the density of a network also affects the velocity of propagation of shocks across a network. An interesting metrics in this respect is the so-called diameter of a network. Consider all the shortest path that connect a node to any other nodes in the network and the take the longest path of all the shortest path. This max-min metrics is the "diameter". When a network is segmented in sub-communities without edges bridging them, the diameter is referred to the largest sub-community. In our case, the diameter of the stock market is 4 and then it is slightly shorter than the one of "human" social networks. [12]

Finally, another concept that can be used to describe the shape of the network is the one of triadic closure. The idea is that, if node A is connected to node B and node B is

[12] A famous example of diameter is the one that says that every individual on the earth is just 6 steps away from the President of the United States.

connected to node C, it is likely that node A is connected to node C, too. In this case, we say that the A, B, C triangle is closed. Note that the transitivity is a possibility, not a certainty. With the index of triadic closure, we then calculate all the possible "triangles" and we check how many triangles are effectively closed. The ratio of the triangle closed over the total number of triangles goes from 0 to 1. In our case, the index of triadic closure was equal to 6.1%.

### 3.3.2 Centrality analysis

The three most common ways of defining the "importance" of a node define the three most common centrality measures: degree, betweenness, and eigenvector centrality. From a network perspective, the importance of a node is in some way related to the connectedness inside the network. The most important nodes are often referred to as the hubs of the network.

The first and simplest way to measure the importance of a node is related to the number of its edges. The number of edges is called the degree of a node and this measure of centrality is then called "degree centrality".

A more complicated but probably more appropriate way of defining the importance of a node consists in evaluating not only the number but also the quality of the connections. For example, a node that is connected to nodes with many connections is more "central" than a node that is connected to nodes with few connections. The PageRank algorithm that is powering the Google search engine and decides which webpages get to the top of its search results is based on this idea. Obviously, there are many ways of defining the "quality" of the connections but all of them starts from basic idea of attributing a score to each node. Connecting to a "target" node with a high score contributes to increase the score of the "source" node.

With the eigenvector centrality measure, the score attributed to a node is equal to the corresponding element of the first eigenvector of the Adjacency matrix representing the network.[13] The intuition behind the use of the first eigenvector is that it can be considered the most explicative linear representation of the Adjacency matrix.[14]

[13] The eigenvector centrality scores are expressed on a [0,1] scale. If we denote with, $v_1$, the eigenvector of A with the greatest eigenvalue, $\lambda_1$, we know that the elements of $v_1$, thanks to the Perron-Frobenius theorem, are all non-negative. Since the elements of an eigenvector are defined up to a multiplicative scalar, they can be arbitrarily rescaled so that the maximum value of the elements of $v_1$ is 1.

[14] The readers accustomed to traditional financial econometrics have probably recognized that this scoring methodology is very similar to the methodology used to decompose the covariance matrix in principal components. The principal components are linear combinations of the underlying variables and are orthogonal among themselves. The associated eigenvalues are the portion of variance explained by that specific linear combination. The first principal component is then the linear combination that explains the largest part of the total variance.

Betweenness centrality (expressed on a [0,1] scale) looks instead at all the shortest paths that pass through a node. In contrast to a hub, we can try to find nodes that act as a broker between distant parts or clustered parts of the network. In this context, a node is important not because it has lots of connections but because it connects different parts of the network.

As one can see, the different centrality measures produce sets of hubs that are similar, but not identical. Which centrality measure to choose? Since these measures have different purposes, the choice depends on the purpose of the investigation.

### 3.3.3 Community detection

As far as our analysis is concerned, we are interested not only in seeing which nodes are more central than others, but also to detect if there are clusters of nodes that are more interconnected inside themselves. These clusters are denominated "communities" in network analysis. The most popular method to detect communities is based on the relative density of the nodes. One starts with a measure of how fractious the network is, which is called modularity, and from this measure one partitions the network in class modules, i.e., in communities, to which the single nodes are attributed. The algorithm that we will use to detect communities is the so-called "greedy modularity maximization" developed by Clauset-Newman-Moore (2004).

Once the communities are found, an interesting analysis consists in verifying if those communities are characterized by some "fundamental" characteristics of the objects represented as node of the network. For example, in the case of the stock market, we would like to investigate if the traditional classification criteria can be associated to the communities that network analysis reveals.

Limiting the analysis to communities constituted by at least 3 nodes, we detected 7 communities in the second quarter of 2020. Using the industry classification of the listed companies, we obtain the following "industrial" characterization of the 7 communities. Note that communities are sorted according to their numerosity and the number of the community is just a label showing the rank of the community.

In Table 6, we showed the industrial composition weighted by market cap of each community. To facilitate interpretation of the numbers, in the last column we showed the industrial composition of the entire market (i.e., of all the listed companies that were selected for our analysis), whereas in the last two rows are indicated the total market cap and the number of members of the communities.

To see if there is an industry characterization of the communities, we searched the community where an industry shows its maximum weight. For example, the Consumer Discretionary industry reaches its maximum capitalization weight in the Community "1", whereas Financials represents a third of the capitalization of the community "7". We can see that, with the notable exception of communities 0 and 6, this methodology helps in discriminating the communities where an industry is more "over-represented".

#### 3.3.3.1 Communities' dynamics in the short run.

The fact that communities change over the short and the long run should not be surprising, because the relationship among stocks depends both on structural trends and fads that together determine the long-term and short-term market dynamics. If we repeat the previous analysis using data of the first quarter of 2020, we see in Table 5 that the number of communities found by the algorithm has increased to 9. Moreover, there has been a generalized increase in their numerosity and market capitalization.

Comparing Table 7 and 6, it is interesting to note that the most numerous community (the "0") detected in the first quarter of 2020 is not characterized by any specific industry. The first quarter of 2020 was characterized for the first two months by "normal" market conditions. Then, the COVID pandemics hit the economy in March, determining a generalized crash in the stock prices. In a certain sense, community "0", one of the most representative in terms of number and weight of stocks, reflects this fact.

Notwithstanding this strong common movements of the stocks, there are still communities with a clear single industry characterization in the first quarter of 2020. Materials is now coupled with Consumer Discretionary and Energy is coupled with IT. Communications Services instead decoupled from Real Estate and Utilities, with these 3 industries forming two distinct communities.

#### 3.3.3.2 Communities' dynamics in the long run.

To go deeper in the analysis, we repeated the community detection analysis for all the quarters of our 1963 sample and we investigated the dynamics of the communities over time. The criterion that we used to create a correspondence between communities in two different quarters is based on the number of shared nodes, not weighted by their market capitalization.

We developed a greedy-matching algorithm that works on the matrix of the intersections between the communities detected in two subsequent quarters. Once the

intersection matrix has been built, the algorithm detects the maximum element of the matrix and its position (row and column number) inside the matrix. The coordinates of the position determine the two corresponding communities. The algorithm then eliminates the row and the column of the maximum element and searches for a new relative maximum. The position of this new element defines the next two corresponding communities. The recursion stops when the intersection matrix is exhausted. In Table 8, we show the results of this "matching" algorithm in describing the dynamics of the 2020Q2 communities.

As we can see, the community number 2 of 2020Q2 is matched with the community number 5 of 2020Q1 which makes sense also from a "fundamental" point of view. In fact, we have seen in the previous tables that this community is mostly characterized by "Industrials" stocks. With the communities matching 5-2 instead there does not seem to be a clear fundamental factor behind.

Another interesting phenomenon is that the further back we go, the lower becomes the matching capability of the algorithm. This is not so much surprising because it reflects the extreme dynamicity of stock markets: top management turnover, technical innovation, changes in consumers tastes … fuel the creative destruction process of competitive, free markets.[15]

## 3.4 Filtering-out the common "market" factor.

It is well known in the statistical literature on the stock market that there is not a single. specific macro systematic factor affecting the single stocks behaviour. For sure, the "market" factor is the dominant one because it summarizes the general economic conditions, the investors' sentiment, the impact of the monetary policy on the "nominal" side of the pricing system. However, companies belonging to certain sectors (financials, pharmaceuticals, utilities …) should show a higher degree of correlation among themselves. And the same is true for big cap with respect to small caps, for companies interpreting certain investment themes (robotics, biotech, Covid-vaccine …), for companies showing specific balance sheet characteristics (value vs growth, high yield vs investment grade …), for companies with a specific geographical or national distribution of their revenues and costs (domestic vs multinational, developed vs emerging countries …).

[15] If we look at big cap stocks, we know that their behavior and their perception by investors changes dramatically over time. One can think of Microsoft that not so many years ago was considered a "dead" hi-tech stock and did not co-move with other tech-darlings such as Apple or Amazon, and now is back in favor.

We then filtered out the common market by regressing the single stock return on the delta log of the index representing the market (in our case, the S&P500). For every *i*-th stock we then ran an OLS regression and we calculated the residuals:

$$\widehat{\varepsilon_i} = r_i - [\hat{\alpha} + \hat{\beta} r_{mkt}] \quad (8)$$

We then calculated the correlation matrix of the residuals and we repeated the above analysis on the latter matrix. We tried different value for the threshold parameters, θ, to find a power-law distribution in the right tail of the degrees' distribution.

It must be noted that, differently from the "normal" correlation matrix, a power-law appears only for relatively high levels of θ (0.6 or higher). However, also for the residuals' correlation matrix, 0.6 is the threshold value that performs uniformly better in generating a power law distribution.

For lower values of 0.6, a bell-shaped curve appears in the log-rank-log-size. This is probably the signal that only a small subset of companies is affected by this secondary network connections. The typical shape with a low and a high value of the threshold are illustrated in the Figure 8.

### 3.4.1 Communities detection based on the residuals' matrix.

If we use the residuals of the regression of the stock-returns over the market-return, we obtain the correlation matrix of the factors orthogonal to the market. The number of nodes that we need to consider seeing a power-law appearing is lower than in the "normal" case. The number of communities detected is just 3 and a sort of super-sector classification seems to appear. Financials and Information Technology dominate the community 0, whereas Communication Services, Health Care, Energy, Industrials, Utilities dominate the community 1. As far as the last community is concerned, Consumer Discretionary represents almost the 50% of it.

## 4. Using Google searches to identify the network structure

Up to now, we have used only the correlation matrix to build the network representation of the stock market. The availability of data concerning the actions performed by the investors can help investigate the network structure of the stock market in a much more meaningful way, adding new "information" to the one contained in the correlation matrix.

In fact, if we use the correlation matrix to detect the network structure, as Boginski et al (2005), we can analyse from a different perspective with respect to traditional clustering techniques the same information contained in the correlation matrix. But the information is always the same.

A way to enlarge the information set consists in obtaining new data, for example from the Google searches. Since retail investors do not have access to expensive database and news' aggregators such as Bloomberg or Thomson-Reuters, this probably implies that what we will unveil is prevalently related to the investment activity of small investors.[16] The financial press underlines the emergence of new trading platforms, such as Robinhood, that were technologically designed for the new generation of native digitals and whose interest was boosted by the fact that during the pandemics sport betting and casino gambling was impossible. Whatever the reason, also if the amount of money played by retail investors was modest compared to the asset under management of the big investment houses, the compounding effect generated by the attention received from the press and by the replicating-anticipating strategies set up by professional investors has been impressive and succeeded in moving not only small caps but also mega-caps such as Tesla or Apple, as their late August reaction to the announced stock-split suggests.[17]

The Network obtained from Google searches (G-Network, from now on) is also structurally Network derived from the Correlation matrix (C-Network, from now on) because "related searches" have a direction, from the original search (the source) to the related ones (the target). Correlation coefficients are, instead, directionless. In the terminology of network analysis, the C-Network can only be of the "undirected" type, whereas the G-Network is of directed type. One can always analyse a directed network as an undirected one, but the reverse is not true.

## 4.1 Building the network.

We considered the Google searches made in USA during the each of the quarters spanning the period 2018Q1-2020Q2. For every ticker (the node) listed on an American stock markets, we looked up if that ticker was the subject of searching

[16] Probably, the carnage caused by the dot.com bubble explosion contributed to the retreat of retail investors and the rise of collective investment vehicles and advisory services dedicated to retail investors (actively managed mutual funds, ETF, …). We needed to wait for a new generation of retail traders to see their come-back on the market and this probably happened in 2020 with the reaction to the COVID pandemics, also if we need more data and time to confirm the effective importance of retail investors in the determining one of the most impressive bull markets of the history.

[17] According to Citadel Securities, retail investors have accounted for 25% of the stock market's activity in the second quarter of 2020, while in 2019 they accounted for 10% (Business Insider, 9 July 2020).

activity by the user of the Google search engine over the same period. If this is the case, we defined that ticker as a "node" of the market graph. Since our goal is to find "connections" between stocks, for every ticker[18] inside our database we looked at the so-called "related searches".[19] If in the list of the related searches appeared some tickers, we identified them as the target nodes and we registered the relation as an "edge" of the "source" node.[20] Since this is a directed network, each node can have in-bound and/or out-bound edges, depending if it is the target or the source of the connection.

First, we tested for the presence of a power law in the edges' distribution. Using the Gabaix-Ibragimov (2008) test, we rejected the Null hypothesis only in 2018Q3. In all the other quarters the power law cannot be rejected at the 5% level.

## 4.2 The topology of the network

The Network obtained from Google (the Google Network or G-Network, from now on) shows the typical power-law shape in the distribution of the degrees (edges) of the nodes. It is important to note that this is a "natural" characteristic and it has not been derived by the application of some filter as we did with Network derived from the Correlation matrix (the Correlation Network or C-Network, from now on) where we tried different thresholds until the power-law distribution appeared.

The average number of edges per node (i.e. the average degree of the network) is equal to 5.8, while the network density is 0.7%. It is interesting to note that the Google Network gives us a picture of the "hubs" that is nearer to the common perception of which companies set the pace of the market in 2020. In Table 11, one can see that, according to the edge centrality metrics, Tesla, Boeing, Apple … are the stocks more central to the network. If we compare this result with the one obtained from the Correlation Network, where some regional banking stock was at the centre of the network, one can understand how much more insightful the analysis is based on the observation of the investors activity.

With large networks, the typical graph picture is quite difficult to interpret. The graph of figure 10 was built using the "Degree Centrality" measure. We decided then to plot only the legend of the top 5 hubs by degree centrality, colouring the nodes according to

[18] We opted for the ticker and not for the company name, because we need to filter out related searches that might have originated for collecting information about the products or services provided by the company.

[19] In Google Trends jargon, these searches are called "related searches") operated by the same persons that did the original search.

[20] As we said, the list of the node-targets is cleaned from all the words that are not tickers or that might be tickers but, since they are a replica of popular English words (USA, NEW, OLD …), might produce spurious centrality measures.

their industry membership. The size of the nodes is proportional to the number of edges. If, instead, we use the "Betweenness Centrality" measure we obtain the picture of Figure 11. As one can see, in the top 5 hubs, Microsoft is substituted by AutoNation, a company of just 5 billion of market capitalization but that during the post COVID rally more than duplicated its share price, outshining many hi-tech companies.

## 4.3 Community detection

We applied the different algorithm of community detection. and for each one of them we decided to focus only on the 10 largest ones. In Table 12, we listed the top 10 communities by numerosity and their top 30 members. As one can see, we voluntarily stopped to 10 the number of the communities and this is already a big difference with respect to the Correlation Network, where the number of the communities detected is much smaller also if the members of the latter Network are more numerous.

### 4.3.1 Industries and communities

In order to verify if the communities were formed according to an industry classification[21] criterion (i.e. the investor searching for an "utilities" ticker will make related searches of other "utilities tickers") we first tested if the industry distribution of the community was statistically different from the industry distribution of the entire network.

We used the Epps-Singleton test[22] that according to the literature[23] has a greater power than the Kolmogorov-Smirnov in detecting if two samples are generated by the same underlying distribution. Moreover, it does not assume that the distribution is continuous. The test rejects always the null hypothesis that the communities are generated by a common "industrial" matrix. This authorizes us to check if some industries are more represented than others inside the communities. In Table 13 we show the industry composition of the communities and we highlight the communities where each industry has the highest weight.

### 4.3.2 Price co-movements of the members of communities

A second thing that we want to test is if belonging to the same community means a higher correlation level. If we define the members of a community as the "insiders" and we define the other members of the community as the "outsiders", we expect that the

[21] For the industry definition we used the GICS classification.
[22] Epps and Singleton (1986)
[23] See for example Goerg, Kaiser (2009)

insiders shows an average a level of correlation that is higher than the one that they show with the outsiders. Using the correlation as a gauge to confirm the "true" membership, we can also calculate how many members of a community should be expelled from the community because they show levels of correlation with the "insiders" that are lower than the one that they have with the outsiders.

As one can see from Table 14, there are 3 communities (0, 3, and 6) where the average level of correlation within the community is lower than the correlation with the outsiders. However, we can see that in general there is a certain coherence between membership and correlations.

