REM WORKING PAPER SERIES

The 21st Century - Cluster Formation in the S&P 500

Maximilian Göbel and Tanya Araújo

REM Working Paper 043-2018

July 2018

REM – Research in Economics and Mathematics
Rua Miguel Lúpi 20,
1249-078 Lisboa,
Portugal

ISSN 2184-108X

The 21st Century - Cluster Formation in the S&P 500

Maximilian Göbel Tanya Vianna de Araújo*

June 2018

ABSTRACT

Since the beginning of the new millennium, investing in the stock market must have felt like a ride on the roller-coaster. The market went through every state between long-time troughs, trade suspensions and all-time highs. Such turbulent periods give all kinds of portfolio managers a hard time, with correlations of stock returns frequently varying. A deeper understanding of the dynamics within the market and the interlinkages between industry sectors is not only of added value to stock market participants, but also to policy makers, being responsible for implementing measures to counteract possible future macroeconomic imbalances. Therefore, this paper tries to uncover the community structure in the S&P 500 in the years 2000 through 2015 with additional focus on the effect of the *Great Recession* - especially on the financial sector.

*ISEG/UL - Universidade de Lisboa, Department of Economics; REM - Research in Economics and Mathematics; UECE - Research Unit on Complexity in Economics.

1 Introduction

Since the beginning of the 21st century financial markets have undergone turbulent times. Especially stock markets have experienced several ups and downs, while sucessively breaking through all-time highs. Focusing on the period between 01/03/2000 and 12/31/2015, the S&P 500 Index, which is generally regarded as a proxy for the overall stock market, lost almost 50% within two years after the bursting of the so-called "dot-com" bubble. Consequently, it reached its sample-period trough on 10/09/2002 at 677.187 points[1]. Afterwards, it took the S&P 500 almost exactly four years to fully recapture its peak value of the year 2000. The bullish market persisted even another twelve months, until rumors about falling housing prices, precedingly lax lending standards and a fragil financial system made the S&P 500 turn again on 10/09/2007 - exactly five years after its last trough. Within the following 18 months, the Index lost more than 55% of its turning-point value and - on 03/09/2009 - counted only 1.5 points more than its sample-period record low dated from 10/09/2002. Since then, bearish sentiment seems to have vanished, with not only the S&P 500, but stock indices all around the world ever rising. Between its trough in March 2009 and the end of the assessment period, the S&P 500 score rose by 236% in real terms. The bullish sentiment even persisted until early 2018, when the Index reached its all-time high, ranking at about 4.8 times above its value on 03/09/2009 and adjusted for inflation.

A complete assessment of the forces driving these fluctuations during the first decade of this century and the consistently bullish behaviour thereafter, demands the merging of economic, mathematical and psychological sciences. The literature on the behaviour of asset pricing, however, generally assumes random processes to be underlying the dynamics of stock returns (Mantegna, 1999). A vast literature on models, such as Fama & French (1992), tried to identify common factors of single stock returns and to shed further light on the seemingly random behaviour of stock returns.

Nevertheless, one commonly agreed perception about stock prices is their procyclicality (Cochrane, 1997; Kocherlakota, 1996; Mehra & Prescott, 2003). Thus, the correlation of stock returns plays a major role in portfolio construction and financial modelling. As

[1] Adjusted for Implicit Price Deflator: 31.12.2008 = 100.

Marvin (2015) points out, these correlations are, however, not constant over time and may even reverse during times of crisis. The widespread misperception of these dynamics by investors and financial product modellers in the beginning of the 2000s is said to have heavily amplified the severity of the *Great Recession* (Pozsar et al, 2010). Assessing the co-movements of stock returns and time-varying market structures are, thus, key to understanding and reacting to the effects of a changing macroeconomic environment.

Therefore, the purpose of this paper is to shed further light on the community structure of the S &P 500 Index in the years 2000 through 2015, by emphasizing the role of the *Great Recession*. The remainder of the paper is then structured as follows: section 2 gives a short summary of common clustering techniques, while section 3 presents the method of *optimal modularity*. Section 4 introduces the method of *symbolization* in order to facilitate the partition exercise of section 5. The last section concludes.

2 Literature Review

The methods for describing the co-movement of a system's individual components are summarized in the literature under the notion of cluster formation or community structure. Tan et al (2006) describe clustering as the grouping of items based on empirical data. In another study, Newman & Girvan (2004) define community structure as "the division of network nodes into groups within which the network connections are dense, but between which they are sparser" (Newman & Girvan, 2004, p.1.). The literature uses the notions of *nodes* or *vertices* to describe the single data points of a network, such as the individual stocks of a portfolio. *Edges* define the connecting links between these data points. Those methods originated initially in the field of physics, but have found entry into the works of economic and social scientists (Newman, 2003). For clustering time-series data, both the definition of an efficient algorithm and an adequate measure of similarity, respectively distance, between nodes are necessary features (Piccardi et al, 2011). To qualify as an appropriate proxy of distance, the respective measure must fulfill three requirements, necessary for defining a *metric* (Mantegna, 1999). Thus, a distance measure $d(i,j)$ between the nodes i and j can only be accounted for as a metric, if the following

conditions are jointly fulfilled:

$$d(i,j) = 0 \qquad \text{if and only if} \quad i = j \tag{1}$$

$$d(i,j) = d(j,i) \tag{2}$$

$$d(i,j) \leq d(i,k) + d(k,j) \quad . \tag{3}$$

Due to condition (1), the Pearson correlation coefficient cannot be regarded as a metric[2] and thus, does not qualify as a measure of similarity, respectively distance (Mantegna, 1999). Nevertheless, the literature uses it frequently as a basis for computing metric-compliant distance measures (Mantegna, 1999; Tan et al, 2006). For his analysis of traded stock portfolios, Mantegna (1999) transformed the correlation coefficient of stocks into a measure, equating the Eucledian Distance between two data points:

$$d(i,j) = \sqrt{2(1-\rho_{ij})} \, . \tag{4}$$

Mantegna (1999) then proceeded by constructing a *Minimal Spanning Tree* by sequentially linking the nodes with the lowest distance measure. This allows to efficiently assess the intensity of connections between stocks and different industrial sectors within a given portfolio. His analysis of the S&P 500 between 1989 and 1995 revealed strong intra-industry sector and intra-industry sub-sector connections, suggesting that, statistically, those stocks react to the same economic conditions.

In contrast to the correlation of stock returns as a basis for creating a proxy for distance, Marvin (2015) favours company related indicators such as revenues-to-assets ratio or net-income-to-assets ratio, as these are said to show higher persistence across varying economic states.

Tan et al (2006) apply a prototype-based algorithm, of which various designs exist (Marvin, 2015). The *k-means* approach requires each node within a cluster to be closer to the cluster's prototype than to the prototype of any other cluster. In the case of *k-means*, the prototype is the mean of the total sum of squared Euclidean distances between all the nodes within a cluster (Tan et al, 2006). Tan et al (2006) prespecify both the number of centroids and the centroids themselves before running the algorithm. The network's vertices are then connected to that centroid, which minimizes the squared Eucledian distance between the node under consideration and the centroid. Each centroid's value is

[2] $d(i,i) = 1$.

updated after each assignment, which can cause inter-cluster movements of already assigned nodes. The process stops as soon as an update does not affect the existing community structure (Tan et al, 2006). The weakness of the *k-means* approach arises from the necessity to pre-specify the initial centroids and to fix the number of clusters manually.

Nevertheless, the two methods described above have a long tradition in the literature. *K-means* belong to the class of *graph partitioning*, whereas the *Minimal Spanning Tree* technique by Mantegna (1999) is a type of *hierarchical clustering* or *community structure* detection. While the first assumes the number of clusters to be given exogenously and the number of intra-cluster nodes to be at least approximately fixed, the latter will endogenously determine both the number of communities within a network and the number of vertices within each cluster (Newman, 2006; Newman, 2010). Hierarchical clustering can be subdivided into two distinct forms: *agglomerative* and *divisive*. The difference between the two approaches results from the direction of filtering the single clusters: the first originates from an initially empty network and successively adds links between items according to a measure of *similarity*. The divisive approach of hierarchical structuring acts on the assumption of an initially completely specified network, from which the weakest links are removed iteratively (Newman & Girvan, 2004). Newman & Girvan (2004), however, critizise especially the agglomerative partitioning for its incapability to detect already known community structures and the tendency to neglect outliers.

For the analysis of the taxonomy of a portfolio of stocks, hierarchical clustering, respectively community structure detection, offers the appropriate features, as the number of clusters - if any - is determined by the network itself (Newman, 2006). Thus, for examining the community structure of the S&P 500 in the various periods between 2000 and 2015, this paper builds up on the community structure detection approach by Newman (2006) - the method of *optimal modularity*.

3 The Method of Optimal Modularity

The method of *optimal modularity* was initially introduced by Newman & Girvan (2004) as a measure to quantify the quality of a specific cluster formation within a network. Isogai (2015) describes it as a ”sort of distance between the actual density of the edges

and that of a randomized network" (Isogai, 2015, p.12.), which is most frequently used in modern cluster-detection analyses. The underlying principle of *modularity* is not - as stated by Newman (2006) - to minimize the number of edges between clusters, but to generate a structure which exhibits fewer edges than usually expected from a random network structure. Thus, Newman (2006) defines the modularity measure as the difference in the number of edges within a cluster and the number of edges in a random network. A random network is defined as one, in which the number of edges per node is preserved, whereas the connections are spread randomly throughout the network. The crucial assumption is that a random network does not exhibit any community structure (Isogai, 2015), whereas a large modularity suggests a large deviation of the detected community structure from a completely randomized network.

