Market Graph Analysis of the Portuguese Stock Market

by

Luís Pedro Airosa Carvalho Brás

Master's Thesis in Data Analytics

Supervised by

Prof. Dalila B. M. M. Fontes

Faculdade de Economia

Universidade do Porto

2017

Biographical Note

Luís Pedro Airosa Carvalho Brás was born on February 1st 1993, in Oporto, Portugal.

Luís enrolled in the Bachelor in Economics in 2011 at Faculdade de Economia da Universidade do Porto (FEP), accomplishing this degree in 2014 with a final grade of 14 out of 20. During these years of college, was one of the founders of StartUp BUZZ, a non-profit startup accelerator, being the head of the organisation during his last year of college.

In 2015, Luís was admitted in the Master in Data Analytics at FEP, with the main goal of completing his academic curriculum, deepening his knowledge of analytical techniques and tools, to use them as a complementary tool to his professional development.

After completing his Bachelor in 2014, Luís started this professional career in Sonae Sports Fashion as an International Development Project Manager, managing the setup process of new retail operations abroad. Since 2017, Luís works in the Marketing Department of NOS as a Product Manager, managing a portfolio of mobile services.

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Dalila Fontes, for all her availability and preciously supervision, but mainly for presenting me the literature that inspired me to develop this theme in my Dissertation.

To all my Master's colleagues, for all the experiences I have shared with them on all classes and group assignments during these two years.

To my closest friends, for their support and patience. I would like to specially thank to my best friend and girlfriend, Luísa, for always being on my side.

Last, but not the least, to my parents, Vitor and Cristina, my aunt Aurora and my brother Ricardo for all the missed dinners, but mainly for all the attention and patience they have been giving me since 1993.

Resumo

Na última década, a Bolsa de Valores Portuguesa tem-se deparado com alguns acontecimentos que afetaram várias empresas e setores, levando a mudanças dramáticas no seu valor e composição. Estes eventos tiveram impactos na estabilidade global do mercado, resultando numa desvalorização considerável da Bolsa de Valores Portuguesa num período de dez anos.

É do conhecimento geral que os mercados bolsistas geram grandes quantidades de dados, sendo bastante dispendioso analisá-los pelos meios tradicionais. No entanto, esta informação pode ser usada para construir grafos, dos quais, ao analisar as suas características, comportamentos de correlação de preços podem ser estudados e deduzidos.

O objetivo desta dissertação é estudar o comportamento da Bolsa de Valores Portuguesa entre 2000 e 2015, recorrendo ao Modelo Market Graph. Este modelo liga pares de ações com base nas correlações das variações de preço das mesmas, com o objetivo de representar e estudar padrões de correlação de preços ao longo dos anos. Ao usar este modelo, também é possível identificar conjuntos de ações e avaliar a exposição do mercado a acontecimentos internos e externos.

Palavras-Chave: Market graph, Cliques, Quasi-cliques, K-cores, Topological stability

Abstract

Over the past decade, the Portuguese Stock Market has gone through some events that have affected several companies and industries, leading to dramatic changes in its value and composition. These events impacted on the overall stability of the market, resulting on a considerable devaluation of the Portuguese Stock Market over the period of ten years.

It is commonly known that stock markets generate large amounts of data, being rather resource-consuming to analyse these data by the traditional means. Nevertheless, this information can be used to construct graphs from which, by analysing their characteristics, price correlation behaviours can be studied and inferred.

The purpose of this dissertation is to study the Portuguese Stock Market between 2000 and 2015, using the Market Graph Model. This model connects pairs of stocks based on the correlations of their price variations, aiming to represent price correlations patterns over the years. By using this model, it is also possible to identify clusters of stocks and assess market exposure to external and internal events.

Keywords: Market graph, Cliques, Quasi-cliques, K-cores, Topological stability

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Over the past decades, the Portuguese Stock Market has gone through some events that have affected several companies and industries, leading to dramatic changes in its value and composition. These events impacted on the overall stability of the market. In Portugal, there are two national indexes PSI 20 and PSI Geral. PSI Geral is constituted by all listed companies in Portugal, while PSI 20 is the reference index of the portuguese stock market, composed of the stocks of the 20 largest companies in PSI Geral. The aforementioned events include the bankruptcies of some Portuguese financial institutions such as Banco Privado Português (*BPP*), Banco Português de Negócios (*BPN*), Banco Internacional do Funchal (*BANIF*), and Banco Espírito Santo (*BES*), which have had and still have impact on the overall stability of the stock market. This can be seen from+° the drastic devaluation of the PSI 20 index, reported as 44,7 billion euro between December 31st 2007 and April 11th 2016 (Relvas, 2016).