### 4.3.3 Building an efficient portfolio using the G-network communities.

Detecting communities inside a financial network can be interesting not only from a pure statistical point of view, but also on a practical ground to build more efficient investment strategies. We can consider the communities detected inside the G-network as "assets" or "portfolios" and see how we could build an efficient portfolio if it were possible to buy ETFs replicating their statistical behaviour. An index representing a set of stocks can be equally weighted or capitalization weighted. In the following we use the equally weighted methodology that probably reflects better than the capitalization weighted the way in which retail investors build their investment portfolios. Figure 12 shows the result in terms of the efficient frontier estimated over the period (2018-01-01, 2020-06-30).

To evaluate the results obtained from communities belonging to the G-network, we can compare with a more traditional classification such as the one based on industries' membership. As we did with communities, we build an equally weighted price index for each of the industries and we consider them as an "asset". We build the efficient frontier using the same methodology as before and the result is showed in Figure 13.

The efficient frontiers have been obtained using the CAPM model for estimating the expected returns of the respective assets (industries and communities). In Table 15, we show the numerical structure of the efficient frontiers depicted in Fig.12-13.

One can see that with "G-communities", the estimated expected returns of the optimal is higher than with "industries". This is counterbalanced by a higher volatility, but only partially. In fact, the Sharpe ratio is higher for the optimal portfolios built with G-communities than with Industries. If there were investment vehicles, such as ETF, that replicate the behaviour of the G-communities, this result suggests that an investor

might achieve higher Sharpe ratios building a portfolio based on them than building a portfolio using standard Industry-based ETF.

Obviously, this result, as many other in empirical Finance, depends on the sampling period and should be tested on a truly out-of-sample experiment. In fact, we used the communities detected in the last quarter of our dataset and we backward evaluated over the last 10 quarters the performance of the associated portfolios. The correct procedure should be the opposite one: use the communities detected in quarter "t" to evaluate the performance of the corresponding portfolios in quarter "t+1". Then, in quarter "t+1", one reshuffles the original portfolios according to the new communities eventually detected and evaluate the performance in quarter "t+2".

## 5. Comparison between the C-Network and the G-Network representations of the stock market

We can observe that the number of the relations (edges) between stocks (nodes) is lower in the G-Network than in the C-Network and this determines a less dense network structure. This result makes sense if we think that most of the correlation between stocks depends on a common macro-factor, the "market". Think about the impact of an increase in interest rates by the FED on stock prices. There is first a common impact via the discount factor, that does not require by investor any specific investigation. Then, for some stocks (banks, leveraged companies …), the investors need to evaluate the specific impact. This requires a specific analysis and then generate an explicit "search" by the investors. This fact probably also explains the reason why the G-Network appears more "transitive" than the C-Network. The triadic closure index stands at 21%, signalling that there is a greater level of transitivity and clustering.

We can compare these communities with the ones that we obtained partitioning the network via the correlation matrix of the log returns. We use the same methodology employed to characterize the dynamics of the Correlation Network communities over time. We then build the intersection sets between the two Networks' communities and we match the communities with the largest intersection. The results are depicted in Table 17.

As one can see from Table 17, there is a certain degree of correspondence between the most numerous communities in the two Networks. The communities labelled "0" in the two Networks (i.e. the most numerous), for example, share 38 nodes (tickers). The

number of shared nodes can appear as small, but we can also note, by comparison with Table 5, that the order of magnitude of the intersection sets is equal to the one observed for the communities derived only by the correlation matrix in 2020Q2 and 2020Q1 quarters.

## 6. The dynamics of the G-Network

We repeated the analysis contained in section 4 for the last 10 quarters of our data sample, i.e. from 2018Q01 to 2020Q02. We focussed on the eigenvector measure of centrality and we observed the change in the network topology.

In Table 18, we show the usual metrics for describing the topology of the network. As one can see, those metrics remains quite stable during the 2018 and 2019 "normal" times. The average number of connections per node (i.e. the degree) remains in the 3.08-3.30 range, the density slightly decreases until the end of 2019 as do the transitivity (the triadic closure). The picture changes dramatically in 2020 and especially in the second quarter. Not only we see a dramatic increase in the number of nodes, but the network starts to look denser and more transitive. The investors' community appears like a living body under attack: new connections are activated and the level of interaction among the different elements of the network increases.

It should be noted that the Covid-crisis appears as something completely different from a previous "minor" crisis like the one that occurred in the last quarter of 2018, when already pumped-up markets worried about the "normalization" of the monetary policy in the US and in Europe. This is a remarkable fact, because if one looks only at the stock prices, probably the COVID crisis would not be classified as a "crisis", but as a simple "blip" in a parabolic speculative bubble. Notwithstanding the worst economic shock of the post-WWII period and the worst health crisis of the last 100 years, central banks engineered an expansion of the stock indices that registered historical maximum both in absolute and in relative terms (i.e. with respect to traditional multiples).

To investigate if the higher level of connectedness changes the relative importance of the nodes constituting the network, we analysed the dynamics of the most important "hubs" of the network. We defined as "top hub" the node that, in at least one quarter, ranked among the top 10 in terms of eigenvector centrality.

As one can see from Table 19, out of 24 node that belonged to the top 10 in at least one quarter (the shaded cells inside each column of Table 14), only 4 of them (AAPL, AMZN, MSFT, and TSLA) did so consistently in each quarter. Not surprisingly we are

talking of the stars of the technology sector. Other two popular technology stocks, FB and NVDA, appeared in the top 10 in all quarters but one. Whereas it is surprising to observe that GOOG never succeeded to enter the top 10. In Fig.14 we show the rise (until 2020Q1) of MSFT and TSLA in terms of "eigenvector" centrality, and the relative fall of FB.

The top 10 chart is populated not only by the hi-tech elite. As for the music charts, we find also "pop" investment theme that for a limited time period raise the generalized curiosity of the investors but then fade away (LYFT, ROKU, SPCE) or that retrench to a specific niche (BYND). But what is most interesting is that we also find companies that enter the Top 10 as the result of pure investment strategies. Look at what happened in 20Q02. We find in the Top 10 companies that were hammered in March 2020 because of the COVID pandemics (AAL, UAL, CCL, DAL … all belonging to the airline sector) and that were bought in Q2 according to a pure contrarian strategy of "buy-the-dip".

The fact that those strategies were mostly implemented by retail investors is confirmed by numerous financial press articles and by the evidence on the activity on the famous Robinhood online trading platform.[24] In Table 20, we listed the Top10 Hubs on the G-network and we looked at how they ranked in terms of "popularity" among Robinhood users. Popularity is measured as the variations in the outstanding individual positions from the beginning to the end of the quarter. As one can see, the correspondence is strong notwithstanding the fact that Tesla, Apple and Microsoft show relatively "lower"[25] popularity rank because they are historically among the core holdings of US retail investors.

If we extend the comparison to the first 100 hubs of the G-network, the Kendall and the Spearman rank correlation coefficients are 42% and 61%, respectively.

Finally, we can verify if the price performance of the top hubs is superior to the market index (the S&P500). We then built equally weighted portfolios of the top 100 hubs and calculated their performance in each quarter of the sample. To investigate the source of the overperformance, we built also other four equally weighted portfolios, consisting of the top 25 hubs (1 quartile), of the hubs that were ranked 26-50 (2 quartile) etc. The results are in Table 21.

[24] The data are taken from https://robintrack.net/, a website that keeps track on an hourly basis of how many Robinhood users hold a stock over time. The timeseries start in 2018 and ends in August 2020, when Robinhood dismissed their public API.

[25] In this respect, one must take into consideration that on Robinhood were listed 8595 tickers.

As one can see, the Top 100 portfolio overperform the S&P500 in 7 quarters out of 10. The differential in performance is asymmetric in favour of the Top 100, in the sense that the magnitude of the overperformances is significantly larger than the magnitude of the underperformances. If we look at the quartiles' portfolios, we can see that the major cause of the differential is attributed to the behaviour of the first and second quartiles.

The overperformance of the Top Hubs with respect to the S&P500 is enormous during both quarters of 2020. Thanks to the performances of companies like Nikola (+534% in 20Q2) or Moderna (+143%) the 2 quartile overperform the S&P500 by an astonishing +80%!

As far as communities are concerned, we observe a relative stability of the membership across time at least for the most important nodes. Moreover, the most important nodes belong to the community labelled "0", which is also the most numerous. The members of this community appear in the Top100 Hubs from a minimum of 68 times (in 19Q03) to a maximum of 92 times (in 20Q02).

# 7. Conclusions

The representation of the stock market as a Network allows to gain some new insights about the statistical properties of the equity returns. We identified the network structure using two different procedures. The first is based on the correlation matrix of the stocks' returns and the second is based on the Google searches. The first procedure, followed by the literature on the applications of network analysis to the financial markets, consists in testing if a power law in the correlation distribution appears when we apply a simple high-pass filter, i.e., a threshold, on the value of the correlation coefficients necessary for a "connection" to be established. Using different time periods, we found that a value of 0.6 for the filter estimated in the seminal paper of Boginski et al (2005) is correct also when applied to the last 15 years of data. Once a network has been identified, one can apply the battery of statistical tools of network analysis to study, for example, the degree of connectedness of the network or the communities inside it. We found that during "crises" periods the network becomes more active and more diversified. This represents our first innovative contribution to the literature.

However, the use of the correlation matrix to derive the network structure does not enrich the information set and the results that we obtain are simply expressing in new

words what could have been obtained with more traditional statistical tools. For this reason, we decided to go to the source of the underlying network structure and analyse the initial phase of the investment process, when the investor gather information. This type of data did not exist until some years ago, but today, thanks to Internet and the use of search engines, it is possible to gather observations about the searching activity of individual investors. We then looked at all the tickers inside our database and used the "related searches" provided by Google Trends analytics to unveil the underlying network structure.

With the direct observations of the searching activity related to each single "node" (ticker) of the network, the power law in the distribution of the connections emerges naturally, without filtering techniques. Moreover, the number and the meaningfulness of the "communities" increases. For example, we find that belonging to a community implies a higher price correlation with the insiders than with the outsiders. This result can be used for building efficient portfolios and we found that using Communities as portfolios-building blocks one might achieve higher Sharpe ratios than using traditional classification systems, such as the one based on Industries.

We also verified if some metrics concerning the centrality of the nodes of the network are reflected in what investors actually did with their "real" money during the pandemics. Specifically, we verified if the most central nodes (the so-called "hubs") of the network were also the most popular stocks on the iconic Robinhood platform. We found a strong relationship and we found also the fact that the most central nodes have consistently overperformed the market. As a by-product of our major investigation, we found also evidence of a change in retail investing attitude. At least in the first part of the pandemics, retail investors didn't panic and "bought the dip", exactly the opposite of what we are used to hear from the asset management industry.

Albeit one should test the discriminating properties of search-based Networks over longer time horizons and in truly out-of-sample experiments, we believe that this result is promising and encouraging for the use in empirical finance of the new source of data that one can collect from Internet. The network analysis can then be used for inference purposes, not only as an alternative descriptive statistic.

# Tables

Table 1 – Summary statistics about the correlation distribution

| | TIME INTERVAL: 1 DAY | | | | | |
|---|---|---|---|---|---|---|
| | **1963-1969** | **1970-1979** | **1980-1989** | **1990-1999** | **2000-2009** | **2010-2020** |
| **mkt volatility** | 8.3% | 13.2% | 15.4% | 13.6% | 21.4% | 17.1% |
| | **correlation** | | | | | |
| **average** | 16.2% | 29.2% | 9.9% | 5.5% | 14.3% | 17.1% |
| **5% conf. Int.** | 10.9% | 25.1% | 5.6% | 1.1% | 9.9% | 12.9% |
| **95% conf. Int.** | 21.4% | 33.1% | 14.3% | 9.9% | 18.6% | 21.3% |
| | **TIME INTERVAL: 3 DAY** | | | | | |
| | **1963-1969** | **1970-1979** | **1980-1989** | **1990-1999** | **2000-2009** | **2010-2020** |
| **mkt volatility** | 8.0% | 12.6% | 14.8% | 11.6% | 17.0% | 13.8% |
| | **correlation** | | | | | |
| **average** | 20.8% | 34.0% | 14.3% | 7.3% | 15.8% | 18.5% |
| **2.5% percentile** | 14.1% | 28.9% | 8.7% | 1.6% | 10.1% | 13.1% |
| **97.5% percentile** | 27.2% | 39.0% | 19.9% | 13.0% | 21.3% | 23.9% |
| | **TIME INTERVAL: 1 WEEK** | | | | | |
| | **1963-1969** | **1970-1979** | **1980-1989** | **1990-1999** | **2000-2009** | **2010-2020** |
| **mkt volatility** | 10.1% | 16.0% | 16.1% | 13.4% | 20.1% | 16.4% |
| | **correlation** | | | | | |
| **average** | 24.3% | 37.3% | 16.3% | 8.8% | 18.0% | 19.0% |
| **2.5% percentile** | 14.4% | 29.6% | 7.8% | 0.2% | 9.6% | 10.8% |
| **97.5% percentile** | 33.8% | 44.4% | 24.5% | 17.3% | 26.2% | 27.0% |

Table 2 – Top 10 quarters by average correlation

| rank | quarter | average correlation | S&P500 performance | S&P500 volatility |
|---|---|---|---|---|
| 1 | 70Q02 | 46.9% | -19.3% | 18.3% |
| 2 | 66Q03 | 40.8% | -10.6% | 13.1% |
| 3 | 11Q03 | 40.4% | -15.5% | 27.2% |
| 4 | 65Q02 | 38.3% | -2.5% | 10.1% |
| 5 | 87Q04 | 37.5% | -24.5% | 53.0% |
| 6 | 10Q02 | 37.5% | -12.5% | 21.1% |
| 7 | 20Q01 | 37.0% | -20.7% | 42.6% |
| 8 | 08Q04 | 34.8% | -22.2% | 45.8% |
| 9 | 69Q01 | 33.4% | -2.3% | 9.7% |
| 10 | 71Q03 | 32.3% | -1.4% | 11.8% |
| average | | 13.2% | 1.9% | 11.9% |

Table 3 – Number of quarters when the PL hypothesis is accepted or rejected (according to different value of the threshold)

| | Accept | Reject | n.d. |
|---|---|---|---|
| θ=0.5 | 68 | 155 | 7 |
| θ=0.6 | 163 | 48 | 19 |
| θ(L) | 173 | 21 | 36 |
| θ(H) | 136 | 9 | 85 |

Table 4 – The topology of the network (2020Q2)

| **Number of "nodes"** | **1312** |
|---|---|
| Number of "edges" | 17336 |
| Average degree of a node | 26 |
| Density | 2% |
| Diameter | 4 |
| Triadic closure | 6.1% |