3.1 The Algorithm of Modularity

3.1.1 Division into Two Clusters

As stated in the inital paper on *modularity* (Newman & Girvan, 2004), the *modularity* measure Q builds up on Newman's (2003) *assortative mixing*, which is based on correlations of any properties of neighbouring nodes within a network. In the case of the S&P 500 Index, *assortative mixing* would measure the pairwise correlation of stock returns.

Thus[3], the initial network N consists of n nodes, respectively vertices. The algorithm starts by dividing the network into two communities. If a node i belongs to community 1, s_i will take the value of 1. If the node belongs to community 2, s_i will be -1. The number of connections, respectively edges, A_{ij}, between nodes i and j can either be 1 or 0. Every single A_{ij} displays an entry in the *adjacency matrix* of the whole network. Thus, the diagonal elements of the *adjacency matrix* have a value of 0, as:

$$A_{ii} = 0 \,. \tag{5}$$

What the literature calls the *degree of a vertex*, k_i, is the number of connections between each node i and all the other $n-1$ nodes within the network N. Thus, the total number of

[3]The following description will be mostly based on Newman (2006).

edges, m, within the network is:

$$m = \frac{1}{2}\sum_i k_i = \frac{1}{2}\sum_i A_{ij} \, . \tag{6}$$

The expected number of edges between vertices i and j, which can also be thought of as the probability of a random edge connecting vertices i and j, is given by:

$$\frac{k_i k_j}{2m} \, .$$

Remembering the goal of a clustering algorithm being to generate fewer than expected edges between communities of a network, the modularity measure Q is hence a function of the difference between the number of connections linking vertices i and j, $A_i j$, which are detected by the algorithm, and the expected number of edges $\frac{k_i k_j}{2m}$, while preserving the degree of each vertex i and j:

$$Q = \frac{1}{4m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right) s_i s_j \, . \tag{7}$$

If nodes i and j belong to the same cluster, the last term ensures that Q will take on a positive value and vice versa. Equation (7) can alternatively be written as:

$$Q = \frac{1}{4m}\mathbf{s}^T\mathbf{B}\mathbf{s} \, , \tag{8}$$

where $\mathbf{s}$ is the column vector with items s_i. B is a symmetric matrix consisting of the following elements:

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m} \, , \tag{9}$$

which Newman (2006) calls the *modularity matrix*.

As the elements within each row and each column of the *modularity matrix* sum to 0, B_{ij} always displays an eigenvector composed only of *ones* and an eigenvalue equal to 0. This feature accounts for the possibility that any division of the network N into several communities is suboptimal.

Next, all eigenvectors e_i of the *modularity matrix*, B_{ij}, will be normalized to u_i, such that each entry of $e_i = (a,b,c)^T$ will be divided by the length of the vector:

$$u_i = \left(a/\sqrt{a^2+b^2+c^2}, b/\sqrt{a^2+b^2+c^2}, /\sqrt{a^2+b^2+c^2}\right)^T \, . \tag{10}$$

The length of the normalized eigenvectors u_i will thus be equal to unity and two distinct normalized eigenvectors u_i and u_j are in addition orthogonal:

$$u_i \perp u_j \ ,$$

such that each eigenvector u_i is *orthonormal*.

Making use of:

$$\mathbf{s} = \sum_{i=1}^{n} a_i \mathbf{u}_i \ , \tag{11}$$

and the following dot-product:

$$a_i = \mathbf{u}_i^T \bullet \mathbf{s} \ , \tag{12}$$

the modularity measure Q of (7) can be transformed into:

$$Q = \frac{1}{4m} \sum_i a_i \mathbf{u}_i^T \mathbf{B} \sum_j a_j \mathbf{u}_j = \frac{1}{4m} \sum_{i=1}^{n} \left(\mathbf{u}_i^T \bullet \mathbf{s} \right)^2 \beta_i \ , \tag{13}$$

where β_i is the eigenvalue corresponding to the normalized eigenvector $textbf u_i$ of the *modularity matrix* $\mathbf{B}$.

As the network N still exists as a whole and hasn't been divided into any distinct communities, yet, the goal is now the partition into two clusters, such that the modularity measure Q is maximizied. Therefore, the eigenvalues β_i with $i = 1, ..., n$, corresponding to the orthonormal eigenvectors u_i, will be ordered with decreasing value.

Because of:

$$Q = \frac{1}{4m} \sum_{i=1}^{n} \left(\mathbf{u}_i^T \bullet \mathbf{s} \right)^2 \beta_i \ , \tag{14}$$

the maximization of Q would require $\mathbf{s}$ to be parallel to the eigenvector u_i corresponding to the largest eigenvalue β_i, which is $i = 1$ by convention. The reason is, that if $\mathbf{s}$ was parallel to u_1, then - due to orthogonality of two distinct eigenvectors u_i and u_j - the following condition would hold:

$$\mathbf{s} \bullet \mathbf{u}_j = 0 \qquad \text{with } \mathbf{s} \parallel \mathbf{u}_i \quad \text{for } i \neq j \ . \tag{15}$$

As the elements of $\mathbf{s}$ are restricted to ± 1, a parallelization of $\mathbf{s}$ and $\mathbf{u}_i$ is not possible. The closest approximation is a maximization of:

$$max \quad \mathbf{s} \bullet \mathbf{u}_1^T \ .$$

This is done by assigning a value of $+1$ to s_i, if the corresponding entry in $\mathbf{u}_1$ displays a positive prefix and vice versa.

Thus, the vertices i of the network N are assigned to each community according to the prefix of their corresponding value in $\mathbf{u}_1$. The value of an element i in eigenvector $\mathbf{u}_1$ gives an indication of the magnitude of change in modularity Q, if the corresponding node i was assigned to the other community. Thus, assigning vertices with a high value in $\mathbf{u}_1$, will have a larger impact on the maximization of Q than a vertex with lower value would have. This is especially important for continuing the clustering process for a further separation of the network N.

3.1.2 Division into N Clusters

After the first round of clustering, a further partitioning of the two clusters might be possible. The established community structure with all its vertices and both inter- and intra-cluster edges will be preserved and the algorithm will keep dividing each community g into two subgroups. The variable to be maximized is now the change in the modularity measure, ΔQ:

$$\Delta Q = \frac{1}{4m}\left[\sum_{i,j\in g} B_{ij}s_i s_j - \sum_{i,j\in g} B_{ij}\right] \tag{16}$$

$$= \frac{1}{4m}\sum_{i,j\in g}\left[B_{ij} - \delta_{ij}\sum_{k\in g} B_{ik}\right] s_i s_j \tag{17}$$

$$= \frac{1}{4m}\mathbf{s}^T\mathbf{B}^{(g)}\mathbf{s}\,, \tag{18}$$

where δ_{ij} is the Kronecker symbol, which takes on a value of 1 if i and j are located in the same community and 0 otherwise. $\mathbf{B}^{(g)}$ is a $n_g x n_g$ matrix, with n_g being the number of vertices within cluster g, which was generated in the previous round. The modularity matrix of cluster g, $\mathbf{B}^{(g)}$, consists of the following entries, describing the connection between vertices i and j:

$$B_{ij}^{(g)} = B_{ij} - \delta_{ij}\sum_{k\in g} B_{ik}\,. \tag{19}$$

As the rows and columns of B from (8) sum to 0, so do the rows and columns of $\mathbf{B}^{(g)}$. Thus, the same approach as mentioned above can be applied to maximize ΔQ.

The algorithm stops and the optimal number of clusters within network N is detected, if $\Delta Q \leq 0$, respectively any additional inter-cluster movement of vertices within the given community structure will not result in a positive ΔQ. This is equivalent to the leading eigenvalue β_1 of $\mathbf{B}^{(g)}$ taking the value of 0.

3.1.3 Transformation of Modularity for Stock Market Applications

When applying the modularity measure to time-series portfolio analysis, the algorithm has to be slightly adjusted. Following Piccardi et al (2011), equation (7):

$$Q = \frac{1}{4m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) s_i s_j \ , \tag{20}$$

will be transformed into a weighted network analysis, where the nodes i and j can be thought of as two distinct time-series of lenght T. The *degree* of node i, k_i, will be substituted by a measure of *strength* , w_i, of the time-series i. A measure for the *strength of the link*, respectively *similarity*, between the time-series i and j, $w_{ij} > 0$, will replace the adjacency matrix A_{ij} as follows:

$$Q = \frac{1}{4w} \sum_{ij} \left(w_{ij} - \frac{w_i w_j}{2w} \right) s_i s_j \ , \tag{21}$$

where the *total number of edges* within network N is replaced by the *total weight of edges*,w, in the network:

$$w = \sum_i w_i \ . \tag{22}$$

For measuring the *similarity* between time-series i and j the literature does not refer to a unique approach. Some researchers use distance measures, which are based on various techniques to calculate correlation coefficients, such as given in equation (4) (Eleutério et al, 2012).