The main objective of this dissertation is to study the degree of association between the price variations of PSI Geral stocks. At the same time, this project aims to search for groups of stocks with similar behaviours throughout time, regarding price fluctuation. Finally, we assess the degree of exposure of the market to external and internal threats.

1.2 Problem Definition

Stock markets generate large amounts of data due to the high volume of daily transactions there occurring. Thus, it is rather resource-consuming to analyze it through traditional means. However, this information can be used to construct graphs from which, by analysing their characteristics, the market behaviour may be studied and inferred.

The purpose of this dissertation is to analyse market data using the graph model, based on the work of Boginski et al. (2003, 2005); Shirokikh et al. (2013), whose theoretical grounds are explained in Chapter 2. A graph consists of a set of nodes which are connected by a set of edges. Edges may be associated with a number known as edge weight. Following on the above mentioned works, each stock is represented as a node and each pair of nodes may be connected by a weighted edge, depending on the degree of correlation between the corresponding stock-price variation.

There are three main streams of analysis. Firstly, a general graph analysis, where the overall level of connectivity of the market and the distribution of connection per stock is assessed by examining, respectively, edge density and degree distribution of the graph. This study provides an overview on how price fluctuations of each stock influence the behaviours of price fluctuations of the other stocks, as well as a perception on whether their fluctuations are disseminated over the whole market or concentrated on a small set of companies.

Secondly, we study the existence and the dimension of groups of interconnected stocks. The main objective is to identify groups of stocks with similar behaviours in what regards their price evolutions and assess if, throughout the years, these groups tend to grow bigger or smaller, *i.e.*, understand the periods with more accentuated price trends. These groups will be identified by resourcing to graph theorical concepts such as cliques and clusters. These concepts are addressed in Chapter 3.

Finally, the market topological stability is analysed. Specifically, we test the market exposure to external threats, by analysing the overall stability, and the vulnerability of market specific stocks. Conceptual details are presented in Chapter 4. This will be accomplished by understanding how prone the market is to outside events that may affect listed companies.

1.3 Thesis Overview

On the following chapters we will provide the necessary Literature Review on Graph Theory Definitions on Chapter 2, Cliques and Graph Clustering on Chapter 3 and Graph Topological Stability on Chapter 4. Afterwards, we present the Problem Formulation and the Solution Approach for this dissertation on Chapter 5 and the Empirical Study and Analysis on Chapter 6. Finally, we present the main conclusions of this dissertation and possible future works that can be done from this dissertation on Chapter 7.

Chapter 2

Graph Theory Definitions

This chapter presents some introductory definitions and concepts of Graph Theory, which will help in understanding this dissertation. Following that exposition, we will introduce the Market Graph and discuss some of its prespectives.

2.1 Introductory Graph-theoretic Definitions

We can define the *graph model* as follows: Let $G = (V, E)$ be a *simple, undirected graph* where V is a set of nodes and E a set of undirected edges, each connecting a pair of nodes. If two nodes u and v are connected by an edge, such that $(u, v) \in E$, then they are called adjacent or neighbours and the edge is called incident to u and v (Boginski et al., 2003).

A path of length r between nodes u and v in G is a subgraph of G defined by an alternating sequence of distinct nodes and edges $u \equiv v_0, e_0, v_1, e_1, ..., v_{r1}, e_{r1}, v_r \equiv v$ such that $e_i = (v_i, v_{i+1}) \in E$ for all $1 \leq i \leq r$. Two nodes u to v are connected if there is at least one path between them.

The graph G is *connected* if there is a path from any node to any other node in the set of nodes V. If the graph is disconnected, it can be decomposed into two or more connected subgraphs, known as *connected components* of G.