Table 5 – Top 30 Hubs according to different centrality measures (2020Q2: 3days not-overlapping log-returns matrix)

| Edge centrality | | | Eigenvector centrality | | Betweenness centrality | |
|---|---|---|---|---|---|---|
| | Number of edges | Sector | | Sector | | Sector |
| ACCENTURE PUBLIC LIMITED COMPANY | 806 | Software & IT Services | AMERIS BANCORP | Banking Services | ACCENTURE PUBLIC LIMITED COMPANY | Software & IT Services |
| ASBURY AUTOMOTIVE GROUP, INC. | 789 | Specialty Retailers | ALLIANCE DATA SYSTEMS CORPORATION | Professional & Commercial Services | ASBURY AUTOMOTIVE GROUP, INC. | Specialty Retailers |
| AMERIS BANCORP | 783 | Banking Services | AECOM | Construction & Engineering | ALLIANCE DATA SYSTEMS CORPORATION | Professional & Commercial Services |
| ALLIANCE DATA SYSTEMS CORPORATION | 781 | Professional & Commercial Services | ARCHER-DANIELS-MIDLAND COMPANY | Food & Tobacco | AUTOMATIC DATA PROCESSING, INC. | Software & IT Services |
| ARCOSA, INC. | 763 | Construction & Engineering | ASBURY AUTOMOTIVE GROUP, INC. | Specialty Retailers | ACADIA HEALTHCARE COMPANY, INC. | Healthcare Providers & Services |
| AECOM | 760 | Construction & Engineering | ACCO BRANDS CORPORATION | Professional & Commercial Services | ACCO BRANDS CORPORATION | Professional & Commercial Services |
| ARCHER-DANIELS-MIDLAND COMPANY | 759 | Food & Tobacco | ADIENT PUBLIC LIMITED COMPANY | Automobiles & Auto Parts | ARCOSA, INC. | Construction & Engineering |
| ACCO BRANDS CORPORATION | 751 | Professional & Commercial Services | ACI WORLDWIDE, INC. | Software & IT Services | AMERIS BANCORP | Banking Services |
| ACADIA HEALTHCARE COMPANY, INC. | 750 | Healthcare Providers & Services | ACCENTURE PUBLIC LIMITED COMPANY | Software & IT Services | AMEREN CORPORATION | Multiline Utilities |
| AMERICAN AIRLINES GROUP INC. | 745 | Passenger Transportation Services | ARCH CAPITAL GROUP LTD. | Insurance | ACI WORLDWIDE, INC. | Software & IT Services |
| ADVANCED EMISSIONS SOLUTIONS, INC. | 739 | Professional & Commercial Services | ATLANTIC CAPITAL BANCSHARES, INC. | Banking Services | AMERICAN AIRLINES GROUP INC. | Passenger Transportation Services |
| AUTOMATIC DATA PROCESSING, INC. | 734 | Software & IT Services | Aaron's, Inc. | Specialty Retailers | ADVANCED EMISSIONS SOLUTIONS, INC. | Professional & Commercial Services |
| ADIENT PUBLIC LIMITED COMPANY | 729 | Automobiles & Auto Parts | ARCOSA, INC. | Construction & Engineering | ARCHER-DANIELS-MIDLAND COMPANY | Food & Tobacco |
| ALCOA CORPORATION | 728 | Metals & Mining | ACADIA HEALTHCARE COMPANY, INC. | Healthcare Providers & Services | AECOM | Construction & Engineering |
| ACI WORLDWIDE, INC. | 728 | Software & IT Services | ALCOA CORPORATION | Metals & Mining | AMERISOURCEBERGEN CORPORATION | Healthcare Equipment & Supplies |
| ANALOG DEVICES, INC. | 721 | Semiconductors & Semiconductor Equipment | AUTOMATIC DATA PROCESSING, INC. | Software & IT Services | ANALOG DEVICES, INC. | Semiconductors & Semiconductor Equipment |
| ADVANCE AUTO PARTS, INC. | 716 | Specialty Retailers | ADVANCED EMISSIONS SOLUTIONS, INC. | Professional & Commercial Services | ALCOA CORPORATION | Metals & Mining |
| ARCH CAPITAL GROUP LTD. | 710 | Insurance | AMERICAN AIRLINES GROUP INC. | Passenger Transportation Services | ADIENT PUBLIC LIMITED COMPANY | Automobiles & Auto Parts |
| AMEREN CORPORATION | 710 | Multiline Utilities | AMEREN CORPORATION | Multiline Utilities | Aaron's, Inc. | Specialty Retailers |
| Aaron's, Inc. | 706 | Specialty Retailers | ANALOG DEVICES, INC. | Semiconductors & Semiconductor Equipment | ADVANCE AUTO PARTS, INC. | Specialty Retailers |
| AMERISOURCEBERGEN CORPORATION | 695 | Healthcare Equipment & Supplies | AMERISOURCEBERGEN CORPORATION | Healthcare Equipment & Supplies | ATLANTIC CAPITAL BANCSHARES, INC. | Banking Services |
| ATLANTIC CAPITAL BANCSHARES, INC. | 693 | Banking Services | ADVANCE AUTO PARTS, INC. | Specialty Retailers | AUTODESK, INC. | Software & IT Services |
| AUTODESK, INC. | 674 | Software & IT Services | AEGION CORPORATION | Construction & Engineering | ARCH CAPITAL GROUP LTD. | Insurance |
| AEGION CORPORATION | 610 | Construction & Engineering | AUTODESK, INC. | Software & IT Services | AEGION CORPORATION | Construction & Engineering |
| AGCO CORPORATION | 24 | Machinery, Equipment & Components | AGCO CORPORATION | Machinery, Equipment & Components | AGCO CORPORATION | Machinery, Equipment & Components |
| AAR CORP. | 24 | Aerospace & Defense | AAR CORP. | Aerospace & Defense | AAR CORP. | Aerospace & Defense |
| ALASKA AIR GROUP, INC. | 24 | Passenger Transportation Services | ALASKA AIR GROUP, INC. | Passenger Transportation Services | ALASKA AIR GROUP, INC. | Passenger Transportation Services |
| BARNES GROUP INC. | 24 | Machinery, Equipment & Components | BARNES GROUP INC. | Machinery, Equipment & Components | BARNES GROUP INC. | Machinery, Equipment & Components |
| CIT GROUP INC. | 24 | Banking Services | CIT GROUP INC. | Banking Services | CIT GROUP INC. | Banking Services |
| CURTISS-WRIGHT CORPORATION | 24 | Aerospace & Defense | CURTISS-WRIGHT CORPORATION | Aerospace & Defense | CURTISS-WRIGHT CORPORATION | Aerospace & Defense |

Table 6 – Industrial features of the top 7 communities (2020Q2: 3days not-overlapping log-returns matrix)

| | communities | | | | | | | | market |
|---|---|---|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | |
| **Communication Services** | 33% | 14% | 3% | 13% | 3% | 26% | 4% | 1% | 15% |
| **Consumer Discretionary** | 6% | 30% | 25% | 15% | 9% | 19% | 13% | 11% | 17% |
| **Consumer Staples** | 0% | 4% | 1% | 0% | 12% | 1% | 14% | 2% | 3% |
| **Energy** | 9% | 2% | 1% | 0% | 12% | 4% | 1% | 5% | 4% |
| **Financials** | 8% | 14% | 23% | 16% | 23% | 6% | 20% | 33% | 16% |
| **Health Care** | 11% | 3% | 2% | 12% | 1% | 18% | 2% | 2% | 7% |
| **Industrials** | 13% | 13% | 21% | 7% | 11% | 5% | 15% | 20% | 13% |
| **Information Technology** | 5% | 15% | 10% | 31% | 13% | 15% | 11% | 17% | 15% |
| **Materials** | 2% | 3% | 5% | 3% | 10% | 6% | 11% | 7% | 4% |
| **Real Estate** | 1% | 0% | | 0% | 0% | 0% | | 0% | 0% |
| **Utilities** | 11% | 1% | 8% | 3% | 6% | 2% | 8% | 1% | 5% |
| **total market cap** | 3,686,673 | 4,768,357 | 2,113,263 | 3,297,817 | 1,891,117 | 897,817 | 812,072 | 1,070,103 | 18,537,219 |
| **number of members** | 234 | 218 | 213 | 205 | 123 | 121 | 101 | 97 | 1312 |

http://dx.doi.org/10.25428/2532-554X/10

Table 7 – Industrial features of the communities (2020Q1: 3days not-overlapping log-returns matrix)

| | communities | | | | | | | | | | market |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | |
| **Communication Services** | 7% | 1% | 6% | 7% | 3% | 1% | 13% | 41% | 5% | 3% | 8% |
| **Consumer Discretionary** | 12% | 12% | 4% | 5% | 11% | 18% | 14% | 4% | 20% | 12% | 11% |
| **Consumer Staples** | 17% | 3% | 0% | 8% | 3% | 1% | 2% | 8% | 1% | 54% | 6% |
| **Energy** | 2% | 3% | 6% | 5% | 2% | 4% | 4% | 0% | | 1% | 3% |
| **Financials** | 9% | 21% | 12% | 5% | 19% | 6% | 18% | 15% | 17% | 1% | 13% |
| **Health Care** | 22% | 6% | 13% | 24% | 22% | 5% | 12% | 12% | 13% | 7% | 13% |
| **Industrials** | 8% | 7% | 5% | 15% | 13% | 16% | 9% | 11% | 7% | 9% | 10% |
| **Information Technology** | 15% | 40% | 46% | 21% | 19% | 44% | 11% | 8% | 27% | 10% | 29% |
| **Materials** | 5% | 6% | 4% | 2% | 1% | 3% | 3% | 0% | 9% | 1% | 3% |
| **Real Estate** | 0% | 0% | 0% | 0% | 0% | 0% | 1% | 0% | | 0% | 0% |
| **Utilities** | 4% | 1% | 3% | 6% | 7% | 2% | 14% | 0% | 1% | 2% | 4% |
| **total market cap** | 3,176,654 | 4,275,444 | 4,039,089 | 2,465,665 | 2,199,545 | 4,151,027 | 1,891,877 | 2,403,490 | 755,343 | 512,307 | 25,870,441 |
| **number of members** | 277 | 227 | 221 | 219 | 206 | 205 | 166 | 129 | 76 | 53 | 1779 |

Table 8 – The dynamics of the communities of 2020Q2
(communities obtained from 3days not-overlapping log-returns matrix, quarter by quarter)

| | | communities correspondance based on largest intersection | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **2020Q2** | **community** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| 2020Q1 | community | 0 | 1 | 5 | 3 | 4 | 2 | 6 | 7 |
| | shared nodes with 2020Q2 | 33 | 30 | 26 | 24 | 15 | 12 | 8 | 8 |
| 2019Q4 | community | 1 | 0 | | | | | | |
| | shared nodes with 2020Q2 | 52 | 36 | | | | | | |
| 2019Q2 | community | 0 | 1 | | | | | | |
| | shared nodes with 2020Q2 | 81 | 7 | | | | | | |
| 2018Q2 | community | 0 | | | | | | | |
| | shared nodes with 2020Q2 | 84 | | | | | | | |
| 2017Q2 | community | 0 | 2 | 1 | 3 | | | | |
| | shared nodes with 2020Q2 | 32 | 31 | 31 | 24 | | | | |
| 2016Q2 | community | 3 | 1 | 0 | 2 | 4 | | | |
| | shared nodes with 2020Q2 | 28 | 27 | 23 | 17 | 14 | | | |
| 2015Q2 | community | 0 | | | | | | | |
| | shared nodes with 2020Q2 | 81 | | | | | | | |

Table 9 – Number of quarters when the PL hypothesis is accepted or rejected
(according to different value of the threshold)

| | **Accept** | **Reject** | **n.d.** |
|---|---|---|---|
| θ=0.5 | 53 | 147 | 30 |
| θ=0.6 | 182 | 8 | 40 |
| θ=0.8 | 110 | 9 | 111 |

Table 10 – Industrial features of the top communities (2020Q2: 3days not-overlapping residuals)

| | **communities** | | | **market** |
|---|---|---|---|---|
| | **0** | **1** | **2** | |
| Communication Services | 3% | 33% | 15% | 18% |
| Consumer Discretionary | 5% | 6% | 44% | 22% |
| Consumer Staples | 1% | 2% | 5% | 3% |
| Energy | 8% | 10% | 2% | 6% |
| Financials | 40% | 17% | 8% | 19% |
| Health Care | 4% | 9% | 1% | 5% |
| Industrials | 5% | 10% | 7% | 8% |
| Information Technology | 26% | 4% | 11% | 13% |
| Materials | 5% | 3% | 4% | 4% |
| Real Estate | 0% | 0% | 0% | 0% |
| Utilities | 3% | 5% | 3% | 4% |
| total market cap | 2,935,124 | 3,841,650 | 4,885,002 | 11,661,776 |
| number of members | 311 | 298 | 296 | 905 |

Table 11 – Top 30 Hubs of the G-Network in 2020 (according to different centrality measures)

| Edge centrality | | | Eigenvector centrality | | Betweenness centrality | |
|---|---|---|---|---|---|---|
| **Hub** | **Number of edges** | **Sector** | **Hub** | **Sector** | **Hub** | **Sector** |
| TESLA, INC. | 806 | Automobiles & Auto Parts | THE BOEING COMPANY | Aerospace & Defense | THE BOEING COMPANY | Aerospace & Defense |
| THE BOEING COMPANY | 789 | Aerospace & Defense | TESLA, INC. | Automobiles & Auto Parts | TESLA, INC. | Automobiles & Auto Parts |
| APPLE INC. | 783 | Computers, Phones & House | APPLE INC. | Computers, Phones & House | AUTONATION, INC. | Specialty Retailers |
| MICROSOFT CORPORATION | 781 | Software & IT Services | Carnival Corp | Hotels & Entertainment Se | Carnival Corp | Hotels & Entertainment Se |
| Carnival Corp | 763 | Hotels & Entertainment Se | MICROSOFT CORPORATION | Software & IT Services | APPLE INC. | Computers, Phones & House |
| AMERICAN AIRLINES GROUP INC. | 760 | Passenger Transportation | AMERICAN AIRLINES GROUP INC. | Passenger Transportation | FACEBOOK, INC. | Software & IT Services |
| AMAZON.COM, INC. | 759 | Diversified Retail | AMAZON.COM, INC. | Diversified Retail | AMERICAN AIRLINES GROUP INC. | Passenger Transportation |
| ADVANCED MICRO DEVICES, INC. | 751 | Semiconductors & Semicond | UNITED AIRLINES HOLDINGS, INC. | Passenger Transportation | ADVANCED MICRO DEVICES, INC. | Semiconductors & Semicond |
| DELTA AIR LINES, INC. | 750 | Passenger Transportation | DELTA AIR LINES, INC. | Passenger Transportation | THE ALLSTATE CORPORATION | Insurance |
| ZOOM VIDEO COMMUNICATIONS, INC. | 745 | Software & IT Services | ZOOM VIDEO COMMUNICATIONS, INC. | Software & IT Services | THE ESTEE LAUDER COMPANIES INC. | Personal & Household Prod |
| UNITED AIRLINES HOLDINGS, INC. | 739 | Passenger Transportation | ADVANCED MICRO DEVICES, INC. | Semiconductors & Semicond | SOUTHWEST AIRLINES CO. | Passenger Transportation |
| FACEBOOK, INC. | 734 | Software & IT Services | Virgin Galactic Holdings, Inc. | Holding Companies | MICROSOFT CORPORATION | Software & IT Services |
| GENERAL ELECTRIC COMPANY | 729 | Industrial Conglomerates | FACEBOOK, INC. | Software & IT Services | MARRIOTT INTERNATIONAL, INC. | Hotels & Entertainment Se |
| EXXON MOBIL CORPORATION | 728 | Oil & Gas | INOVIO PHARMACEUTICALS, INC. | Biotechnology & Medical R | GENERAL ELECTRIC COMPANY | Industrial Conglomerates |
| THE WALT DISNEY COMPANY | 728 | Media & Publishing | THE WALT DISNEY COMPANY | Media & Publishing | MANPOWERGROUP INC. | Professional & Commercial |
| GILEAD SCIENCES, INC. | 721 | Biotechnology & Medical R | ROYAL CARIBBEAN CRUISES LTD. | Hotels & Entertainment Se | DELTA AIR LINES, INC. | Passenger Transportation |
| INOVIO PHARMACEUTICALS, INC. | 716 | Biotechnology & Medical R | GILEAD SCIENCES, INC. | Biotechnology & Medical R | LOWE'S COMPANIES, INC. | Specialty Retailers |
| Virgin Galactic Holdings, Inc. | 710 | Holding Companies | NVIDIA CORPORATION | Semiconductors & Semicond | EXXON MOBIL CORPORATION | Oil & Gas |
| MODERNA, INC. | 710 | Biotechnology & Medical R | BEYOND MEAT, INC. | Food & Tobacco | ALASKA AIR GROUP, INC. | Passenger Transportation |
| NVIDIA CORPORATION | 706 | Semiconductors & Semicond | MODERNA, INC. | Biotechnology & Medical R | AMAZON.COM, INC. | Diversified Retail |
| MGM RESORTS INTERNATIONAL | 695 | Hotels & Entertainment Se | JPMORGAN CHASE & CO. | Banking Services | CVS HEALTH CORPORATION | Healthcare Providers & Se |
| OCCIDENTAL PETROLEUM CORPORATION | 693 | Oil & Gas | Alibaba Group Holding Limited | Software & IT Services | FORD MOTOR COMPANY | Automobiles & Auto Parts |
| ROYAL CARIBBEAN CRUISES LTD. | 674 | Hotels & Entertainment Se | STARBUCKS CORPORATION | Hotels & Entertainment Se | VISA INC. | Software & IT Services |
| BEYOND MEAT, INC. | 610 | Food & Tobacco | EXXON MOBIL CORPORATION | Oil & Gas | AT&T INC. | Telecommunications Servic |
| NIO INC. | 24 | Automobiles & Auto Parts | WALMART INC. | Food & Drug Retailing | CITIGROUP INC. | Banking Services |
| SOUTHWEST AIRLINES CO. | 24 | Passenger Transportation | NORWEGIAN CRUISE LINE HOLDINGS LTD. | Hotels & Entertainment Se | Teladoc Health, Inc. | Healthcare Providers & Se |
| UBER TECHNOLOGIES, INC. | 24 | Software & IT Services | MGM RESORTS INTERNATIONAL | Hotels & Entertainment Se | AMC ENTERTAINMENT HOLDINGS, INC. | Hotels & Entertainment Se |
| Alibaba Group Holding Limited | 24 | Software & IT Services | GENERAL ELECTRIC COMPANY | Industrial Conglomerates | NORDIC AMERICAN TANKERS LIMITED | Oil & Gas Related Equipme |
| MARATHON OIL CORPORATION | 24 | Oil & Gas | BANK OF AMERICA CORPORATION | Banking Services | THE SOUTHERN COMPANY | Electrical Utilities & IP |
| NORWEGIAN CRUISE LINE HOLDINGS LTD. | 24 | Hotels & Entertainment Se | NETFLIX, INC. | Software & IT Services | Dine Brands Global, Inc. | Hotels & Entertainment Se |