Others, such as Piccardi et al (2011), model the *similarity* between two distinct time series, w_{ij}, as a non-linear function of the Euclidean distance ed_{ij}:

$$w_{ij} = f\left(ed_{ij}\right) \qquad , \qquad \frac{\partial f\left(d_{ij}\right)}{\partial ed_{ij}} < 0 \ . \tag{23}$$

By transforming the Euclidean distance measure, the authors assign a higher importance to highly similar time-series. However, all the various distance measures, d_{ij}, fulfill jointly conditions (1) through (3).

3.2 Criticism on Modularity

By refering to Fortunato & Barthélemy (2007), Isogai (2015) mentions the problem of resolution limit, which describes the weakness of modularity-based algorithms to detect relatively small communities within a network. The *modularity* measure struggles with distinguishing a single cluster from a combination of weakly interdependent small clusters. Isogai (2015) tries to circumvent this issue by first identifying the optimal community structure by global modularity maximization and then proceeding by applying the same approach for each detected cluster individually on a localized level.

Isogai (2015) also gives reason for caution concerning the Pearson correlation coefficient: as the distribution of stock returns might not be normal but exhibit fat tails - especially during times of crises as volatility increases - the linear correlation coefficient may display upwardly distorted values. Thus, Isogai (2015) first applies a GARCH to separate volatilities and returns. The generated residuals are then used to compute the correlation matrix for clustering. However, Isogai (2015) remarks that the level of *modularity*, whether based on the linear correlation matrix or on the GARCH-filtered one, does not reveal large differences. Only the statistical significance of the *modularity* Q is reduced, when based on the linear correlation coefficient.

4 Methodology

4.1 The *Method of Symbolization*

After discussing *modularity* as the benchmark according to which our network will be divided into distinct communities, this section will shed further light on how to prepare the raw data for community-structure analysis, before section 5 examines the actual clustering dynamics of the portfolio of 296 S&P 500 stocks.

It is important to keep in mind that the *method of symbolization* is only a technique to prepare the underlying raw data for the priorly described algorithm of *modularity*, which is the main criterion of for this paper's community structure detection procedure.

As already mentioned in Section 2, a community structure analysis of a portfolio of stocks requires an appropriate distance measure between individual stocks. The following

assessment will thus be based on the methodology of *symbolization* as applied by Brida & Risso (2007), Brida & Risso (2010) and Piccardi et al (2011). In particular, the two-step *symbolization* methodology of the latter paper will serve as the benchmark for the following data transformation:

Given the one-day öpgged differences of stock prices, r_{it}, of n stocks of time-series-length T, calculated as:

$$r_{it} = log\left(\frac{p_{i,t}}{p_{i,t-1}}\right) \quad , \tag{24}$$

- $p_{i,t}$ being the stock price of company i at time t - the first step is to generate a cumulative normal probability distribution for each of the n stocks over time T. The second step then calls for *symbolization*: according to pre-defined thresholds, the normalized logged stock returns, r_{it}, will be assigned a specific *symbol* as described in Piccardi et al (2011):

$$s_{it} = \begin{cases} 1 & \text{if } \mathrm{P}(r_{it}) \leq \frac{1}{3} \\ 2 & \text{if } \frac{1}{3} < \mathrm{P}(r_{it}) \leq \frac{2}{3} \\ 3 & \text{otherwise.} \end{cases} \tag{25}$$

This three-fold classification allows for a better differentiation between the different states of the economy and capturing the resulting volatility (Brida & Risso, 2007; Brida & Risso, 2010). Intervalls of equal length are furthermore said to optimally account for noisy time-series data (Molgedey & Ebeling, 2000). As proposed by Piccardi et al (2011), we proceed by computing the Euclidean distance between the time-series of any two stocks in our portfolio, based on the *symbolization*-algorithm explained above:

$$d_{ij} = \left(\sum_t \left(s_{it} - s_{jt}\right)^2\right)^{\frac{1}{2}} \tag{26}$$

As *modularity*, which determines the final community structure of our portfolio of S&P 500 companies, measures the *similarity* between individual stocks, the Euclidean distances, d_{ij}, have to be further transformed, such that high values of d_{ij} symbolize a low degree of *similarity* and vice versa. Piccardi et al (2011) remarks that unlike other networks, in which not every single node exhibits a theoretical relationship with any other node of the system,[4] a portfolio of stocks displays edges between any two of its constituents. This makes the detection of a community structure much more difficult and

[4] For example social network analyses.

requires the elimination of the weakest links between any two stocks to allow for a meaningful partition. In order to differentiate between weaker and stronger relationships, the Euclidean distances, d_{ij}, will be transformed by a non-linear and downward-sloping function $f(\bullet)$ resulting in the second round of *symbolization*.

The non-linear transformation of the $(n(n-1))/2$ Euclidean distances, d_{ij}, is achieved by computing a cumulative distribution function, such that:

$$\Pi(d_{ij}) = \mathrm{P}\left(d_{ij} \leq d\right) = \frac{\text{number of } d_{ij} \leq d}{n(n-1)} \quad . \tag{27}$$

Symbolization now adopts the downward-sloping feature of $f(\bullet)$:

$$w_{ij} = f(d_{ij}) = \begin{cases} 1 & \text{if} \quad \Pi(d_{ij}) \leq 0.025 \\ 0.1 & \text{if} \quad 0.025 < \Pi(d_{ij}) \leq 0.05 \\ 0.01 & \text{if} \quad 0.05 < \Pi(d_{ij}) \leq 0.1 \\ 0.001 & \text{otherwise.} \end{cases} \tag{28}$$

This transformation now allows to clearly differentiate between the degrees of *similarity* between any two stocks of our portfolio. Thus, the following analysis will omit the weakest interlinkages, depicted by $w_{ij} = 0.001$, in order to display a meaningful partition of our network. The drawback of this approach is the loss of information incorporated in 90% of the initial edges.

Nevertheless, the strongest 10% of relationships in this portfolio now form the weights for the *modularity* algorithm to reveal the portfolio's community structure.

4.2 *Gephi's* Modularity Measure

Having applied *symbolization* as a filtering technique to allow only the strongest relationships among companies to enter the community structure detection algorithm, this section briefly explains the *modularity* algorithm implemented in the software-package *Gephi*. It is based on Blondel et al (2008) and resembles a modification of the initial *modularity* measure as presented in section 3.1. *Modularity* is a widely used measure in the research literature, but its practical application is computationally extensive, especially for network structures with more than a million nodes (Blondel et al, 2008). For this reason, Blondel et al (2008) suggest an approximation procedure, which bypasses this obstacle.

Blondel's et al (2008) modification of Newman's (2006) *modularity* is a two-stage approach: at first, each single node i in the network N is assigned to a different community. Thus, initially the number of communities will equal the total number of nodes n in the network N. The algorithm proceeds by comparing the increase in *modularity*, if node i was attached to each of its neighbouring communities. Optimizing *modularity* thus requires to merge node i with the adjacent community, which generates the largest positive increase in *modularity* Q.[5] The first stage terminates as soon as no positive increase in Q can be detected. As Blondel et al (2008) give to consider, researchers being sensitive to computation time, shall be cautious about the order of node evaluation. As this paper's analysis only comprises a couple hundred of nodes, time issues are not a concern.

The second states builds up on the already established community structure throughout stage one. However, the nodes i are now represented by the communities detected in the first stage. Therefore, the weight of a link between two of the new nodes i is the sum of the weights of all the edges connecting nodes of the two first-stage-communities. Working with these new weights, the algorithm of stage one is now applied again, where the nodes of the second stage are defined as the communities of the first stage. In short, the second stage of the algorithm can be described as a merging of communities which maximizes the increase in *modularity* Q until no further improvement can be detected. Then once again, a new network is constructed, with the new nodes being the communities detected in the previous round and weights correspond again to the sum of the weighted links between two communities. Adjacent communities are merged so as to maximize the increase in *modularity*.

Blondel et al (2008) emphasize in particular the advantage of the short running time needed for appropriate results and the quick reduction of the number of communities. Furthermore, the *resolution limit problem*, which describes the struggle of *modularity* to detect small clusters, is mitigated by the algorithm's first stage (Blondel et al, 2008).

Thus, this paper's methodological approach is two-fold: at first the method of *symbolization* is applied to filter out the roughly 10% strongest relationships among companies based on the Euclidean Distance between the one-day logged differences of their stock prices. Contrary to Granovetter's (1973) theory, which holds rather the weak links

[5] see section 3.1 for definition of variables.

between nodes accountable for stronger and more widespread spill-over effects of information, this paper's analysis follows Piccardi et al (2011) by focusing on the strong relationships only. The second stage of the assessement comprises the help of *Gephi's modularity* algorithm, based on Blondel et al (2008), to identify cluster formation in the underlying network.

5 Clustering in the *Standard and Poor's 500* Index

5.1 Period: 01/03/2000 - 12/31/2015

5.1.1 Descriptive Statistics

The underlying data set comprises 296 companies of the *Standard&Poor's 500* Index, covering the period from 01/03/2000 to 12/31/2015. As shown in Table 1, these companies sort themselves into 11 distinct industrial sectors, with *Consumer Discretionary* representing the largest number of companies, such as *Amazon*, *Ford*, *Nike* or *Macy's*. The *Telecommunication Services* sector, in contrast, only covers *AT&T*, *Centurylink* and *Verizon Communications*. The strongest time-series correlation of stock returnsbetween two single entities is 0.88 and was detected in the Real Estate sector between *Avalon Communities* and *Equity Residential*.