Let $S \subseteq V$ be a subset of nodes of G. The induced subgraph $G[S]$ is the graph defined by the node set S and the edge set $E' \subseteq E$ that includes all edges in E that connect nodes in S

The *edge density* $\rho(G)$ is the ratio between the number of edges in G and the number of edges in a complete graph with the same number of nodes. An undirected graph with n nodes has at most $\frac{n \times (n-1)}{2}$ edges. If indeed the graph has $\frac{n \times (n-1)}{2}$ edges then the graph is complete, that is all node are pairwise adjacent.

The *Degree* of a node $v \in V$ denoted by $deg_G(v)$ is the number of edges incident to this node.

The *degree distribution* of a graph is a function (represented as $P(k)$) that shows

the probability that a randomly selected node is connected to k edges (Huang et al., 2009). This function can be represented as $P(k) \propto k^{-y}$, or, on a more convenient form: $logP(k) = - \, y \log k + const$, where k represents the degree and y is the slope of the linear function in log-log scale that characterizes the power-law distribution.

The *complement graph* of $G = (V, E)$ denoted by is $\bar{G} = (V, \bar{E})$, consists of the same set of nodes and of edges connecting them not present in E, that is, if an edge $(i, j) \in E$, then $(i, j) \notin \bar{E}$, and if $(i, j) \in \bar{E}$ then $(i, j) \notin E$ (Pardalos and Xue, 1994)

2.2 Market Graph

The Market Graph is a graph-based representation to study financial systems. In this model, each individual stock is represented as a node in a large-scale graph where every two nodes are connected by an edge if, in a specified period of time, the correlation between their price fluctuations exceeds a predefined threshold θ (Boginski et al., 2003).

In this graph each stock is represented as a node, while edges are assigned to each pair of nodes whenever the correlation of their price variations surpasses the threshold θ.

For example, suppose that, for a certain graph, $\theta = 30\%$. At the same time, assume two stocks, stock A and B represented by the nodes a and b. If the correlation of the price variations of both stocks is higher than 30%, let us say, 55% then there is an edge connecting a and b. If on the other hand, the correlation is 20%, then a and b will not be connected.

The main advantage of this technique is that it can display both general and specific characteristics of the market, being a good tool to summarize and extract information from financial data, which is rather complex and heavy to be studied by more traditional means. This way one can understand how correlated are stock prices by studying the proportion of existing edges (edge density) ans also how relevant a single or a set of stocks are by analysing the number of edges they aggregate.

In more detail, the aforementioned characteristics are translated in some graph-theoretic parameters, such as edge density and degree distribution, to analyze of the graph as whole, as well as clusters and cliques, to identify and study highly similar groups of stocks.

2.3 Power-Law Model

The first approaches to graph representations were developed by Erdös and Rényi (1959, 1960, 1961) based on the concept of *uniform random graphs*, being further developed some years later by Bollobás (1978) and Bollobás and Thomason (1985).

The main idea of the uniform random graphs was to randomly and independently assign edges to each pair of nodes, given a predefined probability p. What must be pointed as an initial point of discussion is that there is no rational for assigning an edge to a specific pair of nodes, rather than mere chance.

Nevertheless, despite capturing some properties of real graphs with the same number of nodes and edges, uniform random graphs showed to be significantly different, being unable to represent some relevant properties like clustering and degree distribution (Boginski et al., 2003). Although uniform random graphs showed to capture some properties of real graphs, in the end they failed to represent the clustering property of real graphs (Watts and Strogatz, 1998; Watts, 1999).

This property states that the probability of two nodes being connected (*clustering coefficient*) is higher if both are connected to a third node, while in uniform random graphs this probability is constant and equal to p, i.e., independent of all other nodes. Watts and Strogatz (1998) and Watts (1999) proved that the value of the clustering coefficient is higher than uniform probability value p for the same number of nodes and edges.

Another misrepresentation when using this type of graphs is the degree distribuition. By using uniform random graphs, edge density would follow a Poisson distribution, with a parameter equal to the average degree of a vertex ($n \times p$). However, the same authors have shown that in real graphs the degree distribution obey to a power law. This property has been observed in real graphs such as the Call Graph and the Web Graph. More recently, Boginski et al. (2003) proved the applicability of the power law model to the market graph.