Table 12 – Top 10 communities (and top 30 members) for the G-Network

| | communities | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **5** | **7** | **8** | **9** | **10** |
| **members** | **189** | **101** | **97** | **95** | **80** | **59** | **58** | **54** | **29** | **9** |
| **1** | AMGN | QD | SEE | XIN | ULTA | VAR | ESTE | AMP | DIN | MEET |
| **2** | AGEN | GE | CAT | PEG | REX | APD | CHUY | LEAF | COMM | NJR |
| **3** | MU | LPI | NAT | PPG | WW | HONE | SIMO | COUP | CAKE | PPL |
| **4** | MRKR | NWL | NBR | NL | CC | GS | Y | CAL | QUAD | BL |
| **5** | SKYW | JWN | SP | PD | W | SPWR | SY | ROAD | BG | PRU |
| **6** | OSTK | LVS | EPAY | YIN | SKY | PING | ES | UPS | DRI | ACN |
| **7** | MATX | EOG | TK | CHE | LDL | PLUG | MOS | REV | SAGE | UGI |
| **8** | PFE | MGA | SIRI | MCC | CMCSA | BEN | ORA | AVID | VOYA | MET |
| **9** | BNTX | CTL | NEWR | AME | VC | INS | ESTA | SCS | FARM | IBM |
| **10** | SQ | TILE | FRO | CE | BCC | MTN | TRV | CLNE | ASH | |
| **11** | LKQ | SWN | SHW | PARR | RAD | ZG | AIN | H | TREE | |
| **12** | MUR | WPX | SNAP | TRI | DGX | GERN | HPQ | VRTX | CORT | |
| **13** | VZ | OKE | IQ | DECK | ARES | AM | HA | ADVM | TSC | |
| **14** | COF | NCLH | AC | BOOM | S | WORK | XERS | RP | BRO | |
| **15** | HON | WYNN | AGM | HOFT | TMUS | ADP | BLK | SSD | TCS | |
| **16** | AAL | CPE | EW | JAX | TC | JAZZ | CASA | DOOR | NET | |
| **17** | SGEN | HTZ | ABG | BAP | GTS | PAYS | SI | CNC | BURL | |
| **18** | VNOM | HBAN | DHT | DAO | ETM | THOR | FOE | GPS | J | |
| **19** | AMZN | SDRL | MAN | ECHO | LOW | LITE | POR | FE | RM | |
| **20** | JMIA | WLL | CNA | TELA | FORM | CAMP | DL | XPO | LAD | |
| **21** | NKE | TGI | SJW | SATS | HUM | HES | MD | SAIA | LX | |
| **22** | NFLX | KSS | PUMP | UIS | ZS | VERY | BLL | SOI | FIT | |
| **23** | PCG | USFD | ADES | BAH | HAE | NBRV | LUNA | VG | FLO | |
| **24** | TPR | CHK | SLAB | TOUR | T | CHEF | GOLD | DT | EAT | |
| **25** | DLTR | EV | ETN | FOLD | FN | FLY | HPE | DAC | LL | |
| **26** | BAC | MTDR | MSA | CORE | JT | PM | MPX | OCUL | CENT | |
| **27** | TWLO | BBBY | HOV | BOX | D | HE | GEN | PROS | DQ | |
| **28** | MCD | RRR | EMR | FIVE | ZEUS | GLUU | NAK | ZEN | YELP | |
| **29** | VMW | VSLR | ALL | APO | CCC | GIS | TX | DELL | KOD | |
| **30** | CVX | TGTX | RUN | ATVI | LEA | ENPH | MITO | TOWN | | |

Table 13 – Industry's composition of the G-Network communities

| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **ALL** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Consumer Discretionary** | 19.5% | 20.8% | 5.7% | 4.4% | 10.8% | 1.4% | 14.5% | 1.7% | 33.5% | 0.0% | 16.1% |
| **Financials** | 4.7% | 17.3% | 7.5% | 6.5% | 9.4% | 10.1% | 14.5% | 11.9% | 27.5% | 18.3% | 7.0% |
| **Materials** | 0.1% | 1.5% | 3.2% | 8.4% | 1.2% | 6.5% | 13.8% | 0.9% | 4.8% | 0.0% | 1.5% |
| **Real Estate** | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| **Consumer Staples** | 3.2% | 8.3% | 17.5% | 8.8% | 14.1% | 11.7% | 11.4% | 0.1% | 8.9% | 0.0% | 5.4% |
| **Health Care** | 14.3% | 2.4% | 10.4% | 3.4% | 6.5% | 19.6% | 22.5% | 15.5% | 5.3% | 0.0% | 13.0% |
| **Utilities** | 0.0% | 0.7% | 0.2% | 4.4% | 6.3% | 25.1% | 3.7% | 2.7% | 0.0% | 8.4% | 2.1% |
| **Communication Services** | 15.7% | 2.2% | 11.8% | 13.7% | 36.5% | 2.5% | 9.4% | 0.6% | 2.0% | 0.0% | 14.4% |
| **Energy** | 1.2% | 37.8% | 0.9% | 0.2% | 0.2% | 1.8% | 0.3% | 0.1% | 0.0% | 0.0% | 3.0% |
| **Industrials** | 5.1% | 9.1% | 32.1% | 7.0% | 11.6% | 2.0% | 2.2% | 16.2% | 12.9% | 0.0% | 6.8% |
| **Information Technology** | 36.3% | 0.0% | 10.5% | 43.3% | 3.3% | 19.2% | 7.7% | 50.3% | 5.1% | 73.4% | 30.6% |

Table 14 – Coherence between correlation among stocks and community membership

| | Average Levesl of Correlation | | | Members of the Community | | |
|---|---|---|---|---|---|---|
| community | S&P500 | Insiders | Outsiders | Number | "Extraneous" | % Extraneous |
| **0** | 46.8% | 22.8% | 26.3% | 180 | 6 | 3.3% |
| **1** | 62.5% | 50.8% | 36.7% | 88 | 9 | 10.2% |
| **2** | 58.1% | 37.7% | 34.6% | 87 | 12 | 13.8% |
| **3** | 53.1% | 29.4% | 31.3% | 81 | 10 | 12.3% |
| **4** | 58.4% | 38.1% | 35.0% | 65 | 12 | 18.5% |
| **5** | 59.7% | 37.7% | 35.1% | 50 | 13 | 26.0% |
| **6** | 55.0% | 31.2% | 32.3% | 53 | 11 | 20.8% |
| **7** | 62.3% | 44.0% | 37.2% | 44 | 9 | 20.5% |
| **8** | 63.3% | 45.9% | 38.2% | 24 | 4 | 16.7% |
| **9** | 76.8% | 67.5% | 46.1% | 7 | 1 | 14.3% |

Table 15 – Optimal portfolios built using “Industry” and “Community” clustering

| Industries | CAPM Model | | | Communities | CAPM Model | |
|---|---|---|---|---|---|---|
| | Expected Return | Weights (Market Portf.) | | | Expected Return | Weights (Market Portf.) |
| Health Care | 5.4% | 9.06% | | Community 0 | 6.0% | 10.00% |
| Materials | 6.1% | 9.19% | | Community 1 | 8.3% | 10.29% |
| Consumer Discretionary | 6.5% | 9.42% | | Community 2 | 6.4% | 9.87% |
| Industrials | 6.4% | 9.14% | | Community 3 | 6.0% | 9.82% |
| Financials | 6.1% | 9.39% | | Community 4 | 6.4% | 10.17% |
| IT | 5.5% | 9.12% | | Community 5 | 6.0% | 9.84% |
| Consumer Staples | 4.5% | 8.20% | | Community 6 | 5.9% | 9.65% |
| Utilities | 4.6% | 8.92% | | Community 7 | 6.6% | 10.38% |
| Communications | 5.9% | 8.88% | | Community 8 | 6.6% | 10.08% |
| Energy | 8.2% | 9.33% | | Community 9 | 5.9% | 9.91% |
| Real Estate | 6.1% | 9.35% | | | | |
| | Optimal Portfolio (tangent) | | | | Optimal Portfolio (tangent) | |
| Expected return | 5.0% | | | Expected return | 6.0% | |
| Volatility | 21.3% | | | Volatility | 26.8% | |
| Risk-free rate | 2.0% | | | Risk-free rate | 2.0% | |
| Sharpe ratio | 14.1% | | | Sharpe ratio | 14.9% | |

Table 16 – Characteristics of the Network structure of the Stock Market (2020Q2)

| | **C-Network** | **G-Network** |
|---|---|---|
| Number of eligible "nodes" | 1312 | 821 |
| Number of "edges" | 17336 | 2395 |
| Average degree of the nodes | 26 | 5.8 |
| Density of the Network | 2% | 0.7% |
| Diameter of the Network | 4 | 12 |
| Triadic closure (transitivity) | 6.1% | 21.2% |

Table 17 – Communities correspondence between C-Networks and G-Networks

| **2020Q2** | **communities correspondence based on largest intersection** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **communities labels** | | | | | | | |
| **Correlation Network** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| **GoogleTrends Network** | **0** | **4** | **1** | **2** | **3** | **5** | **7** | |
| *shared nodes* | *38* | *22* | *22* | *12* | *10* | *8* | *5* | |

Table 18 – Changes in the network topology

| | **18Q01** | **18Q02** | **18Q03** | **18Q04** | **19Q01** | **19Q02** | **19Q03** | **19Q04** | **20Q01** | **20Q02** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Num. Nodes** | 458 | 463 | 468 | 451 | 463 | 479 | 486 | 520 | 606 | 615 |
| **Avg. Degrees** | 3.25 | 3.24 | 3.30 | 3.28 | 3.08 | 3.20 | 3.11 | 3.08 | 4.05 | 5.37 |
| **Density** | 0.71% | 0.70% | 0.71% | 0.73% | 0.67% | 0.67% | 0.64% | 0.59% | 0.67% | 0.87% |
| **Triadic closure** | 21.6% | 25.3% | 23.1% | 23.9% | 20.2% | 25.2% | 20.3% | 18.4% | 22.5% | 28.6% |
| **Diameter** | 14 | 17 | 14 | 14 | 15 | 14 | 14 | 14 | 15 | 14 |

Table 19 – Top 10 hubs in the quarters: 18Q01-20Q02 (eigenvector centrality)

| | 18Q01 | 18Q02 | 18Q03 | 18Q04 | 19Q01 | 19Q02 | 19Q03 | 19Q04 | 20Q01 | 20Q02 |
|---|---|---|---|---|---|---|---|---|---|---|
| **AAPL** | 34.0% | 31.0% | 29.6% | 33.7% | 34.8% | 31.6% | 36.0% | 35.6% | 30.5% | 21.1% |
| **AMZN** | 31.3% | 31.1% | 31.0% | 35.2% | 35.5% | 29.7% | 31.6% | 30.9% | 22.7% | 17.5% |
| **FB** | 29.5% | 25.8% | 30.9% | 24.5% | 26.3% | 27.0% | 20.0% | 19.4% | 14.7% | 11.8% |
| **NVDA** | 23.9% | 22.2% | 18.4% | 21.8% | 20.7% | 21.1% | 22.5% | 20.2% | 14.3% | 14.4% |
| **MSFT** | 23.3% | 23.3% | 20.7% | 18.8% | 22.4% | 23.3% | 26.3% | 28.1% | 25.5% | 17.4% |
| **BABA** | 23.3% | 20.9% | 23.6% | 17.5% | 14.7% | 20.6% | 20.1% | 19.2% | 11.3% | 11.9% |
| **MU** | 21.6% | 25.0% | 22.1% | 13.5% | 15.0% | 12.6% | 11.0% | 9.1% | 8.7% | 5.4% |
| **GE** | 19.1% | 15.1% | 0.0% | 19.6% | 14.0% | 7.5% | 5.8% | 11.3% | 7.6% | 9.1% |
| **NFLX** | 18.4% | 19.7% | 22.0% | 20.1% | 20.2% | 20.2% | 22.4% | 17.3% | 12.3% | 9.9% |
| **TSLA** | 17.4% | 20.6% | 19.5% | 22.9% | 22.1% | 27.9% | 24.1% | 24.5% | 32.8% | 21.1% |
| **TWTR** | 17.3% | 20.3% | 15.9% | 17.5% | 16.3% | 13.1% | 15.9% | 18.1% | 9.3% | 8.7% |
| **BAC** | 14.0% | 20.0% | 16.2% | 14.7% | 16.2% | 12.3% | 6.3% | 13.2% | 11.7% | 12.8% |
| **AMD** | 10.8% | 15.8% | 26.1% | 21.6% | 19.2% | 22.6% | 0.0% | 22.4% | 16.8% | 11.1% |
| **BA** | 16.8% | 10.3% | 7.2% | 9.3% | 25.5% | 23.1% | 15.2% | 14.9% | 26.3% | 26.9% |
| **LYFT** | | | | | | 21.4% | 0.1% | 3.6% | 0.7% | 2.7% |
| **ROKU** | | 1.0% | 7.4% | 0.0% | 9.6% | 5.2% | 25.0% | 16.7% | 7.7% | |
| **BYND** | | | | | | | 21.5% | 12.8% | 11.6% | 12.9% |
| **DIS** | 11.1% | 7.7% | 10.0% | 7.1% | 6.8% | 16.2% | 16.5% | 13.1% | 16.9% | 13.4% |
| **XOM** | 7.7% | 15.2% | 11.2% | 11.5% | 11.3% | 8.5% | 5.0% | 6.2% | 16.2% | 13.3% |
| **SPCE** | | | | | | | | | 16.0% | 9.0% |
| **AAL** | 2.3% | 1.6% | | | 0.8% | 2.0% | 5.0% | 1.6% | 12.2% | 21.1% |
| **UAL** | 2.1% | 1.3% | | | 0.1% | | | | 10.7% | 20.6% |
| **CCL** | 0.0% | 1.9% | | | 0.0% | | | | 14.2% | 20.5% |
| **DAL** | 2.1% | 0.0% | | | 1.8% | | | | 10.6% | 19.9% |

Table 20 – The popularity of the Top 10 hubs among Robinhood users (2020Q02)

| | | 2020Q02 | | |
|---|---|---|---|---|
| | | G-Network | Robinhood Individual Positions | |
| ticker | Short Name | Eigenv. Centrality | Delta | Popularity Rank |
| **BA** | **BOEING** | 26.9% | 168,612 | 12 |
| **TSLA** | **TESLA** | 21.1% | 112,689 | 24 |
| **AAPL** | **APPLE** | 21.1% | 178,048 | 10 |
| **AAL** | **AMERICAN AIRLINES** | 21.1% | 440,745 | 1 |
| **UAL** | **UNITED AIRLINES** | 20.6% | 235,601 | 6 |
| **CCL** | **CARNIVAL** | 20.5% | 311,896 | 4 |
| **DAL** | **DELTA AIR LINES** | 19.9% | 403,134 | 2 |
| **AMZN** | **AMAZON** | 17.5% | 181,999 | 9 |
| **MSFT** | **MICROSOFT** | 17.4% | 109,147 | 26 |