In terms of average market-value[6], *Exxon Mobile* was with $352,797.54 billion the largest company, followed by *General Electric* with $298,148.52 billion. The smallest company according to the market-value indicator was the *Information Technology* company *Flir Systems* averaging $2,776.80 billion over the fifteen-year period.

Applying the filtering techniques described in section 4.1, the data set for the final clustering analysis reduces to 263 companies showing at least one interconnection with any of the other entities. As Table 2, the number of industry sectors was preserved, with *Consumer Discretionary* still forming the largest group of companies followed by *Financials* as displayed in Table 2. In terms of interlinkages, Table 3 proves *Financials* to be the most interconnected sector, forming relationships with 36.53% of all companies

[6]Defined as stated by *Bloomberg*: total current market value of all of a company's outstanding shares; capitalization is a measure of corporate size.

within the network. Financial companies are therefore twice as highly interconnected as the second largest industry sector, *Industrials*, representing 17.73% of all degrees.

On the single entity level, the company with the most interlinkages is the insurance and investment company *Lincoln National Corporation* with 165 degrees, followed by *American International Group* and *The Hartford Financial Services Group, Inc.* with 164 interlinkages remaining after having deleted the 90% weakest links from the initial network. Ranking as the 211th largest company in terms of market-capitalization[7], *Lincoln National Corporation* counts 76 degrees more than the second largest company *General Electric* (89) and 144 more than the leading *Exxon Mobile* (21). This already suggests, that a clear-cut relationship between a company's size and its interconnectedness within the network does not exist. Indeed, the 5% largest companies only represent 4.18% of the network's edges.

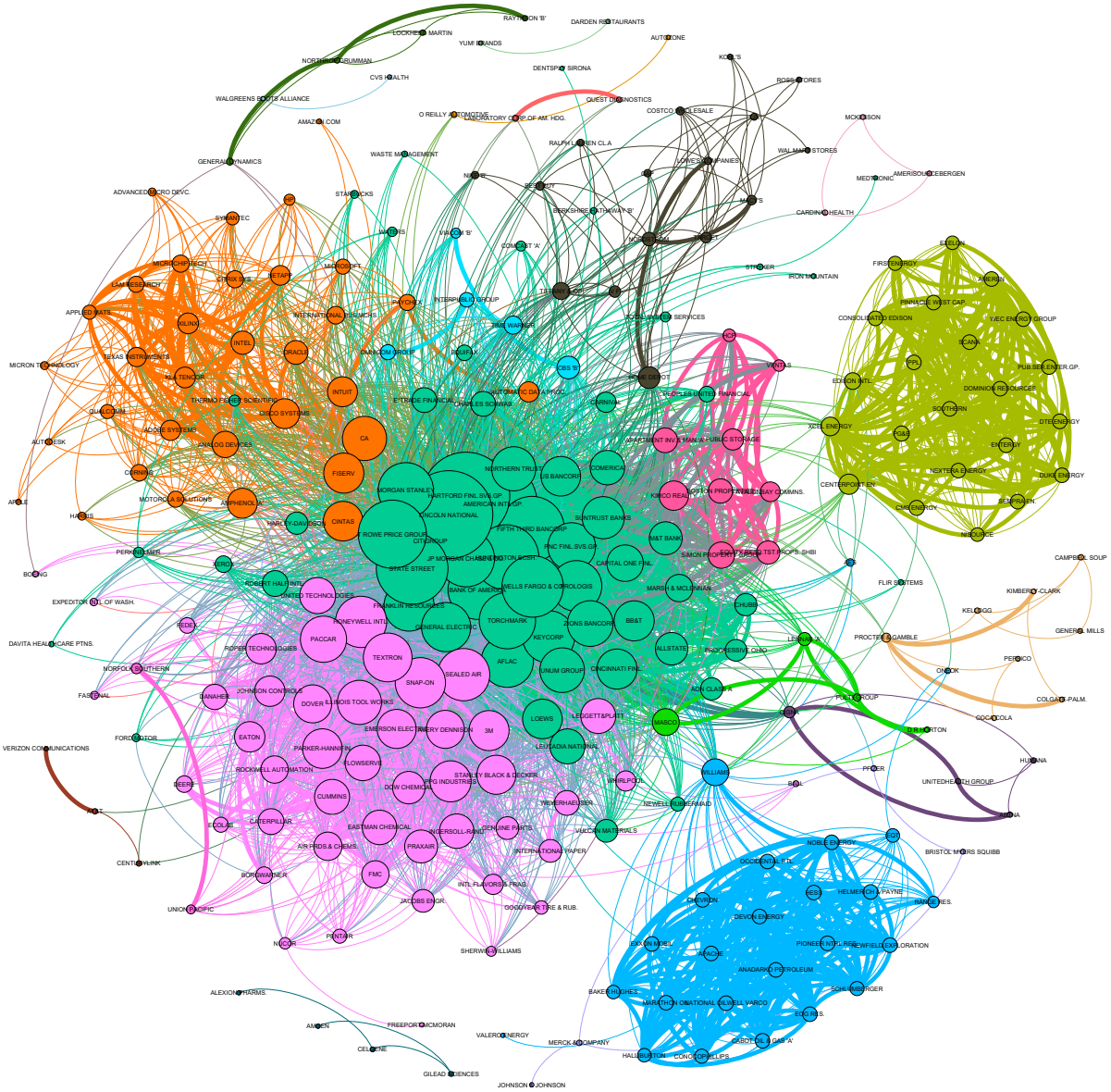
On an aggregate level, the average degree of an entity within the system is slightly higher than 33, whereas 13 companies are left with being connected to only one of the other 262 firms within the system. A comparison of the average degree with its median value of 21 suggests that the portfolio consists of a relatively small number of companies being highly interconnected with the other companies within the system: the five most interconnected companies - which all correspond to the financial sector - comprise almost 9% of the network's interlinkages. With only one of them - *CitiGroup* - ranking among the top-ten largest companies, this is further evidence against a direct relationship between size and interconnectedness.

5.1.2 Community Structure

As already stated in the introductory paragraph, a closer look at the community structure of a portfolio of S&P 500 companies shall not only shed further light on portfolio composition strategies, but also provide further insights into industrial networking.

Figure 1 shows the clustering result of all 263 companies for the period 01/03/2000 to 12/31/2015. The colours represent the 21 different clusters within the overall network resulting in a *modularity* of 0.595. This measure is in line with previous research as already

[7]Defined as stated by *Bloomberg*: total current market value of all of a company's outstanding shares; capitalization is a measure of corporate size.



Note: Colours display different clusters. Colours do not have any interpretational purpose, but are just randomly applied for a more convenient visualization.

Newman & Girvan (2004) suggested that typical *modularity* measures range between 0.3 and 0.7. In his analysis of 30 time-series of the *Dow Jones Industrial Average* between 2001 and 2006, Piccardi et al (2011) detected 7 clusters with a *modularity* of 0.679.

The clustering depicted in Figure 1 is characterized by an agglomeration of several clusters which seem to exhibit not only strong intra- but also inter-cluster linkages, whereas the two clusters in black and orange are quite isolated. The network is complemented by a large number of smaller clusters with apparently weaker links to the bigger communities.

As Figure 1 only differentiates between different communities, a closer look at the

distribution of the several sectors among clusters might provide a depper understanding of intra-industry co-movements of companies.

Figure 2 depicts again a partition in patterns of clusters. Whereas the sectors portrayed by Figures 2c, 2d, 2e and 2i are quite concentrated on forming a cluster on their own, the stock returns of sectors such as *Consumer Discretionary* or *Health Care*, Figures 2a and 2f, are spread out among several clusters. Thus, the latter two sectors do not seem to be affected by intra-industry developements but are rather influenced by extra-industry-sector shocks.

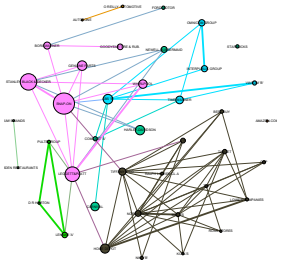
After an overview of the community structure formation and the distribution of industry sectors among clusters, a look at the specific composition of the single clusters might allow for a more thorough insight into industry sector dynamics: therefore, the colours in Figure 3 now represent the different sectors and the partition of Figure 4 shows the composition of the largest communities within the network. Worth noting is again the two-fold picture in Figure 3: on the one hand, strong inter-industry sector linkages of a subset of sectors including *Financials* and *Industrials*, and on the other hand two seemingly isolated sector-specific clusters of the *Energy* and *Utilities* sector. Although only hardly visible, the *Real Estate* seems to be highly interconnected with the financial sector. However, only an extension of the assessment period permits an analysis of possible reasons for this formation, such as the contribution of increased securitization of mortgages. Nevertheless, its central positioning already indicates that a crisis hitting the financial system may infiltrate other sectors or even the whole real economy.

Figure 4 presents the largest nine clusters and the remaining ones as an aggregate in Figure 4j. Preserving the sector colours, Figure 4a confirms the interconnectedness of financial companies not only with own-sector entities and proves its role as a catalysator of sector-specific risk. Figures 4d and 4e emphasize the impression of Figure 3 as largely isolated communities, whereas also at least some *Real Estate* companies form a unique community as referenced by Figure 4g.