Subsequently, the power-law random graph model was used to describe real-life graphs, being widely considered as a well-adjusted representation of many graphs. The basic idea of this model $P(\alpha, \beta)$ is as follows: if y is the number of nodes with degree x, then per this model $y = e^{\alpha}/x^{\beta}$, meaning that the number of nodes with degree x will vary as a power of the degree value x.

The relationship between the parameter (β) and the size of the largest connected component is very relevant for the study of graphs, as it has been theoretically shown that in a power-law graph, a giant connected component a.a.s. emerges at $\beta = 3.47875$, and a graph a.a.s. becomes connected when $\beta < 1$. (Aiello et al., 2001).

Empirical evidence have showed that when β is rather small, the graph contains many nodes with a high degree. This fact is important to find large cohesive groups of nodes.

Chapter 3

Cliques and Node Clustering

This chapter introduces some concepts and methodologies regarding cliques and cluster discovery. More specifically, it provides definition of Cliques and Independent Sets, and also the financial interpretation of these concepts. Afterwards, we will introduce alternatives to cliques, namely, quasi-cliques and k-core decomposition. These two concepts will also be used for clustering purposes, however they are less restrictive than cliques.

3.1 Cliques and Independent Sets

A *clique* can be defined as a group of nodes that are completely interconnected. Given a subset $S \subseteq V$, we denote by $G(S)$ a subgraph induced by S. A clique C is a subset $C \subseteq V$ such that $G(C)$ is a complete graph (Bomze et al., 1999). A closely related concept is that of an independent set as it consists of a set of nodes in a graph such that no two nodes are adjacent, *i.e.*, is a set of nodes without connections. We can also define this as being a subset $I \subset V$ such that the induced graph $G(I)$ has no edges.

Within this thematic, we have the *maximum clique problem*, which seeks for the largest clique in a graph, *i.e.*, the clique of maximum cardinality. The clique number of G, denoted by $\omega(G)$ is the number of nodes of the maximum clique. Analogously, we can define the *maximum independent set problem* as the independent set of maximum cardinality, *i.e.*, the largest set of nodes without connections. The size of an independent set is known to be the *stability number* of a graph G - $\alpha(G)$. These two are complementary since a clique in G is an independent set in $\bar{G}$. Therefore, finding the maximum clique in G corresponds to finding the maximum independent set in $\bar{G}$.

However, for specific graphs the complementary solving of these two problems may be significantly different: for example, for sparse graphs, the maximum clique as a bound that can be found in polynomial time (Chiba and Nishizeki, 1985), but

finding a maximum independent set is NP-hard. Since the market graph is a sparse graph and thus finding a maximum clique is computationally easier than finding a maximum independent set.

The maximum clique problem has been proven to be NP-hard (Gary and Johnson, 1979). Nevertheless, some authors have developed methods to address this problem. A comprehensive survey on algorithms for the maximum clique problem is provided by Pardalos and Xue (1994).

For this dissertation, In this dissertation and following on Shirokikh et al. (2013), we use a greedy constructive heuristic to find a clique. The basic idea of this heuristic is to add one node at the time to the incumbent clique. The node chosen to add is the one having the largest number of adjacent nodes not yet included in the clique. It, obviously, has to connect to all nodes already in the clique. Otherwise, by adding it the clique property would be lost. This is repeated until no nodes out of the clique connect to all nodes in the clique.

The size of the clique found using this heuristic (n) is a lower bound on the true size of the maximum clique. Therefore, problem dimmensionality can be reduced by removimng from the original graph all nodes with a degree less than n and thus improving the efficiency of the process of searching for a maximum clique.

Maximum cliques will be studied Chapter 6 with the main goal of finding and analysing cohesive clusters in Portuguese Stock Market (from 2000 to 2015).

3.1.1 Financial Interpretation of Cliques

A clique in the market graph with a positive threshold value θdefines the set of stocks whose price fluctuations exhibit a similar behaviour, *i.e.*, a change of the price of any instrument in such a clique is likely cause a price fluctuation in the same direction on all other instruments in the clique.

Thus, finding cliques in the market graph is a very useful technique of classifying information of financial instruments and of supporting investment decisions either to build a diversified portfolio or to find similar instruments (Huang et al., 2009; Boginski et al., 2005).