Table 21 – The performance of the Top 100 hubs

| | 18Q01 | | 18Q02 | | 18Q03 | | 18Q04 | |
|---|---|---|---|---|---|---|---|---|
| | avg return | delta wrt SPX | avg return | delta wrt SPX | avg return | delta wrt SPX | | |
| 1 quartile | 5.4% | 7.4% | 13.5% | 10.5% | 11.3% | 4.1% | -22.4% | -8.4% |
| 2 quartile | 7.4% | 9.5% | 22.5% | 19.5% | 2.3% | -4.9% | -18.2% | -4.2% |
| 3 quartile | -4.5% | -2.4% | 0.0% | -2.9% | 6.0% | -1.2% | -17.3% | -3.3% |
| 4 quartile | -5.2% | -3.2% | 1.7% | -1.2% | 9.7% | 2.5% | -15.1% | -1.1% |
| top100 | 0.8% | 2.8% | 9.4% | 6.5% | 7.3% | 0.1% | -18.2% | -4.3% |
| SPX | -2.04% | | 2.93% | | 7.20% | | -13.97% | |

| | 19Q01 | | 19Q01 | | 19Q01 | | 19Q01 | |
|---|---|---|---|---|---|---|---|---|
| | avg return | delta wrt SPX | avg return | delta wrt SPX | avg return | delta wrt SPX | avg return | delta wrt SPX |
| 1 quartile | 20.5% | 7.4% | 0.3% | -2.9% | -1.1% | -2.8% | 25.6% | 17.1% |
| 2 quartile | 14.3% | 1.3% | 1.2% | -2.0% | 0.2% | -1.5% | 26.7% | 18.2% |
| 3 quartile | 19.2% | 6.1% | -2.8% | -6.0% | -7.8% | -9.5% | 7.7% | -0.8% |
| 4 quartile | 20.3% | 7.3% | 6.2% | 3.0% | -3.8% | -5.6% | 6.9% | -1.7% |
| top100 | 18.6% | 5.5% | 1.2% | -2.0% | -3.1% | -4.9% | 16.7% | 8.2% |
| SPX | 13.07% | | 3.19% | | 1.77% | | 8.53% | |

| | 20Q01 | | 20Q02 | |
|---|---|---|---|---|
| | avg return | delta wrt SPX | avg return | delta wrt SPX |
| 1 quartile | -10.7% | 9.3% | 27.9% | 7.9% |
| 2 quartile | -0.3% | 19.7% | 100.1% | 80.1% |
| 3 quartile | -17.2% | 2.8% | 59.1% | 39.2% |
| 4 quartile | 1.4% | 21.4% | 53.6% | 33.6% |
| top100 | -6.7% | 13.3% | 60.2% | 40.2% |
| SPX | -20.00% | | 19.95% | |

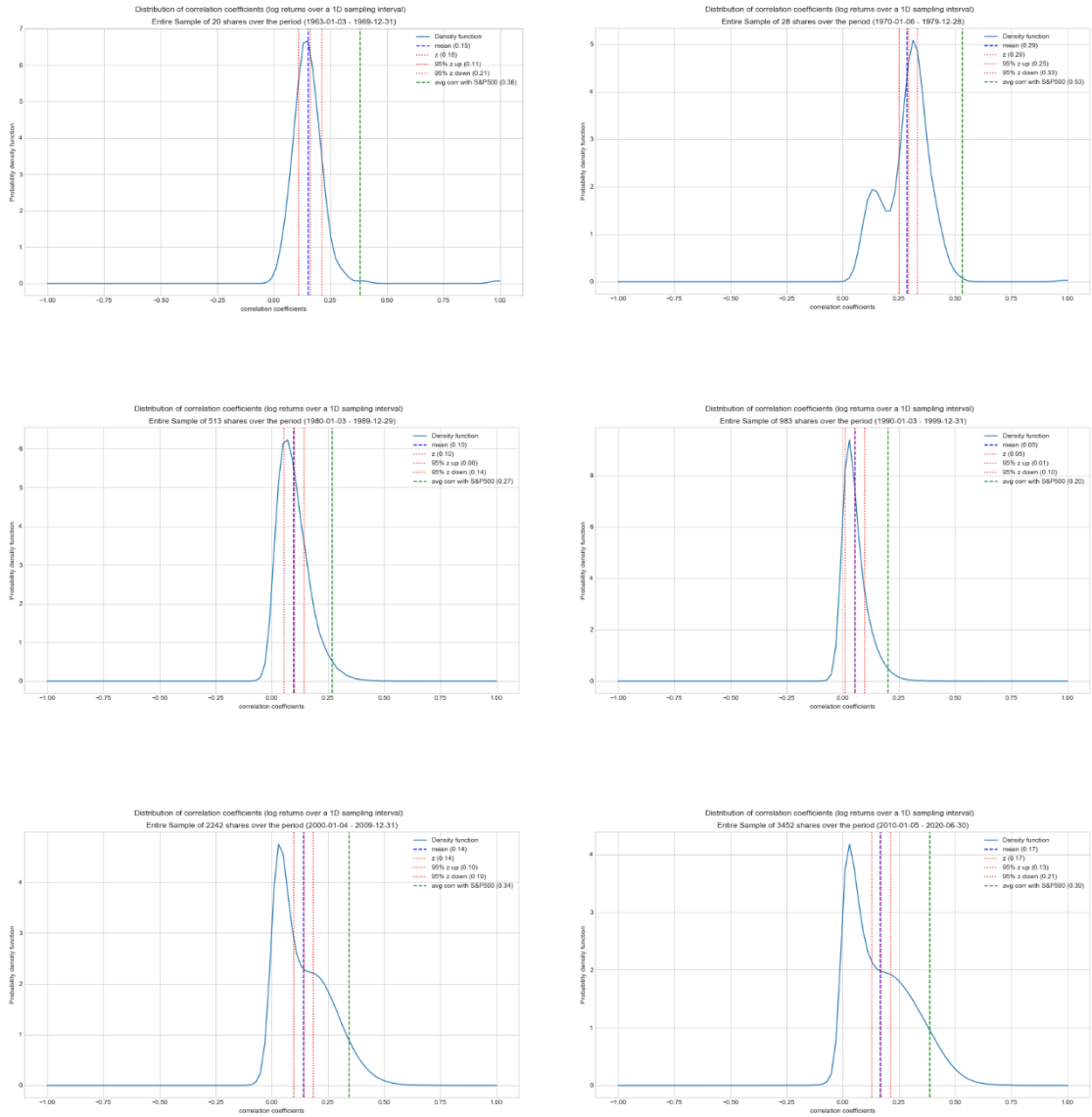
Table 22 – Community membership of the Top 100 hubs

Community membership of the Top100 Hubs

| 18Q01 | | 18Q02 | | 18Q03 | | 18Q04 | | 19Q01 | | 19Q02 | | 19Q03 | | 19Q04 | | 20Q01 | | 20Q02 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| label | n. in Top100 | label | n. in Top100 | label | n. in Top100 | label | n. in Top100 | label | n. in Top100 | label | n. in Top100 | label | n. in Top100 | label | n. in Top100 | label | n. in Top100 | label | n. in Top100 |
| 0 | 76 | 0 | 73 | 0 | 82 | 0 | 76 | 0 | 71 | 0 | 70 | 0 | 68 | 0 | 69 | 0 | 86 | 0 | 92 |
| 1 | 4 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 5 | 1 | 1 | 1 | 5 | 1 | 1 | 1 | 1 | 2 | 3 |
| 2 | 1 | 2 | 3 | 2 | 4 | 2 | 2 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 1 | 4 | 1 | 6 | 2 |
| 5 | 5 | 4 | 1 | 4 | 2 | 3 | 3 | 3 | 8 | 4 | 2 | 3 | 4 | 3 | 1 | 5 | 1 | 8 | 2 |
| 6 | 1 | 5 | 3 | 6 | 1 | 4 | 4 | 4 | 1 | 5 | 9 | 4 | 15 | 4 | 1 | 6 | 1 | 10 | 1 |
| 7 | 1 | 7 | 3 | 7 | 1 | 5 | 1 | 5 | 5 | 6 | 1 | 6 | 1 | 5 | 2 | 7 | 3 | | |
| 8 | 2 | 8 | 1 | 8 | 1 | 6 | 3 | 6 | 1 | 8 | 3 | 7 | 2 | 6 | 2 | 8 | 4 | | |
| 9 | 2 | 9 | 1 | 9 | 4 | 7 | 1 | 8 | 4 | 9 | 3 | 9 | 1 | 7 | 4 | 10 | 2 | | |
| 11 | 1 | 11 | 3 | 11 | 1 | 8 | 1 | 20 | 1 | 10 | 2 | 17 | 1 | 8 | 1 | 21 | 1 | | |
| 12 | 2 | 12 | 1 | 22 | 1 | 9 | 1 | 37 | 1 | 14 | 2 | | | 9 | 10 | | | | |
| 14 | 3 | 14 | 1 | 23 | 1 | 10 | 1 | | | 18 | 1 | | | 10 | 1 | | | | |
| 17 | 1 | 20 | 2 | 31 | 1 | 24 | 2 | | | 21 | 1 | | | 11 | 4 | | | | |
| 19 | 1 | 21 | 1 | | | 25 | 1 | | | 22 | 1 | | | 18 | 2 | | | | |
| | | 22 | 1 | | | 27 | 1 | | | 24 | 1 | | | 22 | 1 | | | | |
| | | 23 | 1 | | | 44 | 1 | | | | | | | | | | | | |
| | | 28 | 1 | | | | | | | | | | | | | | | | |
| | | 33 | 1 | | | | | | | | | | | | | | | | |
| | | 37 | 1 | | | | | | | | | | | | | | | | |

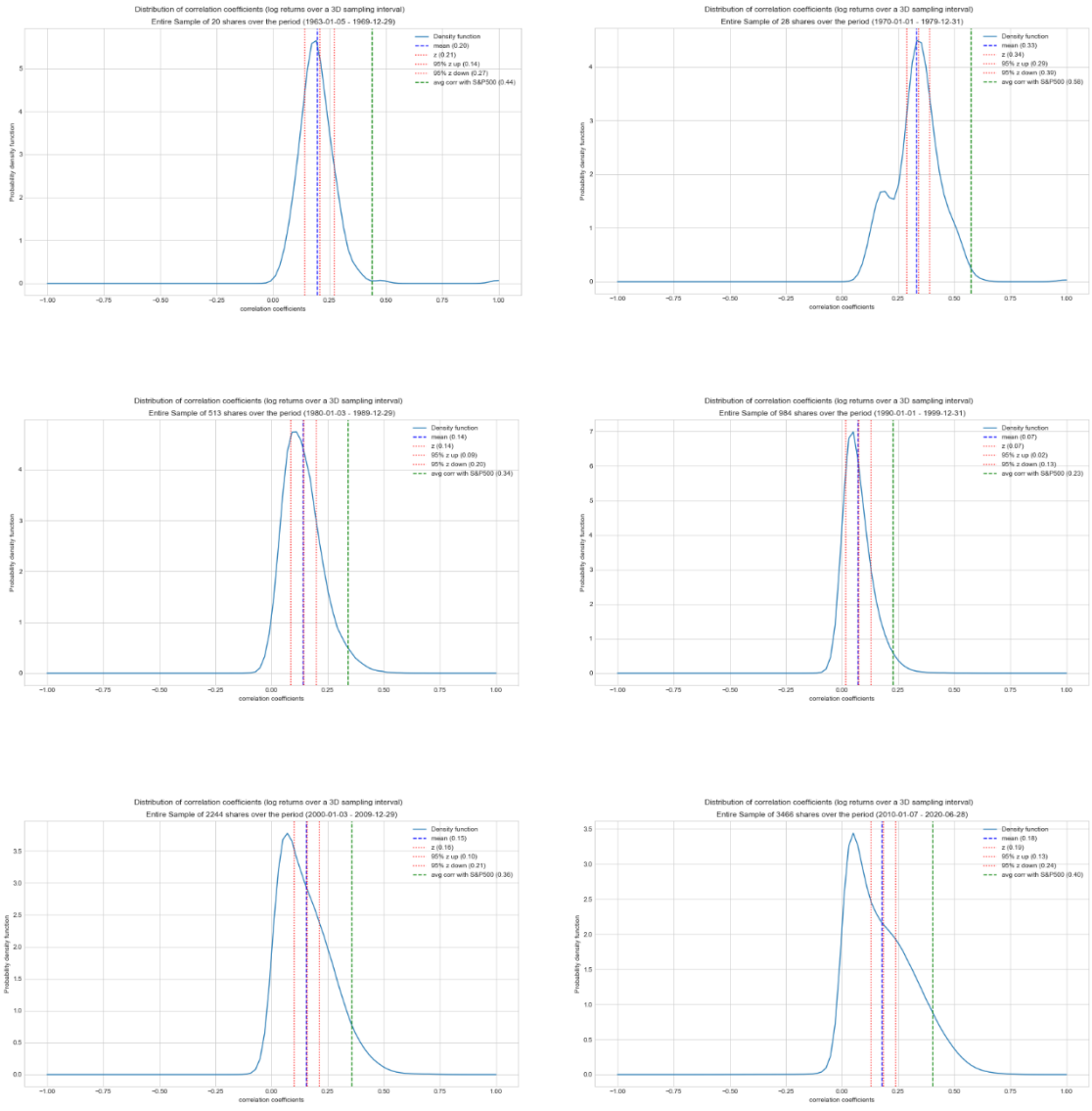
# Figures

Fig. 1 - The distribution of the correlation coefficients among 1-day log returns (1963-2020 by decades)



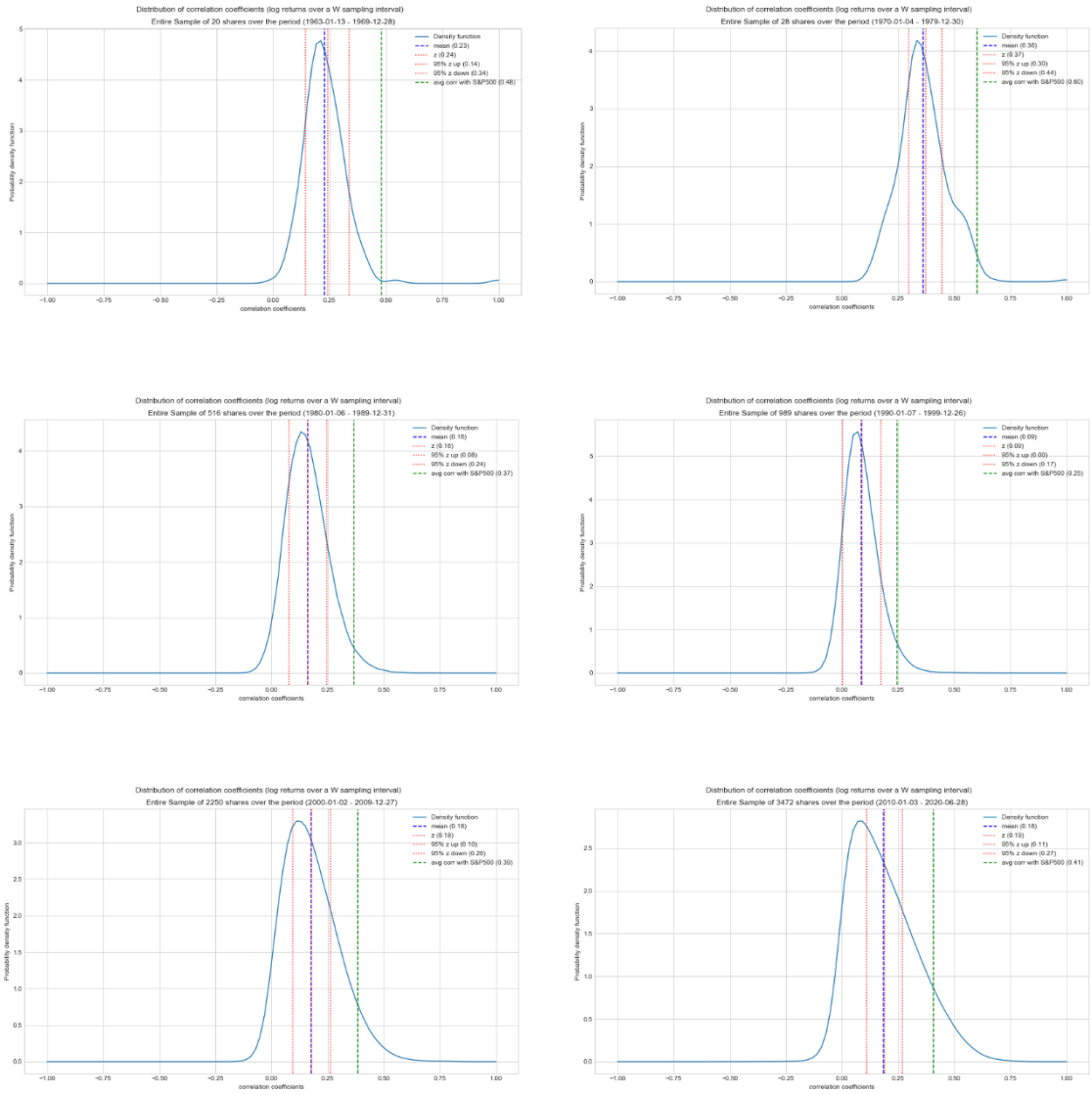

The vertical lines represent the mean correlation (violet) +/- 1 std deviation (orange), and the average correlation with the S&P500 (green).