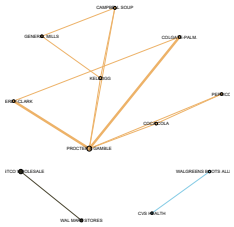
5.2 Period: 12/01/2007 - 06/30/2009

Having examined the cluster formation of a subset of the S&P 500 Index for the whole first 15 years of the 21st century, this section provides an assessment of the community

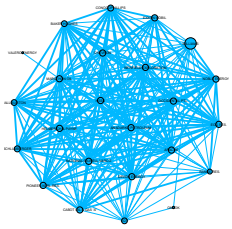
Figure 2: Composition of Industry Sectors by Clusters



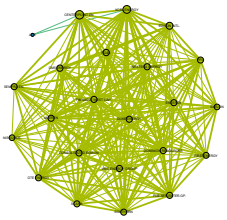
(a) Consumer Discretionary (1.76%)



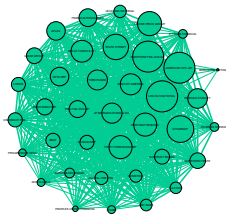
(b) Consumer Staples (0.3%)



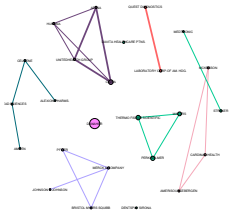
(c) Energy (5.16%)



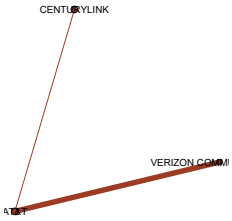
(d) Utilities (5.27%)



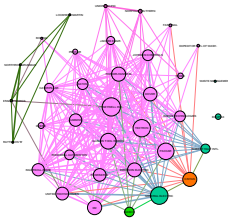
(e) Financials (14.74%)



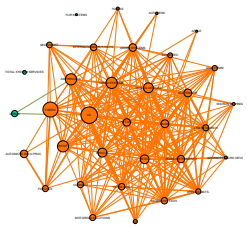
(f) Health Care (0.48%)



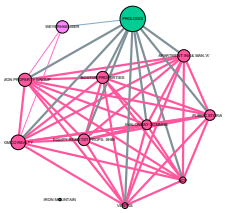
(g) Telecommunications (0.05%)



(h) Industrials (5.98%)



(i) Information Technology (6.07%)



(j) Real Estate (1.1%)

(k) Materials (1.67%)

Note: Numbers in brackets indicate the ratio of intra-industry sector edges to total number of edges within the network.

Figure 3: Industry Sector distribution of 263 Companies of the *S&P 500* Index

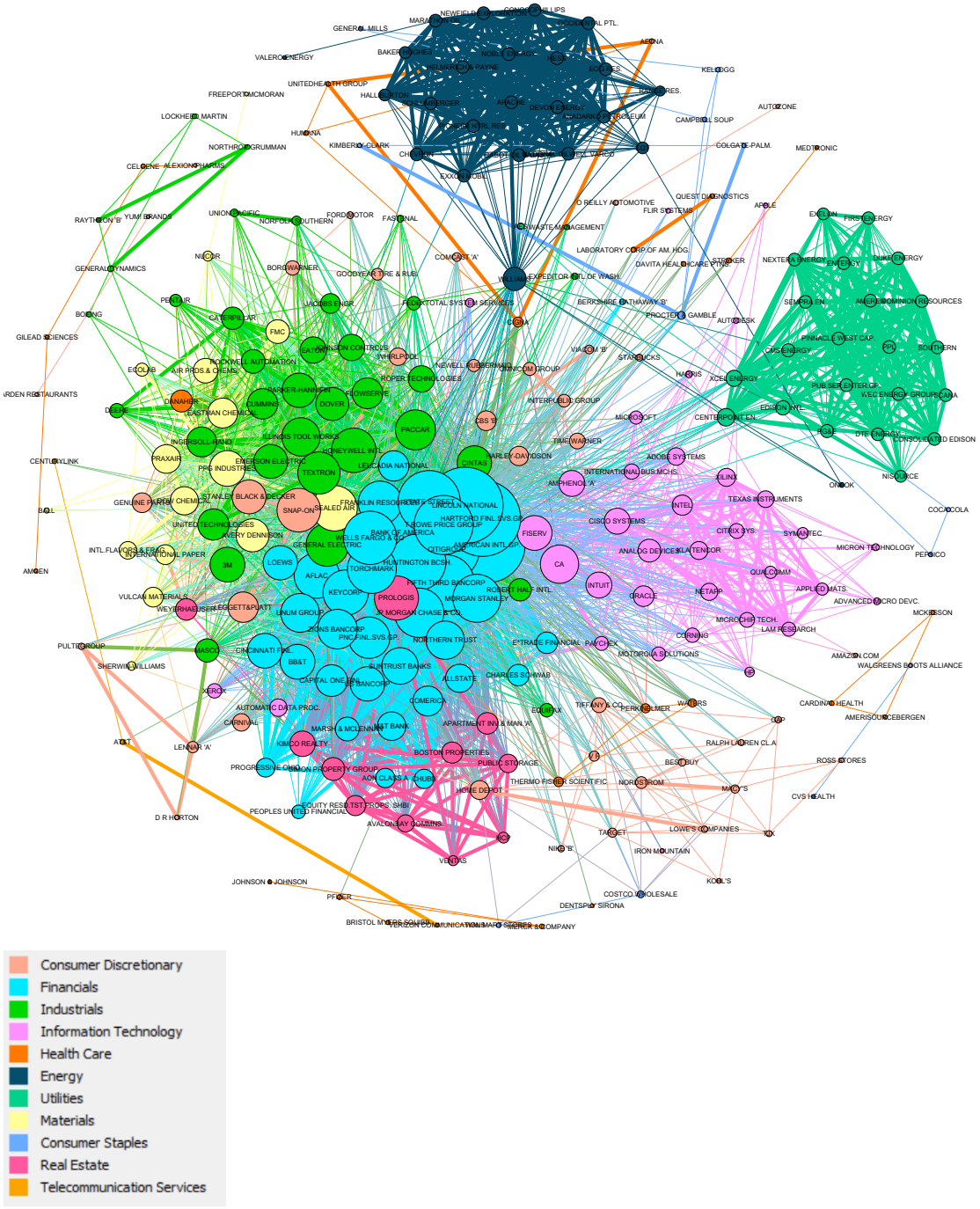
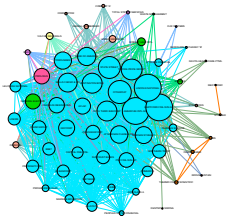
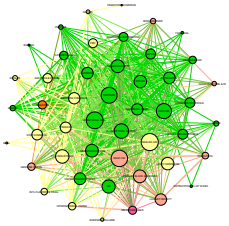


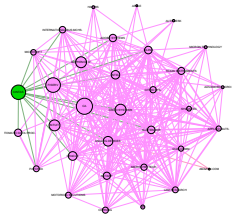
Figure 4: Composition of Clusters by Industry Sectors



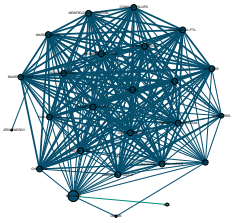
(a) Largest Cluster: 61 Nodes; (21.77%)



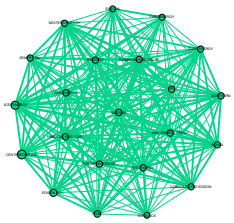
(b) 2nd largest Cluster: 46 Nodes; (13.04%)



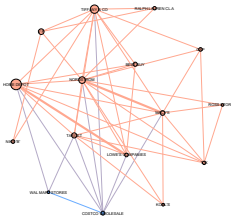
(c) 3rd largest Cluster: 33 Nodes; (6.39%)



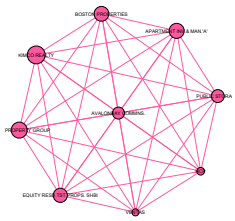
(d) 4th largest Cluster: 25 Nodes; (5.18%)



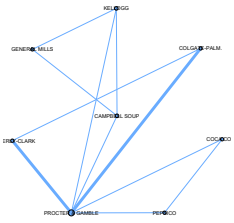
(e) 5th largest Cluster: 22 Nodes; (5.23%)



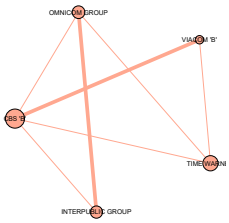
(f) 6th largest Cluster: 16 Nodes; (1.1%)



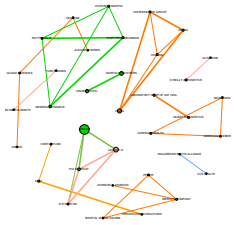
(g) 7th largest Cluster: 9 Nodes; (0.83%)



(h) 8th largest Cluster: 8 Nodes; (0.25%)



(i) 9th largest Cluster: 5 Nodes; (0.25%)



(j) 12 smallest Clusters combined: 35 Nodes; (0.83%)

Note: Numbers in brackets indicate the ratio of intra-cluster edges to total number of edges within the network.

structure during the *Great Recession* - the biggest economic turmoil since the *Great Depression* - with a particular emphasis on the financial sector.