3.2 Graph Clustering

Strongly related with the concept of clique, we have the notion of *cluster*, *i.e.*, a group of highly similar elements. The clustering property can be found on real graphs and states that the probability of the event that two given nodes are connected by an edge is higher if these nodes have a common neighbour (Watts and Strogatz, 1998; Watts, 1999).

The process of discovering clusters can be naturally done by finding connected components: a disconnected graph has a unique decomposition into maximal con-

nected subgraphs. Although decomposing a graph into connected components provides one of the most intuitive graph-based clustering techniques, it may retrieve very few information, as clusters may be extremely large, almost reaching the length of the whole graph (Shirokikh et al., 2013).

Cliques, by definition, can be too restrictive for clustering purposes as it is required that all nodes in a clique are pairwise adjacent, and so, some relaxations have been proposed. Among these, we focus on two different relaxations, namely Quasi-cliques and K-Core Decomposition. The first is centered on the behaviour of the cluster as a whole, while the latter focus on the behaviour of each individual node.

3.2.1 Quasi-Cliques

As stated before, a clique is the strictest concept of a cluster, where all nodes are pairwise connected and the absence of just one link violates its structure and, so, may produce inconsistent clustering results.

Therefore, other structures, which bring flexibility to the concept of clique have been proposed. These structures some how "relax" the requirements of a clique. In the context of the market graph, it is reasonable to relax the requirement of the edge density of a clique, performing a so-called density-based clique relaxation (Shirokikh et al., 2013).

One of the most used relaxations is referred to as the a γ-quasi-clique (Cγ), being formally defined as a subset of V such that the induced subgraph $G(C_\gamma)$ is connected and has at least $\gamma\frac{q(q-1)}{2} with \gamma \in [0,1]$ edges (Pardalos and Rebennack, 2010), where q is the number of nodes in the subset, *i.e.*, $|C_\gamma| = q$. For the lower bound value of $\gamma = 0$, $G(C_\gamma)$ may have no edges, while for the bound of $\gamma = 1$, C_γ is a clique in G.

Nevertheless, working with ratios may originate some misleading results, as it may lead to result in a group with highly cohesive regions involving a high volume of direct interactions, coupled with very sparse regions, relying mostly on indirect interactions with the rest of the group (Seidman, 1983). Additionally, the process of discovering quasi-cliques is similar to the one of finding the maximum clique, meaning that it is still computationally challenging, The maximum quasi clique problem is also a NP-hard problem (Bomze et al., 1999; Pattillo et al., 2013).

Quasi-cliques are a more reasonable clustering method as they are more inclusive than the clique, without violating the basic idea of a cluster, since it ensures a certain minimum ratio γ between the number of existing edges and the maximum possible number of edges within the group (Namaki et al., 2011) - a group of 10 nodes where only two or three edges are missing in a total of 45 ($\frac{10\times9}{2}$) is stil a group of highly related elements.

3.2.2 K-Core Decomposition

Seidman (1983) proposed the *K-Core Decomposition* that can be defined as the simplest clique relaxation technique. In comparison to the quasi-clique concept just introduced, the K-Core Decomposition allows for finding clusters with a more uniform intra cluster similarly, since a minimum level of direct connections in imposed (Shirokikh et al., 2013; Shahinpour and Butenko, 2013). A k-core is a subset of nodes that induce a subgraph with minimum node degree of at least k, meaning that this subgraph is composed only by nodes that have at least k incident edges. The degeneracy of a graph G, denoted by $\delta(G)$ is the largest value k for which G has a nonempty k-core (Pattillo et al., 2013).

Similarly, to the procedure of finding a maximum clique, a greedy algorithm can be applied to recursively remove all nodes with degree less than k from the graph until each of the nodes in the core have a sufficient number of incident edges.

While a k-core guarantees a certain minimum number of neighbours of each node in the group, the number of non-neighbours within the group may still be much higher than k, indicating a low level of familiarity in the group (Pattillo et al., 2013). However, this notion is easier to compute than that of the quasi-clique.