Fig. 2- The distribution of the correlation coefficients among 3-day log returns (1963-2020 by decades)



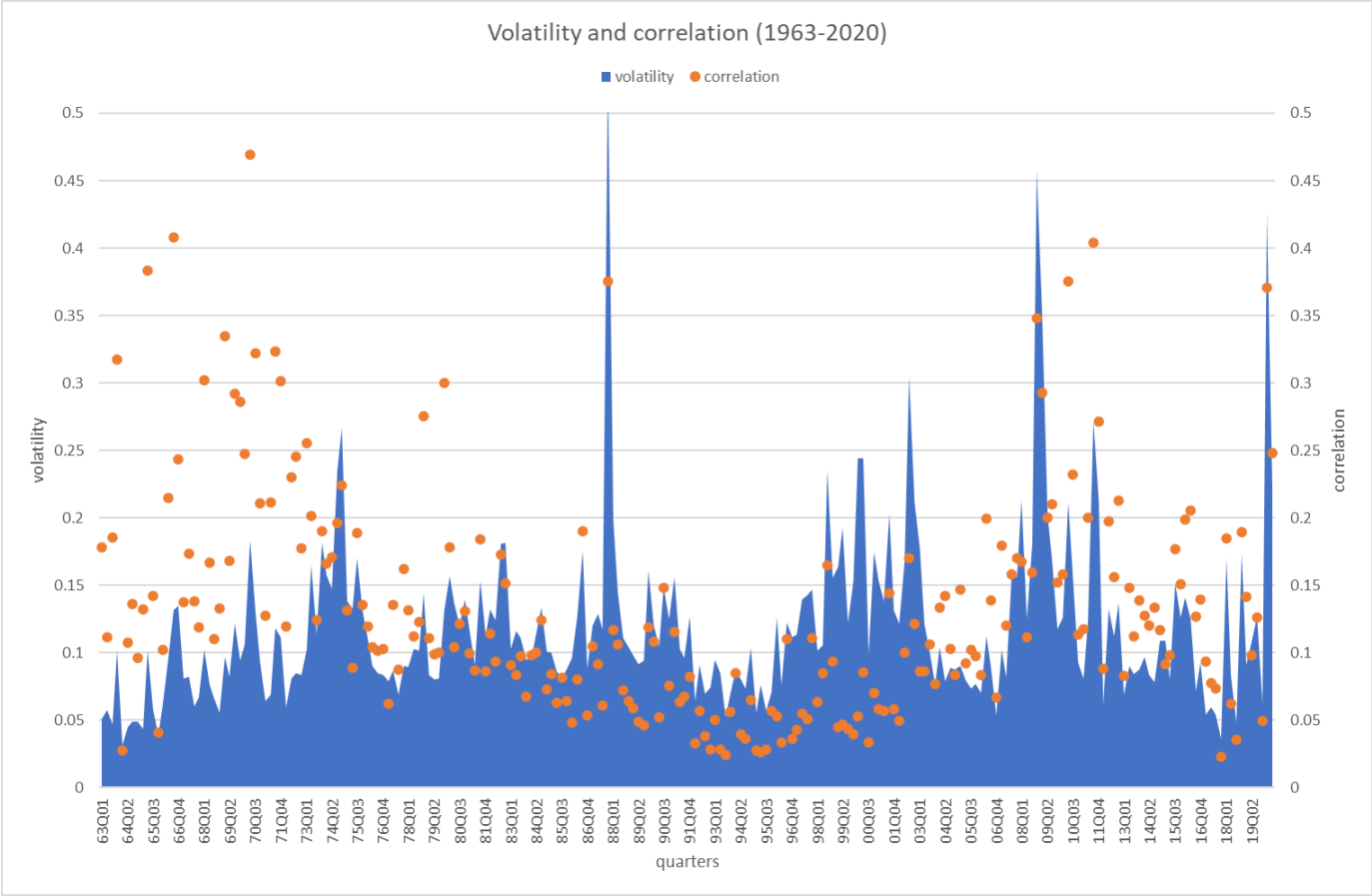

The vertical lines represent the mean correlation (violet) +/- 1 std deviation (orange), and the average correlation with the S&P500 (green).

Fig. 3- The distribution of the correlation coefficients among weekly log returns (1963-2020 by decades)




The vertical lines represent the mean correlation (violet) +/- 1 std deviation (orange), and the average correlation with the S&P500 (green).

Fig. 4 – The dynamics of volatility and correlation among stock returns




http://dx.doi.org/10.25428/2532-554X/10

Fig. 5 – The correlation density before, during and soon after a major crisis

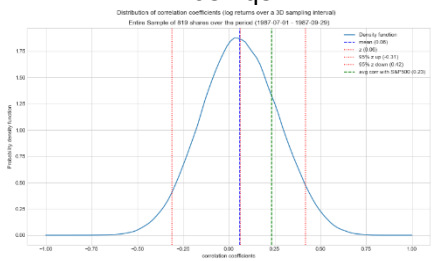
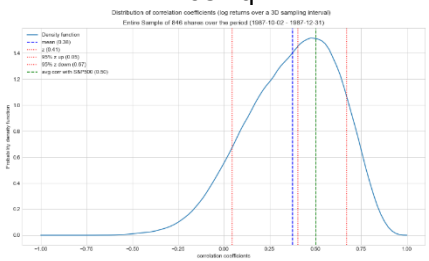
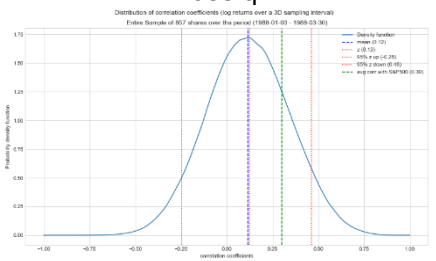
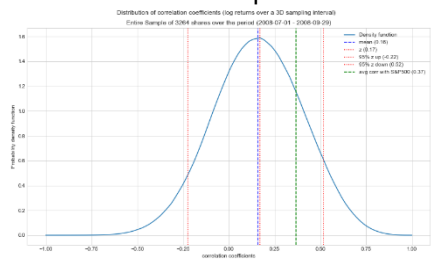
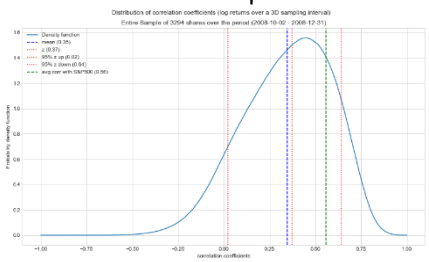
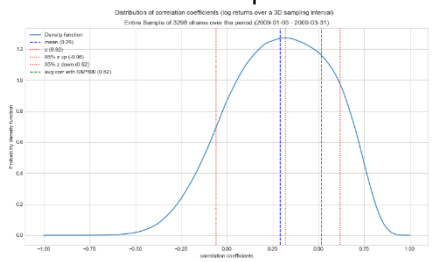
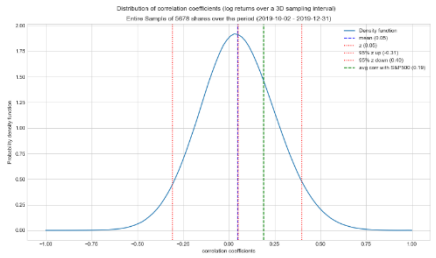
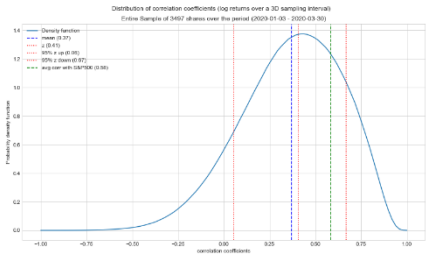
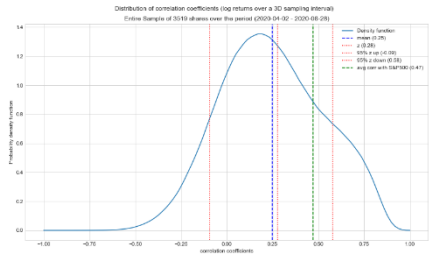
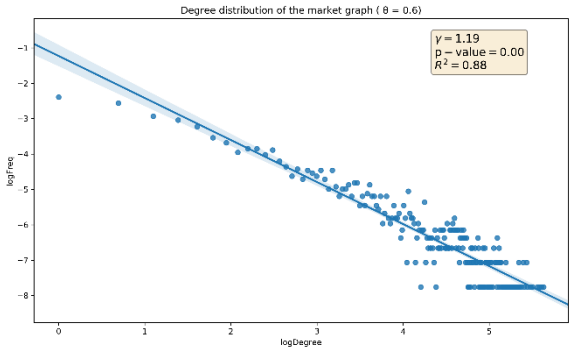
| Before the shock | At the shock | After the shock |
|---|---|---|
| 1987 q3 | 1987 q4 | 1988 q1 |
|  |  |  |
| 2008 q3 | 2008 q4 | 2009 q1 |
|  |  |  |
| 2019 q4 | 2020 q1 | 2020 q2 |
|  |  |  |

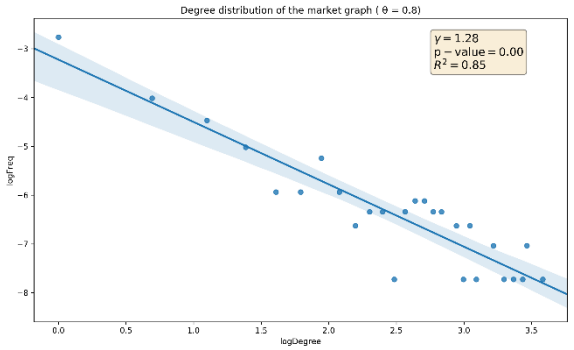
Fig.6- Typical degrees' distribution before the monetary policy regime shift occurred with the Lehman Brothers default

"normal" times (2001 – q2)

$\theta_L$= 60%

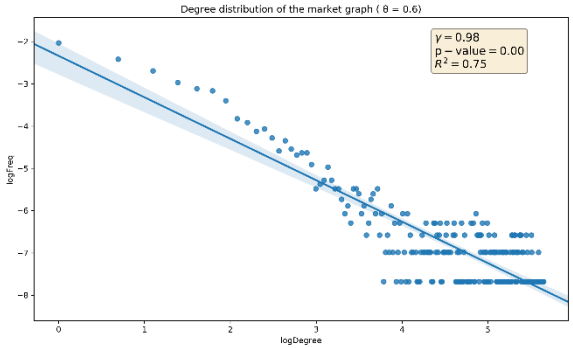



$\theta_H$=80 %

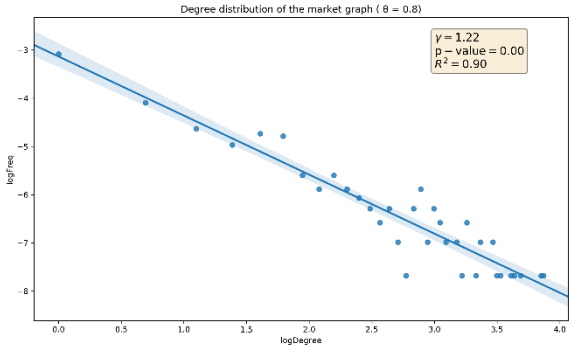



"crises" times (2000 – q2)

$\theta_L$= 75%




$\theta_H$=90 %

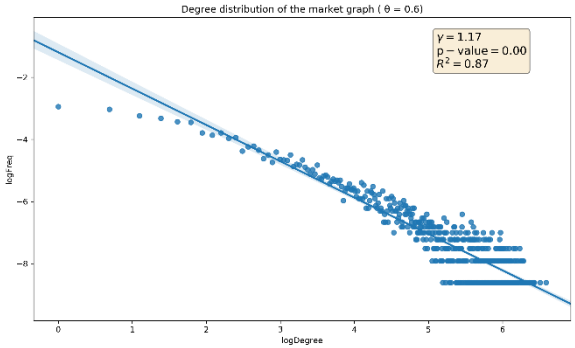



http://dx.doi.org/10.25428/2532-554X/10

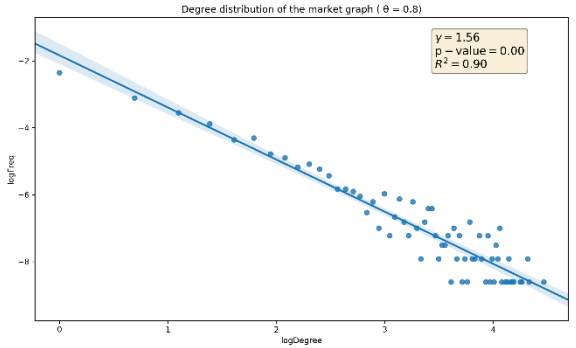
Fig.7- Typical degrees' distribution after the monetary policy regime shift occurred with the Lehman Brothers default

"normal" times (2019 – q2)

$\theta_L$= 60%

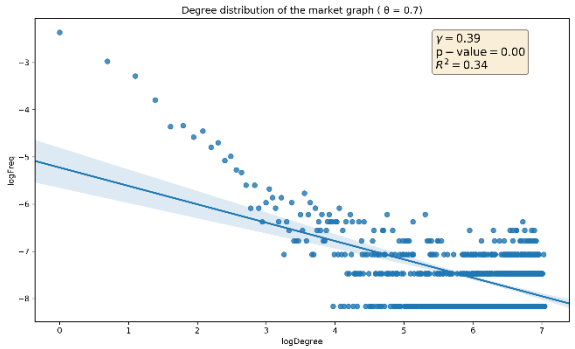



$\theta_H$=80 %




"crises" times (2020 – q2)

$\theta_L$= 70%




$\theta_H$=85 %

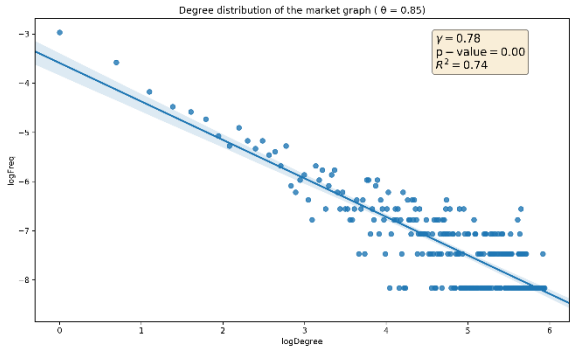


Fig. 8 - Typical degrees' distribution for the returns' residuals

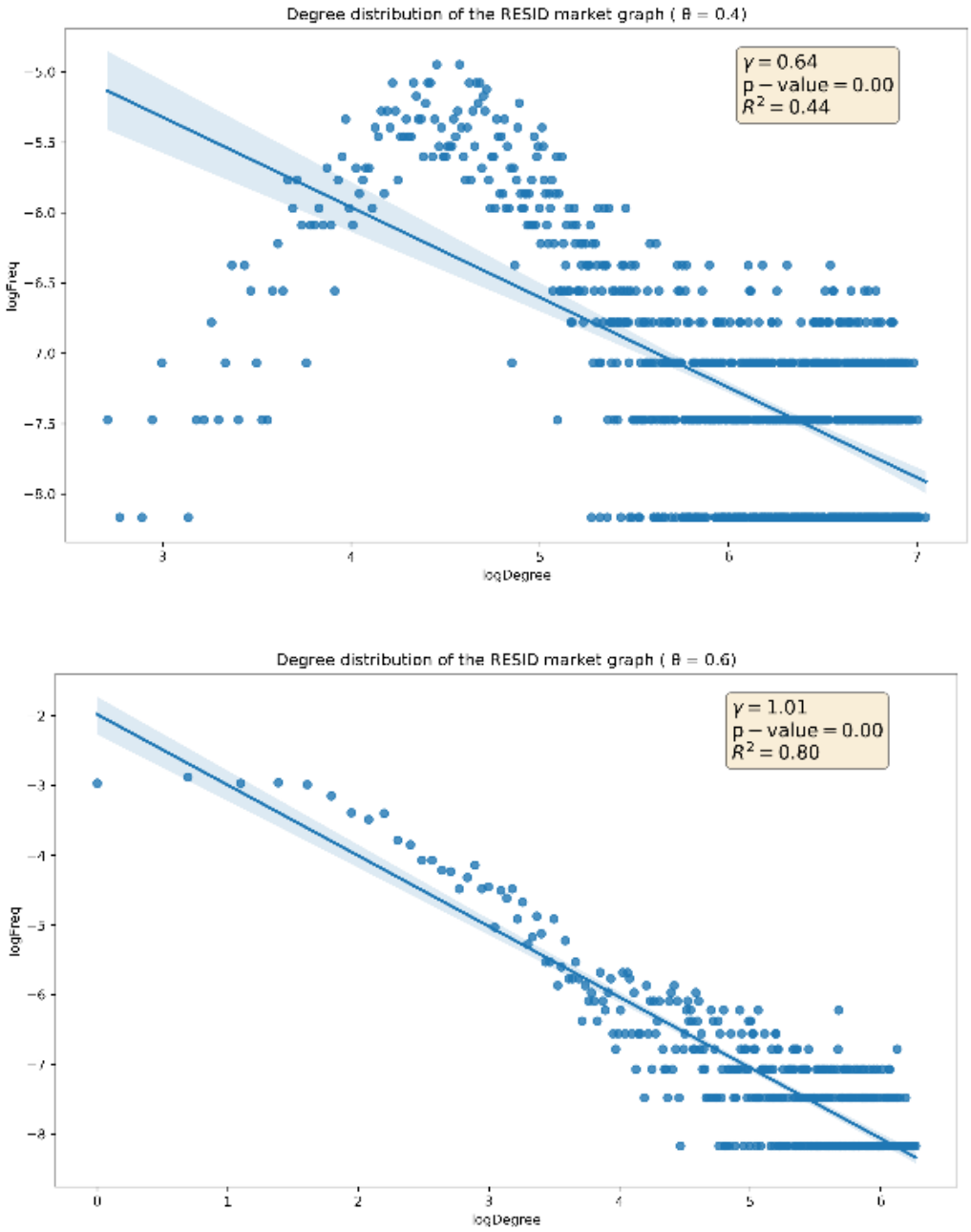


Fig. 9 - The distribution of the “inside” and “outside” correlation for each community