However, the accompanying question adresses the definition of the crisis period: was the trigger of the crisis the fall in housing prices, being represented by the turning point in the widely used *Case-Shiller Index*? Was it the onset of tightening refinancing conditions in the interbank market in the summer of 2007, when *money market mutual funds* struggled to not *break the buck*, or was it the run on the investment bank *Bear Stearns* in March 2008? Another, more quantitative criterion to identify abnormal periods, especially in the stock market, is the Volatility Index (VIX) of the *Chicago Board Options Exchange*. The VIX represents the 30-day expected volatility of the U.S. stock market, using prices of call- and put-options on the S&P 500 Index as a basis for calculation (CBOE, 2018). Cerutti et al (2017) define a period to be characterized by financial stress, if the close-of-quarter value of the VIX surpasses the value of 30. Evaluating this proxy for risk-aversion and uncertainty of market participants (Rey, 2015) on a monthly basis accordingly, identifys the crisis period to be starting in September 2008 and finishing in April 2009. Although the stock market is said to be strongly correlated with the business cycle, using the VIX as a criterion for crisis identification might cause biased results, as the calculation of the VIX uses data derived from the same variables serving as raw data for this paper's assessment process.

Therefore, referring to the recession-dating indicator of the *National Bureau of Economic Research*, this paper assumes the *Great Recession* to be described by the 19 consecutive months between December 2007 and the end of June 2009.

5.2.1 Descriptive Statistics

Before investigating the community structure of the same initial 296 stocks within the above-defined period of economic turmoil, a look at some stylized facts may already give an idea about the exposure of certain industry sectors to the recent financial crisis.

After applying the *symbolization* and filtering techniques of section 4.1, again 263 companies form the input for the clustering based on the *modularity* measure. Nevertheless, Figure 7 demonstrates the effect of a period characterized by financial distress on the dynamics of one-day logged differences of stock returns in the S&P 500. Both, the

shift in correlations to higher levels and the overall reduction in the pairwise Euclidean distances, proxying the *similarity* of the portfolio's companies, are an indication of increased co-movement of stocks during the*Great Recession*. This also caused the set of companies under investigation to slightly differ from the previous analysis, as only 252 companies entered both algorithms.

These results are also apparent from Tables 2 and 3: once again all 11 industry sectors were covered in the analysis with *Consumer Discretionary* still covering the largest number of companies (42), but now followed by both *Financials* and *Industrials* with 37 entities. In terms of interconnectedness, the sector of *Financials* suffered a drop of almost 15 percentage points to 22.96%, which underscores the aforementioned increased instability of financial institutions during the recent financial crisis. Furthermore, the sector's representation in the top-five interconnected companies dropped to one, while *Consumer Discretionary* covered 60% and one company of the *Industrials* sector complements that group. Despite being the industry sector to represent most companies in the sample, *Consumer Discretionary* was only weakly interconnected within the system relative to other sectors. However, it was that sector, which experienced the largest rise of interlinkages with more than 70%. The *Real Estate* sector could experience an equal rise in strong relationships.

Zooming in on the individual company level, it was *Loews Corporation* which became the most interlinked company during the crisis. Being classified as part of the *Financials* by Standard & Poor's, *Loews Corporation* describes itself as "one of the largest diversified companies in the United States, with businesses in the insurance, energy, hospitality and packaging industries" (Loews, 2018). Nevertheless, it only ranked number 126 in terms of overall period's market value. A look at the five most interconnected companies during the crises and their ranking according to the full period's average market value tempts to infer a negative correlation of a companies *size*, proxied by its market capitalization, and *similarity*, proxied by the Euclidean Distance, during times of crisis. The Pearson correlation coefficient of 0.13, however, invalidates such a conclusion.

With the average degree of a company having slightly fallen to 32 and the median degree having risen to 23, the distribution of the degrees seemed to have become more center-driven. Even though, the number of firms keeping only a single interrelation with

any of the remaining 262 firms, rose from 13 to 19. However, also the share of total degrees covered by the five most interconnected companies dropped from almost 9% to 6.5%. Even though these figures might suggest that the *Great Recession* might have countributed to a higher level of isolation on a single company basis, Figure 7b clearly shows the overall tendency of company-dynamics to converge to similarity during times of economic turmoil.

5.2.2 Community Structure

This subsection is intended to shed further light on the effect of the recent financial crisis on the community structure of a sub-portfolio of the S&P 500 Index by means of *modularity*. In particular, the following analysis focuses on whether a period of increased financial stress tends to make companies stick together, resulting in less clusters, or whether companies try to seperate in order to isolate themselves from possible negative spill-over effects. An increased variety of clusters would as well indicate a growing risk-sensitivity and willingness for in-depth company analysis by stock market investors during times of bearish economic outlooks.

Compared to the whole assessment horizon, the *modularity* increased during the crisis period from 0.595 to 0.621, indicating a slightly more clear-cut cluster formation. Remarkable, however, is the decrease in the number of identified clusters from 21 to 12. This leaves plenty of room for causal interpretation: on the one hand, the reduction in clusters is evidence for the impact of the *Great Recession* not being limited to an individual market segment, as the the crisis is said to have emerged from the housing sector, spilling over to financial instiutions and to the whole real economy, . On the other hand, the shift in clusters as well as the stock return dynamics depicted in Figure 7 reflect the effects of fire-sales triggering the shift to safe havens, due to increased risk-aversion and the fear of large losses.

Figure 5 depicts the community structure of the 263 companies with each industry sector being coloured differently. Compared to Figure 3, the emergence of the *Consumer Discretionary* sector is apparent. A second look suggests a slight shift and a tendency towards isolation by the *Financials*. The *Energy* sector seems to have moved from a quite isolated position in Figure 3 to a more interlinked community, especially with *Industrials*.

The increased *modularity*, indicating a more clear-cut differentiation between clusters, is hardly visible and can only be confirmed by a closer look. Nevertheless, both figures confirm the non-uniform community structure of the stock market across varying states of the economy.

Last but not least, Figure 6 allows a comparison of the financial sector's interconnectedness with the system as a whole. A first look suggests the sector's loss of its central role as doubts about its stability became widespread among market participants. Even though a loss of aggregate interconnectedness, symbolized by the amount of degrees as displayed in Table 3, might not be apparent from the visualizations, the size of the nodes being connected to the financial sector has changed. Whereas the nodes were rather small in Figure 6a, the company-specific amount of connections of entities being linked to the financial sector has risen in the months representing the crisis period.

Thus, the financial sector seemed to have either shifted its interlinkages to companies, which are highly interconnected themselves, or the companies, to which the financial sector had already established relationships with, became highly interconnected throughout the period of increased financial stress themselves.

Understanding the reasons and drivers of the generally changed landscape of the network structure is crucial for understanding the dynamics of not only stock markets, but also of investors' decision making across varying economic states. A more in-depth assessment is open to future research.

6 Conclusion

The first 15 years of the 21st century have been marked by several ups and downs in the stock market. As the stock market is said to mimic the course of the business cycle and thus the behaviour of the overall economy, a more thorough understanding of the market's dynamics may allow policy makers to base their decisions on improved economic forecasts. Undoubtably, knowledge about the differences in aggregate stock market behaviour and the interplay of various industry sectors during times of crises and normal times is of great importance to portfolio managers and stock market participants.

This paper, hence, tried to shed further light on the market dynamics, proxied by 296

Figure 5: Industry Sector distribution of 263 Companies of the *S&P 500* Index (12/01/2007 - 06/30/2009)