Chapter 4

Graph Topological Stability

Graph topology can be defined as the arrangement of the various nodes of a graph (Groth and Skandier, 2005). The topological stability of a graph against random failures and attacks is considered as a very important research aspect of complex graphs. Understand the correlation patterns among stocks through its study can be a good guide for risk management of stock investment (Huang et al., 2009).

In stock markets, listed companies can be exposed to several types of risks, such as bankruptcy or delisting, *i.e.*, market exit. In both cases, these events are represented in the market graph model by the removal of nodes and/or edges, which will forcedly lead to topological changes o in the graph. Topological stability of stock market graphs mainly measures the effect of node attacks and edge attacks. If the graph properties do not change drastically, then the graph is considered robust against attacks.

Taking into consideration the characteristics of a stock market graph, one can apply the node attack method, measuring the risk related to delisting or bankruptcy (Huang et al., 2009). Within the market graph model, it is applied by removing some nodes and all their connecting edges.

Node removal can be done in both stochastic and selective ways: while stochastic removal corresponds to randomly cut out nodes from the graph, selective removal means removing the nodes with a predetermined purpose, such as opting by the highest degree nodes.

The *Maximum connected component size* reflects the connectivity of a stock correlation graph. Graph stability can be evaluated by comparing the relative changes of the maximum connected component size before and after node removal. Consider the maximum component size of a graph ($|CO_{max}|$). The node removal, either stochastic or selective, is represented as f_N, where N is the number of nodes, f is the proportion of nodes removed. $R_{|CO|}$ is the ratio between the maximum component size of the new graph and the original maximum component size ($0 < R_{|CO|} \leq 1$).

The option between stochastic and selective removal, alongside the value for the removal proportion f, can lead to different results regarding graph stability. For

that reason, the RR index is the average of the differences between the effects of stochastic and selective removal methods, and it is given by the expression:

$$RR = \frac{\sum_f (R_{|CO|(r,f)} - R_{|CO|(d,f)})}{m},$$

where $R_{|CO|}(r, f)$ and $R_{|CO|}(d, f)$ represent, respectively, the influencing effects of stochastic and selective removal for a certain value of f and m is the number of different removal proportions chosen.

Huang et al. (2009) conclude that connectivity of a stock market graph becomes more and more fragile due to increasing threshold in selective removal, being always robust against stochastic removal, regardless of the threshold. This can be explained by the fact that edges become less frequent and the degree distribution gets farther away from uniformity as θ increases. At the same time, high degree nodes represent a core spot on large components, and thus their removal implies the removal of a large number of edges, making the component to reduce abruptly.

On the one hand, the robustness against stochastic edge removal indicates that some stochastic events such as delisting or bankruptcy will not have an essential influence on the whole price fluctuation correlation of the graphs. On the other hand, the frangibility to selective removal demonstrates that high degree stocks play a very important role in the whole market correlation. These arguments suggest that only listed companies associated with high degree nodes should be followed closely as only they may impact significantly the graph topology. This findings are very important in areas such as portfolio investment and risk management, as allows for investors to concentrate their efforts focusing on the most relevant targets.

Chapter 5

Problem Formulation and Solution Approach

As already mentioned, the case study in this work refers to the Portuguese stock market and the data used is that of the PSI Geral index, with closing price data being collected from all the 49 listed companies, between the years 2000 and 2015.

5.1 Data Description and Preprocessing

The data were collected from Thomson Reuters database, which was used to retrieve historical prices of these companies for the period stated above. The choice of the period had the rational of representing the last 16 years since the turn of the century, which were marked by a set of relevant events for the Portuguese financial panorama, almost all of them regarding the bankruptcy of financial institutions.

Following on the work by Shirokikh et al. (2013) and to reveal dynamic trends of the market, sixteen different graphs were constructed, corresponding to consecutive non-overlapping yearly periods from January 2000 to December 2015. Note that the number of stocks may be different in each time period, since not all the stocks were traded during this sixteen-year period. The choice of these yearly periods allows to avoid the lack of statistics while keeping reasonable a length, allowing one to capture the changes in sufficient detail. Calendar dates corresponding to each time period are given in Table 5.1.

The study is conducted based on time series sets of logarithmic returns, meaning that the datasets were transformed into a daily returns data, as it possesses the *scalability property*, this way avoiding prices magnitude differences in stocks. At the same time, by using a logarithmic scale autocorrelation corrections are introduced since it reduces the influences of the previous instances on each instance of a time series, leading to statistical properties which are more attractive to analyse (Campbell et al., 1997).