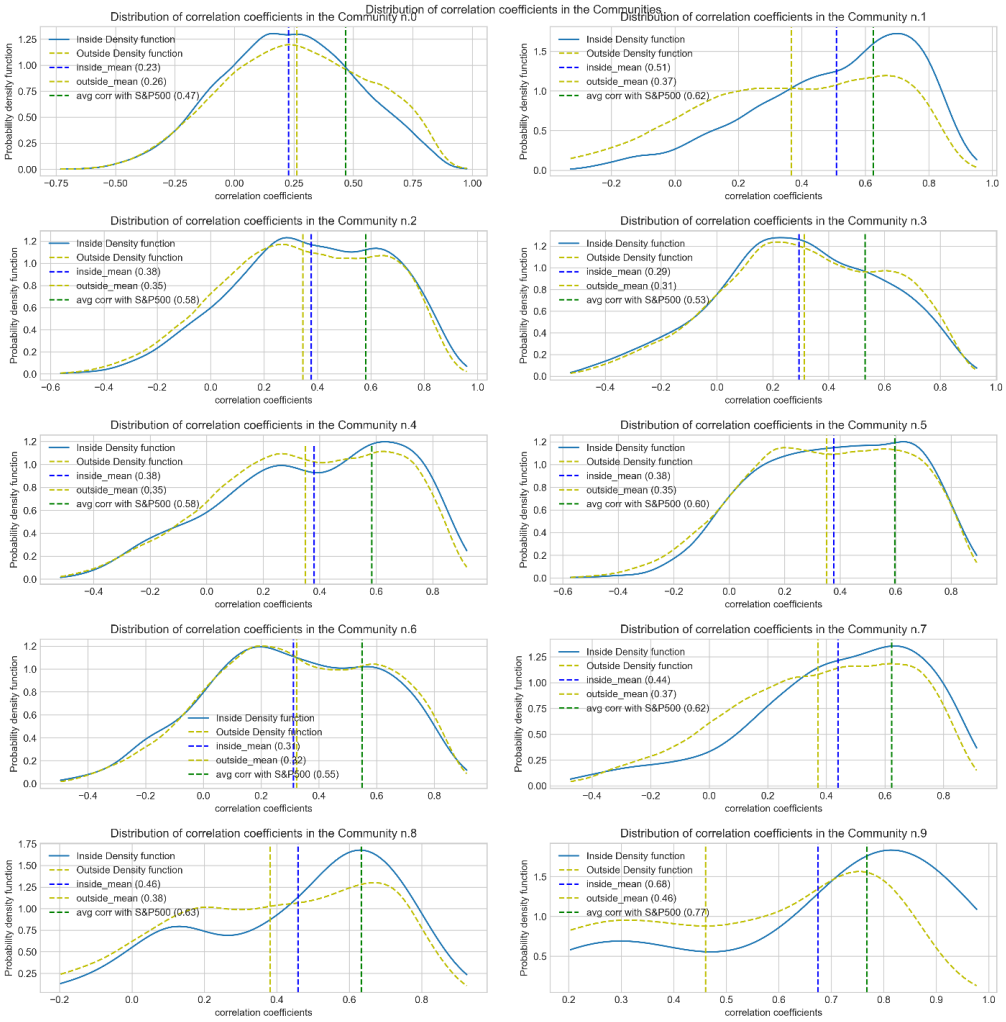


Fig. 10 - The market graph of the G-Network (degree centrality)

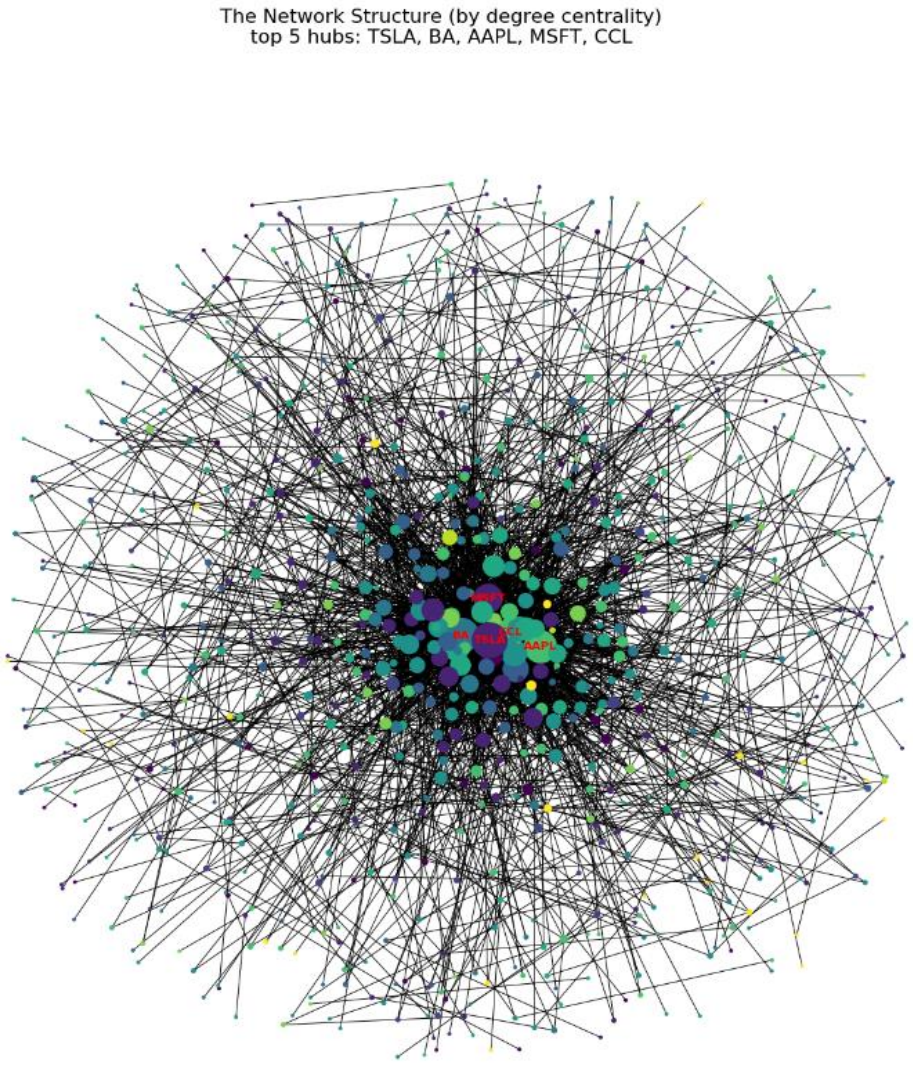


Fig. 11 - The market graph of the G-Network (eigenvector centrality)

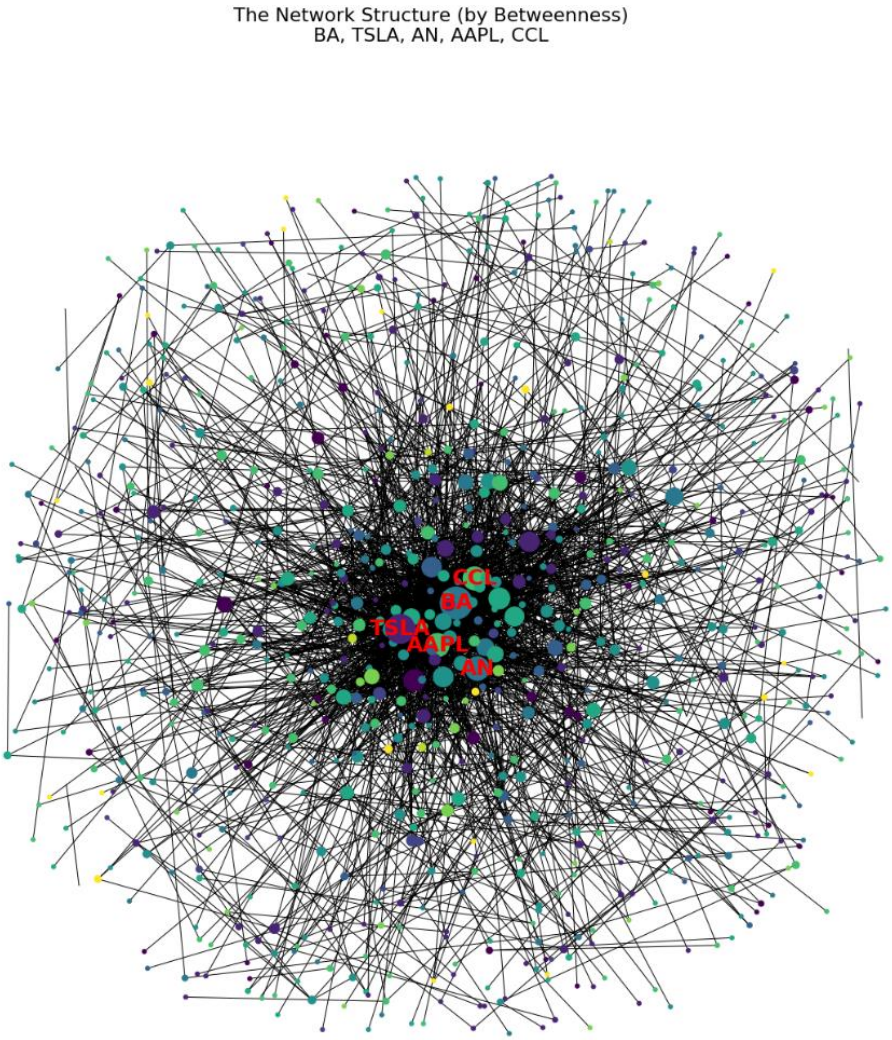


Fig. 12 - The efficient frontier built on the communities of the G-Network

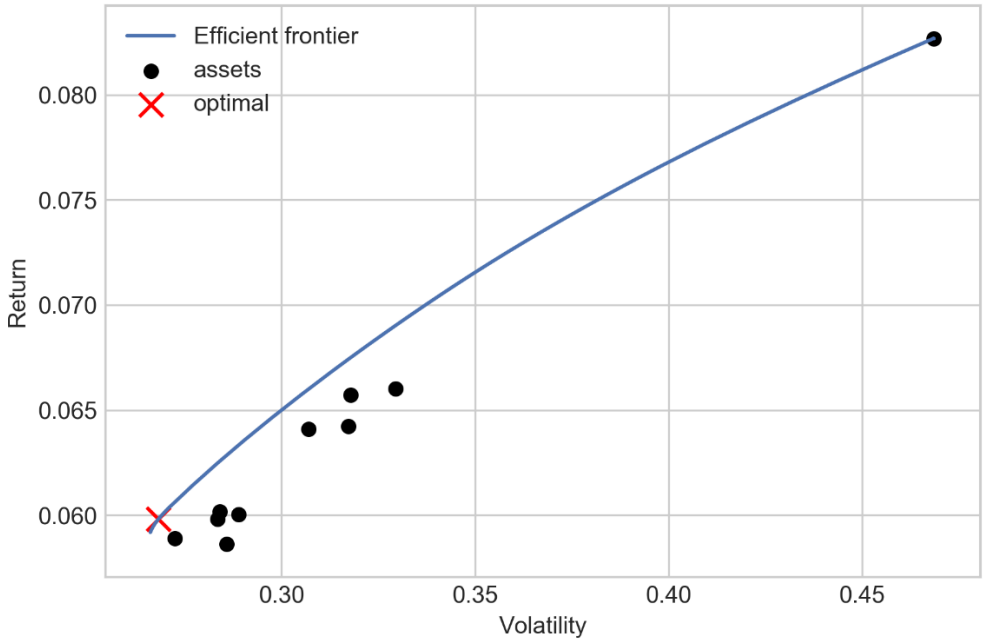



Fig. 13 - The efficient frontier built on the “Industries”

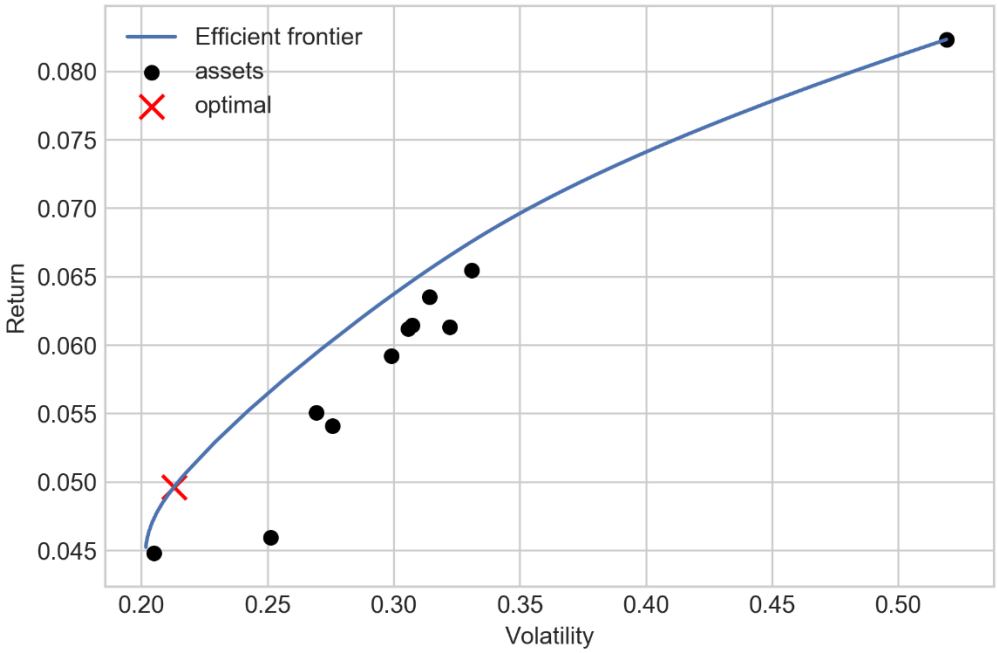
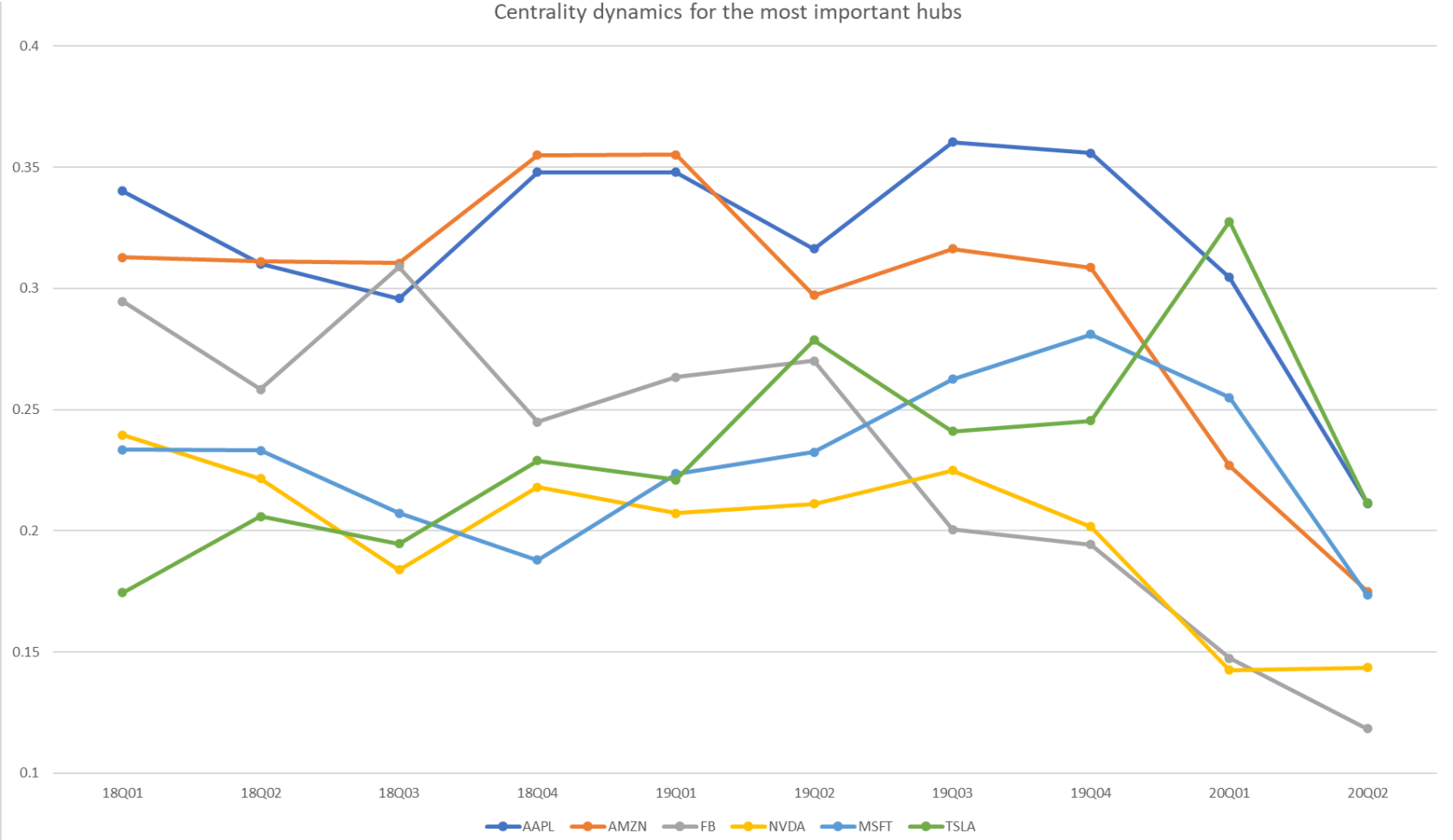