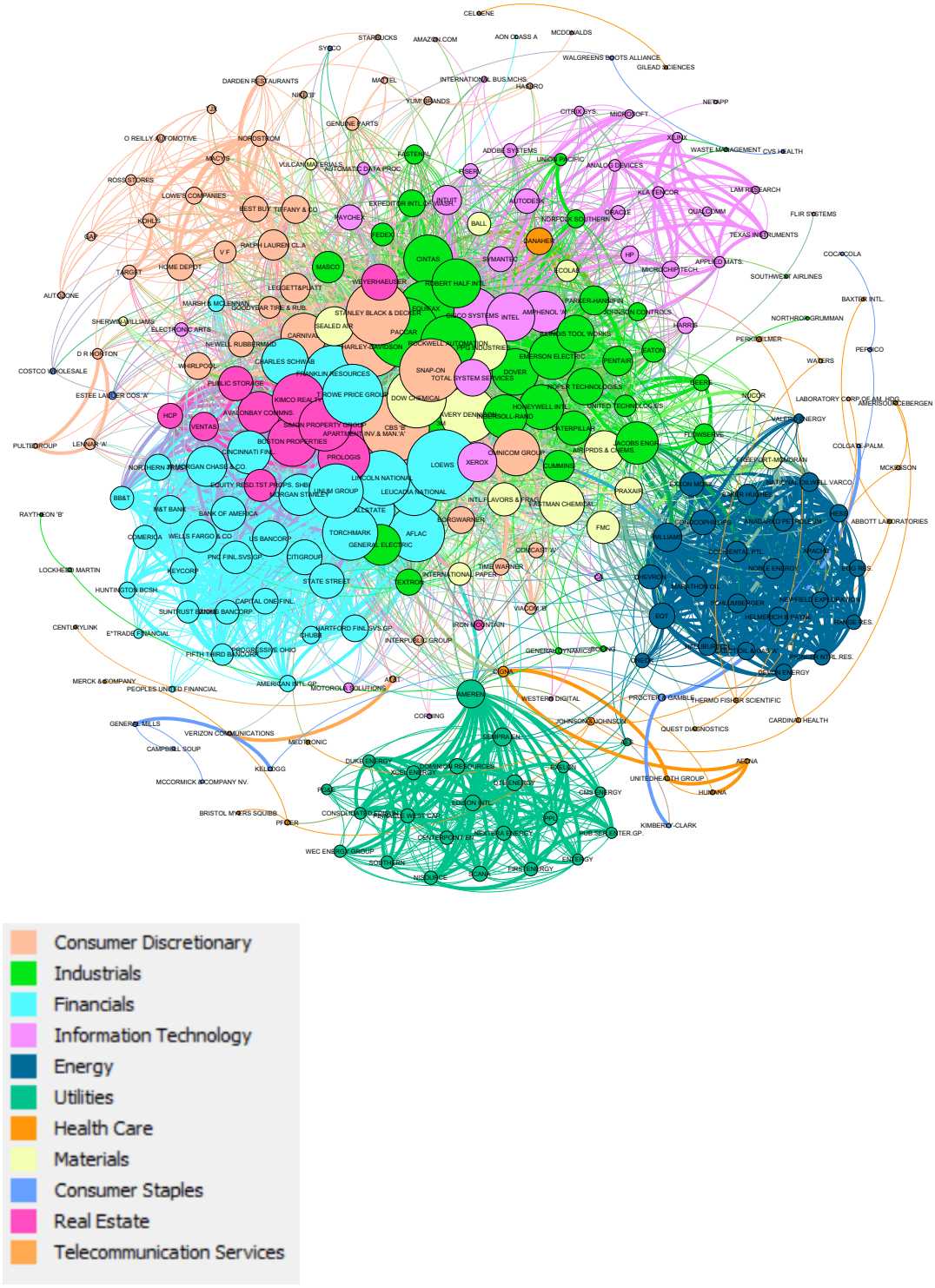
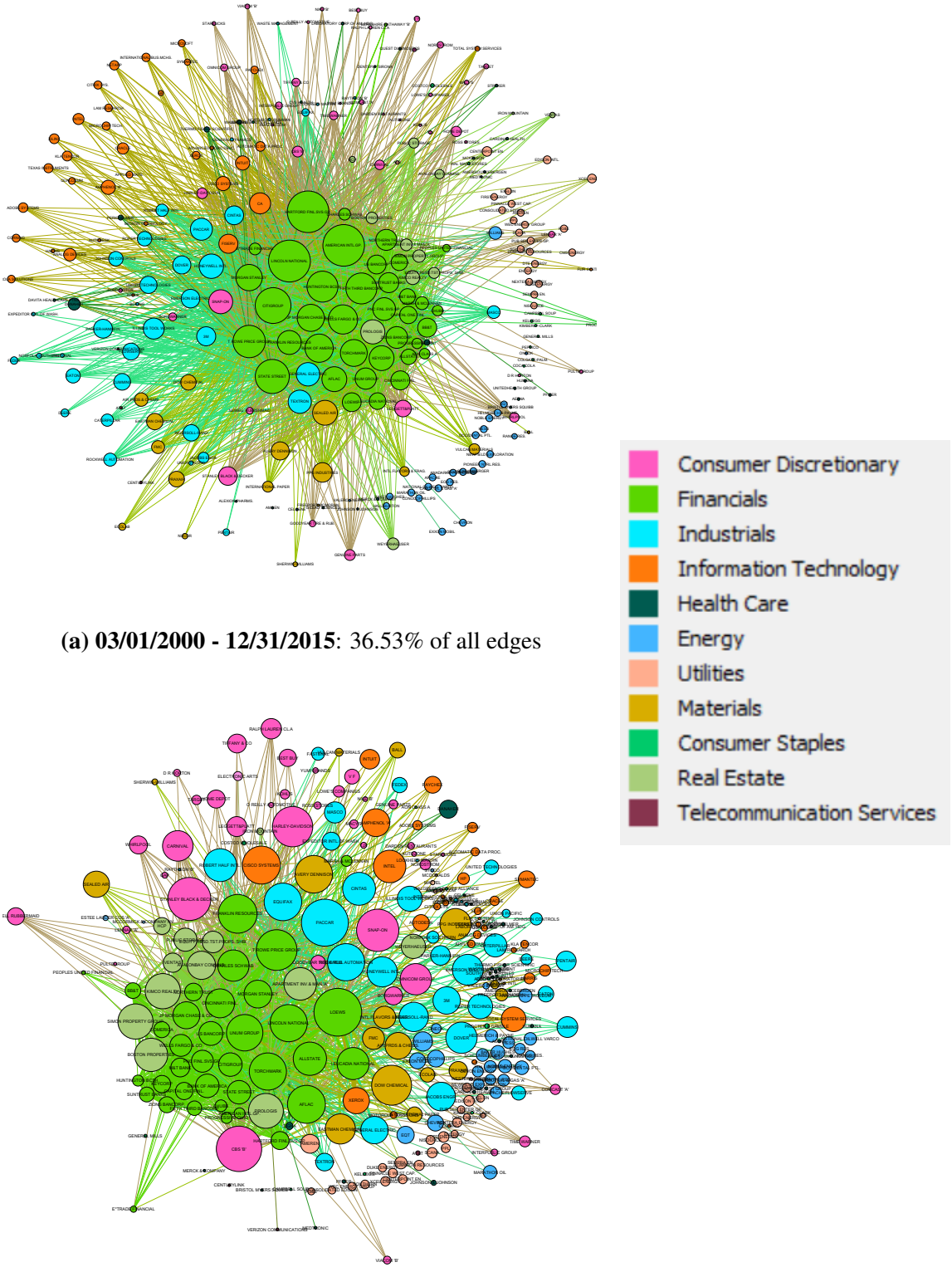


Figure 6: Interconnectedness of the Financial Sector



(a) 03/01/2000 - 12/31/2015: 36.53% of all edges

(b) 12/01/2007 - 06/30/2009: 22.92% of all edges

companies of the S&P 500 between 2000 and end of 2015. Using the measure of *modularity* combined with *symbolization* allowed a detailed assessment of cluster formation in the stock market. The Euclidean distance of one-day logged differences of stock prices served as a measure to describe the *similarity* between two single stocks.

The data revealed strong interlinkages of the financial sector with the overall community, whereas the *Great Recession* lead to a sharp decline in the sector's interconnectedness. Moreover, a direct relationship between a company's size and its connectivity to the overall system could not be backed up by the empirical analysis. The value of the *modularity* measure conforms with earlier studies of network analyses and allows to claim that a certain community structure is indeed prevailing in the stock market as well. Even though the visualizations show a quite dense picture with no trivial cluster differentiation, they clearly demonstrate the non-stationarity of partition, varying with the state of the economy.

Although this paper's analysis is limited to a time frame of 15 years, it may serve as an encouragement for further research extending the assessment period and analyzing patterns during other times of economic stress. Furthermore, the determinants driving the specific partitioning and the reasons for their strong non-stationarity are still unknown.

Appendix

Table 1: Number of Companies per Industry Sector

- before *Symbolization* -

	Absolute	Percent
Industrials	38	12.84%
Health Care	35	11.82%
Information Technology	36	12.16%
Utilities	23	7.77%
Financials	38	12.84%
Materials	17	5.74%
Consumer Discretionary	45	15.20%
Energy	24	8.11%
Real Estate	12	4.05%
Consumer Staples	25	8.45%
Telecommunication Services	3	1.01%
Sum	296	100.00%

Table 2: Number of Companies per Industry Sector

- after *Symbolization* -

Industry Sector	Period			
	01/03/2000 - 12/31/2015		12/01/2007 - 06/30/2009	
	Absolute	Percent	Absolute	Percent
Industrials	36	13.69%	37	14.07%
Health Care	25	9.51%	22	8.37%
Information Technology	34	12.93%	33	12.55%
Utilities	23	8.75%	23	8.75%
Financials	38	14.45%	37	14.07%
Materials	16	6.08%	16	6.08%
Consumer Discretionary	40	15.21%	42	15.97%
Energy	24	9.13%	24	9.13%
Real Estate	12	4.56%	12	4.56%
Consumer Staples	12	4.56%	14	5.32%
Telecommunication Services	3	1.14%	3	1.14%
Sum	263	100.00%	263	100.00%

Table 3: Number of Degrees per Industry Sector

- after *Symbolization* -

Industry Sector	Period			
	01/03/2000 - 12/31/2015		12/01/2007 - 06/30/2009	
	Absolute	Percent	Absolute	Percent
Industrials	1547	17.73%	1652	19.60%
Health Care	142	1.63%	103	1.22%
Information Technology	957	10.97%	816	9.68%
Utilities	489	5.60%	425	5.04%
Financials	3188	36.53%	1932	22.92%
Materials	662	7.59%	774	9.18%
Consumer Discretionary	774	8.87%	1281	15.20%
Energy	476	5.45%	665	7.89%
Real Estate	445	5.10%	734	8.71%
Consumer Staples	39	0.45%	39	0.46%
Telecommunication Services	7	0.08%	9	0.11%
Sum	8726	100.00%	8430	100.00%