Period	Start date	End date	Period	Start date	End date
1	03/01/2000	29/12/2000	9	01/01/2008	31/12/2008
2	01/01/2001	31/12/2001	10	01/01/2009	31/12/2009
3	01/01/2002	31/12/2002	11	01/01/2010	31/12/2010
4	01/01/2003	31/12/2003	12	03/01/2011	30/12/2011
5	01/01/2004	31/12/2004	13	02/01/2012	31/12/2012
6	03/01/2005	30/12/2005	14	01/01/2013	31/12/2013
7	02/01/2006	29/12/2006	15	01/01/2014	31/12/2014
8	02/01/2007	31/12/2007	16	01/01/2015	31/12/2015

Table 5.1: Dates corresponding to each time period.

Therefore, if we consider $P_i(t)$ and $P_i(t-1)$ as, respectively, the closing prices of stock i at day t and $t-1$, the logarithmic return time series for stock i (r_i) can be defined as:

$$r_i(t) = ln\frac{P_i(t))}{P_i(t-1)}, t = 2, 3, ..., N$$

where N is the number of days in the period.

In the literature, both the Pearson Linear correlation coefficient and the Spearman Rank Correlation coefficient have been used for the calculation of correlation between financial time series. This dissertation uses the latter one, based on the analysis conducted by Shirokikh et al. (2013), in which they conclude it to be more robust for calculating correlations between financial time series. The authors have shown that the advantage of the Spearman Coefficient since it is a distribution-free statistic and less sensitive to outliers on the series, when compared with the Pearson coefficient.

The authors conclusion is supported mainly by three facts. Firstly, the Spearman correlation coefficient is distribution-free, *i.e.*, it does not assume any probability distribution of the original data. Secondly, it is less sensitive to outliers, in comparison to the Pearson correlation coefficient. Lastly, it provides a higher value of association in the presence of non-linear relationships between the variables.

To show how the Spearman rank correlation is computed, let us consider two time-series X = $\{x_1, ..., x_n\}$ and Y = $\{y_1, ..., y_n\}$. Them obtain rank variables R^X and R^Y by sorting X and Y making R_i^X and R_i^Y equal to the order of the corresponding X_i and Y_i (i = 1, 2, 3, . . ., n). In case there are identical values in the data, the average of the ranks is computed. Finally, compute the Spearman rank correlation coefficient as:

$$\rho = \sum_i^n (R_i^X - \bar{R^X})(R_i^Y - \bar{R^Y}) \sum_i^n \sqrt{(R_i^X - \bar{R^X})^2 (R_i^Y - \bar{R^Y})^2},$$

where $\bar{R^X}$ and $\bar{R^Y}$ are the average values of the corresponding variables.

Financial time series may have, and usually do, some unusual or abnormal data points, whose existence affects the analysis and its outcome.Therefore, following on the work by Boginski et al. (2003), and using the procedure they propose, we detect and correct some of such points. The algorithm used, see Table 6.2., is based on the GARCH (Generalized Autoregressive Conditional Heteroskedasticity) model. For further details, the interested reader is referred to Shirokikh et al. (2013).

Table 5.2: Detection and correction of outliers taken from Shirokikh et al. (2013)

Step 1: Fit the GARCH(1,1) model into the data and obtain the time series of residuals ϵ_t. Assume that $\epsilon_t \sim N(0,1)$
Step 2: Let N be a length of the time series. Generate N standard normal random variables. Choose a threshold k_α^n such that the probability of the maximum among these N random variables to be greater than k_α^n is α
Step 3: Find the maximum in the absolute values of all residuals and if the value is above k_α^n then consider the corresponding observation in the original time series as outlier;
Step 4: Set the corresponding value of the original time series to be equal to μ (assuming a zero value of the residual, which is the most probable for $N(0,1)$ and corresponds to the value of μ in the original time series);
Step 5: Repeat Step 1 for corrected time series and if no further outlier is found, terminate; otherwise, proceed to **Step 2.**

It should be noticed that two types of outliers, namely additive and innovational outliers, can be found in stock return time series (Grané and Veiga, 2010). Pena (2001) has shown that innovational outliers may impact the graph dynamics as they may affect many observations. In contrast, additive ones only affect a single observation and therefore should be removed. Therefore, following on the work by Shirokikh et al. (2013) and using the procedure they propose, we detect and correct such points.