Fig. 14 – Evolution of the importance of the hubs of the G-network (2018Q1-2020Q2)



## Appendix 1 - the CAPM model for stocks' returns

The simplest way of modeling the return of the i-th stock is according to the well-known CAPM26 equation:

$$\boldsymbol{r_i = \beta_i r_{mkt} + \varepsilon_i} \quad \text{(A.1)}$$

The equation contains one parameter ("beta") and two independent random variables: the market return and an error term,$\boldsymbol{\varepsilon}$. Both random variables are assumed to be Normally distributed:

$$\boldsymbol{r_{mkt} \sim N(\bar{r}, \sigma^2)} \quad \text{(A.2)}$$

$$\boldsymbol{\varepsilon_i \sim N(0, \omega_i{}^2)} \quad \text{(A.3)}$$

From this representation it derives that the return of the i-th stock is Normally distributed as:

$$\boldsymbol{r_i \sim N(\beta_i \bar{r}, \beta_i{}^2 \sigma^2 + \omega_i{}^2)} \quad \text{(A.4)}$$

While the correlation between the return of the i-th and j-th stocks is equal to:

$$\boldsymbol{corr(r_i, r_j) = \frac{\beta_i \beta_j \sigma^2}{\sqrt{\beta_i{}^2 \sigma^2 + \omega_i{}^2}\sqrt{\beta_j{}^2 \sigma^2 + \omega_j{}^2}}} \quad \text{(A.5)}$$

The presence of two random terms in the return equation allows for identifying a relation between volatility and correlation. To see this relationship in a clear way, we can make a further simplification assuming that all securities are identically distributed. If this is the case, then the beta term is equal to 1 and equations (A.4) and (A.5) become:

$$\boldsymbol{r_i \sim N(\bar{r}, \sigma^2 + \omega^2)} \quad \text{(A.4a)}$$

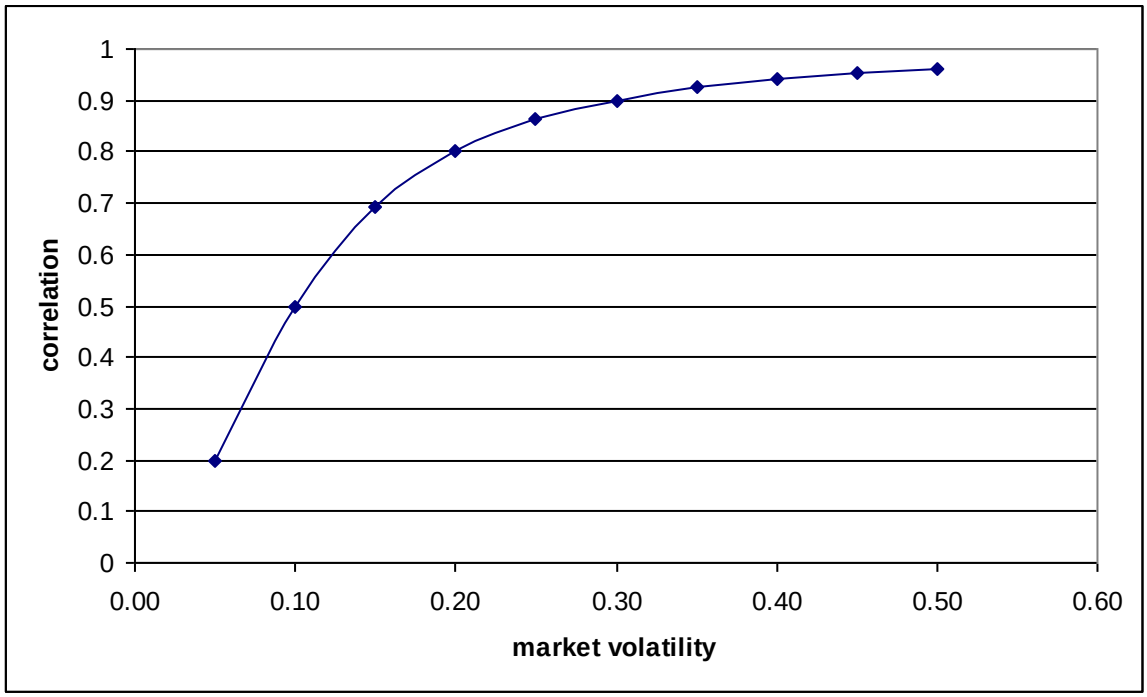
$$\boldsymbol{\rho \equiv corr(r_i, r_j) = \frac{\sigma^2}{\sigma^2 + \omega^2}} \quad \text{(A.5a)}$$

From (A.5a), it is quite easy to see the relationship between the volatility components and the correlation term. In particular, the higher the volatility of the market term (σ) and the higher is the correlation between securities. The opposite holds true if the volatility (ω) of the idiosyncratic term increases.

[26] See Sharpe (1964) and Ross (1977) for the original contributions and Elton et al (2003, ch. 13) as a textbook.

In figure A.1, we show how correlation increases (from 20% to almost 100%) when market volatility increases from 5% to 50%. Note that we assume that idiosyncratic volatility is equal to 10% and remains constant. Those values might seem extreme and "unrealistic", and they are indeed unobserved most of the time. But there are periods during which they become the norm and these periods are what we define as financial crises. And, unfortunately, for the "Normality" hypothesis of the stock markets' returns, they are extremely frequent also in Wall Street, probably the most sophisticated and liquid market of all the times.

Fig. A1.1 – the relationship between market volatility and average correlation between single stocks implied by a simplified CAPM model of returns




It must be noted that this simple model can shed light on another statistical property of stock returns that can seem puzzling at first sight, i.e., the fact that the average correlation among stocks is lower than the average correlation of the stocks with a representative index of the "market".[27]

If we consider a stock market composed of a set of N stocks of identical capitalization, the return of the market index is the average return of the N i.i.d. stocks and its return distribution is simply given by equation (A.2). It derives that the correlation between the i-th stock and the index is:

[27] The average is calculated with respect to the elements of the upper triangular (excluding the diagonal) of the matrix correlation.

$$\boldsymbol{corr(r_i, r_{mkt})} = \frac{\boldsymbol{\sigma}}{\sqrt[2]{\boldsymbol{\sigma^2+\omega^2}}} \tag{A.6}$$

Since we assumed that the stocks are i.i.d. the average correlation is exactly given by eq. (A.6). Now we can compare eq. (A.6) with eq. (A.5a) and we can immediately prove that the former is bigger than the latter. The ratio between the average correlation with the index and the average correlation among stocks is:

$$\frac{\sum_i corr(r_i, r_{mkt})}{\sum_i \sum_{j \neq i} corr(r_i, r_i)}) = \frac{\boldsymbol{\sigma}}{\sqrt[2]{\boldsymbol{\sigma^2+\omega^2}}} \Big/ \frac{\boldsymbol{\sigma^2}}{\boldsymbol{\sigma^2+\omega^2}} = \frac{\sqrt{\boldsymbol{\sigma^2+\omega^2}}}{\boldsymbol{\sigma}} > 1 \tag{A.7}$$

One can also note that the average correlation among stocks is the quadratic power of the average correlation with the index, so the ratio is the inverse of average correlation with the index. Since the correlation is always less than one, the inverse will always be greater than one.

This result also if coming out from an extremely simple model of stock returns can help understanding the nature of a financial crises and the nature of the shock hitting the economy. The increase in volatility that characterize a transition from a phase to another one can in fact derive from two sources. Is then the crisis originating from a macroeconomic shock hitting all the different companies in a very similar way? Or does it consist of a technological or socio-political shock affecting mainly some specific sectors or group of companies? Looking only to general level of the volatility is of no help in answering those questions. If we enlarge the analysis to the correlation matrix, we can have instead an answer.

# Appendix 2 - the data

## The stock market database

The database comprises all the stocks trading on the NYSE and the NASDAQ.[28] We focussed only on the closing prices of single stocks, excluding ETFs and other collective investment vehicles, but including foreign stocks listed as ADR. Since one of the issues that we want to tackle with the network analysis is the identification of the common factors behind the connectedness of the stock market, and an ideal candidate is the "industry" to which the listed companies belong, we will further restrict the database to the tickers for whom we succeeded in finding the industry classification. The resulting database comprises over 3,543 tickers, with the first of the 14,873 data points starting in 1963-01-01 and the last ones ending on 2020-06-30. Obviously, the timeseries for the 3,543 tickers are not complete, since at the beginning of the period only a subset of the companies existed.

Table A2.1 – Summary statistics about the industry distribution of the dataset

| Industry | Count | Mkt Cap (2020 06 30) | Weight |
|---|---|---|---|
| Health Care | 846 | 4,811,507 | 14.2% |
| Information Technology | 548 | 8,751,598 | 25.8% |
| Consumer Discretionary | 482 | 4,887,825 | 14.4% |
| Industrials | 464 | 3,002,623 | 8.9% |
| Financials | 413 | 3,371,699 | 9.9% |
| Energy | 211 | 921,549 | 2.7% |
| Communication Services | 183 | 3,905,977 | 11.5% |
| Materials | 155 | 942,621 | 2.8% |
| Consumer Staples | 142 | 2,288,419 | 6.8% |
| Utilities | 73 | 968,815 | 2.9% |
| Real Estate | 26 | 45,156 | 0.1% |
| **Total** | **3543** | **33,897,789** | **100.0%** |

As far as the variable representing the "market", we choose the S&P500 index.

## Google data

As it is well known, Google is the most popular search engine: it accounted for 87% of the global search market in July 2020, while Bing, the second search engine in terms of market share, accounted for just 6,4%. It is very likely that the activity by individual

[28] The dataset is available on Kaggle, here: https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs. The dataset is updated up to 2020-04-01. For the second quarter of 2020, this dataset was integrated downloading the price data from Yahoo Finance. As far as the S&P500, data were downloaded from …

investors of information gathering to prepare investment decisions is conducted almost entirely through the Google search engine.

It is not possible to directly query the searches database, but Google powers and maintains a website, Google Trends, that provide some insights into the most popular searches related to specific topic or terms across different geographical regions and languages. Obviously, one needs to be confident in the accuracy and timeliness of the results produced by Google Trends because, as we said before, there is no possibility to query the underlying database or to modify the algorithms working behind the UI of the website.

There are no publicly available API to query Google Trends. This implies that the job must be done manually or by using some "scraping" algorithm to automate the task.[29]

Among the analytical tools provided by Google Trends, "related searches" is probably the most interesting, at least for our purposes. For every term or topic, in fact, Google Trends also provides the terms that are most frequently searched in the same search session, within the chosen category, country, or region. Related searches are divided in two groups, "rising" and "top". As one can intuit, the "top" related searches are simply the most frequent over a very long-time horizon. Whereas the "rising" related searches are the one that show the most significant growth in volume in the requested period. We will use both.

[29] We used the open-source Pytrends library to build the scraping algorithm.

# Bibliography

Beirlant J, Goegebeur Y, Segers J, Teugels J. (2004). Statistics of Extremes: Theory and Applications. England: Wiley

Billio, M., Getmansky, M., Lo, A.W., Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. Journal of Financial Economics, 104, 535–559

Boginski V., Butenko S., and Pardalos P. (2005). Statistical analysis of financial networks. Computational statistics & data analysis 48.2, 431-443

Bonanno G., Caldarelli G., F. Lillo, S. Micciché, N. Vandewalle,, R. Mantegna (2004). Networks of equities in financial markets. Euro. Phys. J. B, 38, 363–371

Caraiani P. (2012). Characterizing emerging European stock markets through complex networks: From local properties to self-similar characteristics. Physica A, 391, 3629–3637.

Chi, K., Liu, T.J., Lau, C.M.F. (2010). A network perspective of the stock market. Journal of Empirical Finance, 17, 659–667

Chu J., Nadarajah S., (2017). A statistical analysis of UK financial networks. Physica A, 471, 445-459

Clauset A., Newman M.E.J., Moore C. (2004). Finding community structure in very large networks. Physical Review, E. 70 (6)

Elton E.J., Gruber M.J., Brown S.J., Goetzmann W.N. (2003). Modern Portfolio Theory and Investment Analysis. John Wiley and Sons

Epps T. W., Singleton K. J. (1986). An omnibus test for the two-sample problem using the empirical characteristic function. Journal of Statistical Computation and Simulation 26: 177–203

Gabaix X. (2016). Power Laws in Economics: An Introduction. Journal of Economic Perspectives, 30:1, 185–206

Gabaix X, Ibragimov R. (2008). Rank-1/2: A simple way to improve the OLS estimation of tail exponents. Work Pap., NBER

Goerg S.J., Kaiser J. (2009). Nonparametric testing of distributions—the Epps–Singleton two-sample test using the empirical characteristic function. The Stata Journal, 9(3), 454–465

Konishi M. (2007). A global network of stock markets and home bias puzzle. Appl. Financ. Econ. Lett., 3, 197–199

Huang W. Q., Zhuang X. T., Yao S. (2009). A Network Analysis of the Chinese Stock Market. Physica A

Lee K.E., Lee J.W., Hong B.H. (2007). Complex networks in a stock market. Comput. Phys. Commun. 177, 186.

Li P., Wang B. (2006). An approach to Hang Seng Index in Hong Kong stock market based on network topological statistics. Chin. Sci. Bull. 51, 624–629.

Mantegna R.N., (1999). Hierarchical structure in financial markets. Euro. Phys. J. B , 11, 193–197

Pareto V. (1896) Cours d'Economie Politique. Geneva: Droz

Ross S.A. (1977) The Capital Asset Pricing Model (CAPM), Short-sale Restrictions and Related Issues. Journal of Finance 32(2), 177-190;

Sharpe W.F. (1964) Capital Asset Prices: a Theory of Market Equilibrium Under Conditions of Risk. Journal of Finance 19(3), 425-442

The Economist (2020) Techtonic shifts. September 12-18, p. 61-62

Yong Shi, Yuanchun Zheng, Kun Guo, Zhenni Jin, Zili Huang (2020) The Evolution Characteristics of Systemic Risk in China's Stock Market Based on a Dynamic Complex Network. Entropy 2020, 22, 614

Zhuang X., Min Z., Chen S. (2007) Characteristic analysis of complex network for Shanghai Stock Market. Journal of Northeast. Univ. Nat. Sci., 28, 1053

Zipf G.K. (1949) Human Behavior and the Principle of Least Effort. Addison-Wesley, Cambidge MA

# Biographical note

### Marcello Esposito

Marcello Esposito, born in Milan in 1963, teaches International Financial Markets at the Cattaneo University of Castellanza. From 1990 to 2000 he was an economist at the Research Department of Banca Commerciale Italiana (now Intesa Sanpaolo), where he was head of Financial Markets Research; subsequently, he carried out several positions in the main Italian asset management companies. Since 2014 he has dedicated himself to consulting and has founded several fin-tech start-ups, as well as being engaged in numerous initiatives in the non-profit world. He holds a degree from Bocconi University (DES) and holds the MSc/MPhil in Economics from the London School of Economics.

LIUC
Università Cattaneo