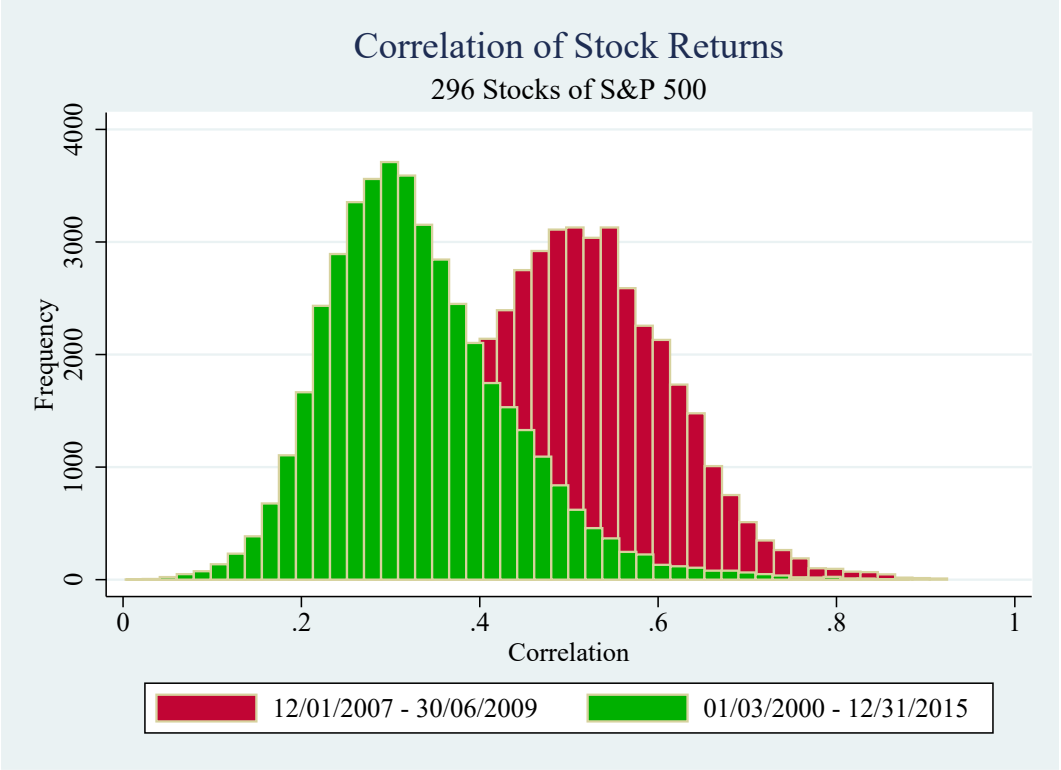
Table 4: Number of Degrees - Summary Statistics

- after *Symbolization* -

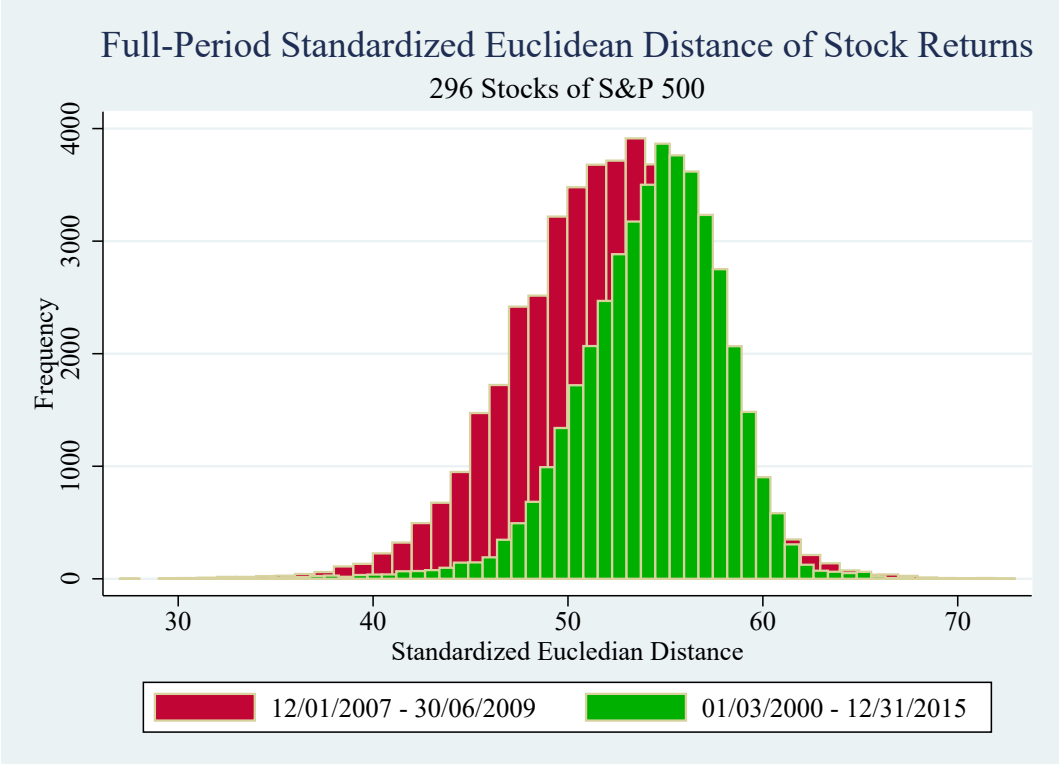
	Period	
	01/03/2000 - 12/31/2015	12/01/2007 - 06/30/2009
Total Degrees	8726	8430
Maximum Degrees	165	117
Minimum Degree	1	1
Frequency of Minimum	13	19
Average Degree	33.18	32.05
Median Degree	21	23

Figure 7: Dynamics of Stock Returns

(a) Correlation of Stock Returns



(b) Euclidean Distance of Stock Returns



References

[1] Blondel, V. D., Guillaume, J.-L., Lamboitte, R. and Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment* 10.

[2] Brida, J. G. and Risso, W. A. (2007). Dynamics and Structure of the Main Italian Companies. *International Journal of Modern Physics C* 18.

[3] Brida, J. G. and Risso, W. A. (2010). Dynamics and Structure of the 30 Largest North American Companies. *Comput Econ* 35, 85-99.

[4] Cerutti, E., Claessens, S. and Rose, A. K. (2017). How Important is the Global Financial Cycle? Evidence from Capital Flows. *IMF Working Paper Series* WP/17/193.

[5] Chicago Board Options Exchange Global Markets. *VIX Index* [Online]. Available from: http://www.cboe.com/vix [Accessed: 05/06/2018].

[6] Cochrane, J. H. (1997). Where is the Market Going? Uncertain Facts and Novel Theories. *NBER Working Paper Series* (6207).

[7] Eleutério, S., Araújo, T. and Mendes, R. V. (2012). Portfolios and the Market Geometry. *Lisbon School of Economics and Managment Working Papers*, No. 0874-4548.

[8] Fama, E. F. and French, K. R. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance* 47(2), 427-465.

[9] Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America* 104(1), 36–41.

[10] Granovetter, M. S. (1973). The Strength of Weak Ties. *The American Journal of Sociology* 78(6), 1360-1380.

[11] Isogai, T. (2015). Clustering of Japanese Stock Returns: Statistical Analysis of the Correlation Structure of Fat-Tailed Returns [Online] Available from: https://dspace.jaist.ac.jp/dspace/bitstream/10119/12872/11/paper.pdf [Accessed: 21/05/2018].

[12] Kocherlakota, N. R. (1996). The Equity Premium: It's Still a Puzzle. *Journal of Economic Literature* 34, 42-71.

[13] Loews Corporation. *Loews Corporation - Overview* [Online]. Available from: https://loews.com/overview [Accessed: 06/05/2018].

[14] Mantegna, R. N. (1999). Hierarchical Structure in Financial Markets. *The European Physical Journal B* 11, 193-197.

[15] Marvin, K. (2015). *Creating Diversified Portfolios Using Cluster Analysis* [Online]. Available from: https://www.cs.princeton.edu/sites/default/files/uploads/karina_marvin.pdf [Accessed: 05/21/2018].

[16] Mehra, R. and Prescott, E. C. (2003). The Equity Premium in Retrospect. *NBER Working Paper Series* (9525).

[17] Molgedey, L. and Ebeling, W. (2000). Local Order, Entropy and Predictability of Financial Time Series. *The European Physical Journal B* 15, 733-737.

[18] Newman, M. E. J. (2003). Mixing Patterns in Networks. *Physical Review E* 67(2), 1-14.

[19] Newman, M. E. J. (2006). Modularity and Community Structure in Networks. *Proceedings of the National Academy of Sciences* 103(23), 8577–8582.

[20] Newman, M. E. J. (2010). *Networks - An Introduction*, 1st Ed. New York: Oxford University Press.

[21] Newman, M. E. J. & Girvan, M. (2004). Finding and Evaluating Community Structure in Networks. *Physical Review E* 69(2), 1-16.

[22] Piccardi, C., Calatroni, L. and Bertoni F. (2011). Clustering Financial Time Series By Network Community Analysis. *International Journal of Modern Physics C* 22(1), 35-50.

[23] Pozsar, Z., Adrian, T., Ashcraft, A. and Boesky, H. (2010). Shadow Banking. *Federal Reserve Bank of New York Staff Reports* 458.

[24] Rey, H. (2015). Dilemma not Trilemma: The Global Financial Cycle and Monetary Policy Independence. *NBER Working Paper Series* (21162).

[25] Tan, P.-N., Steinbach, M. and Kumar, V. (2006). *Introduction to Data Mining*, 1st Ed. Boston: Pearson Education, Inc.