The algorithm detects additive outliers by using GARCH (1,1) model with

$$Y_t = \mu + \varepsilon_t, \varepsilon_t = \sigma_t \epsilon_t,$$

where Y_t is the time series and ϵ_t are randomly distributed errors as in Shirokikh et al. (2013).

The procedure was used for four different levels of significance α: 0.5%, 1.25%, 2.5% and 5%, the results obtained are reported in Table 5.3. As expected, the elimination of the additive outliers induced a generalized data smoothing of each time series, which in turn led to significant reductions of the level of correlation.

Furthermore, larger values of α led to larger degrees of smoothing and thus, implying a larger decrease in the correlation level. More specifically, on average, more than 50% of the pairs of stocks have had correlation decreases after applying

the outlier correction procedure (see Table 5.3.), regardless of the year. The averages of the absolute correlation values for each year also decreased for almost all the values of α, with the exception of one single case: 2014 with $\alpha = 5\%$, where the average correlation value increased marginally: 0.1 p.p. As it can be seen from Table 5.4, $\alpha = 0.5\%$ led to a decrease of 1.9 p.p., $\alpha = 1.0\%$ to 2.0 p.p., $\alpha = 2.5\%$ to 1.8 p.p, and $\alpha = 5.0\%$ to 2.0 p.p.

From the analysis done one can conclude that the outlier correction has had the opposite effect on correlations of that of Shirokikh et al. (2013). Thus, rather than having larger absolute values of correlation we obtain smaller ones, *i.e.*, closer to zero. As a result, we have deciced to use the original dataset without any correction of outiler instances.

5.2 Graph Construction

In the market graph model, two stocks are connected by an edge if their correlation exceeds a certain threshold. In the literature, there are some different approaches to the method of choosing the threshold value θ. On the one hand, we have the definition of an absolute threshold value θ ($|\theta| \in [0, 1[$), used by Boginski et al. (2003, 2005). The idea is to consider pairs of stocks with rather coordinated behaviours, thus detecting sets of correlated and inversely correlated stocks, as opposed to sets of stocks with low or no correlation, *i.e.*, with correlation coefficient values near 0. On the other hand, we have the definition of a single positive threshold ($\theta \in [0; 1]$) (Shirokikh et al., 2013; Huang et al., 2009), which only analyses the time series of instruments which show similar behaviour over time, not considering the pairs of stocks with negative correlations.

In this work, two different analyses are performed, one for positive correlations ($\theta \in [0; 1]$) and another for negative correlations ($\theta \in [-1; 0]$). The goal of these analysis is to separately study sets of correlated and inversely correlated stocks in the market, studying the complete scope of stocks with coordinated behaviours, while being able to differentiate these two behaviour.

To perform these analyses, the threshold values θ to be considered in the graphs are previously defined: θ_{pos} is the threshold for the positive graph ($\theta_{pos} \in [0; 1]$) and θ_{neg} for the negative graph ($\theta_{neg} \in [-1; 0]$). In fact, the threshold value θ controls the edge density of a graph (Shirokikh et al., 2013), and so, both thresholds will be defined based on the frequency of correlation coefficients throughout time (represented in Figure 5.1), choosing the values that allow an adjusted and consistent number of edges for every period.

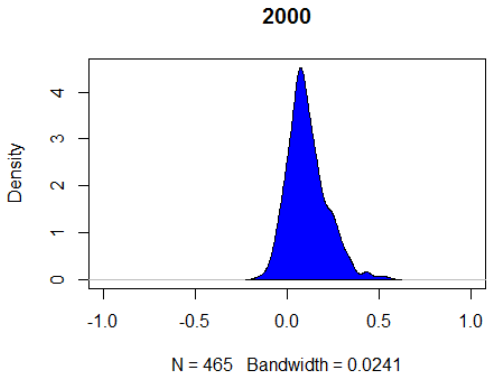
5.3 Solution Approach

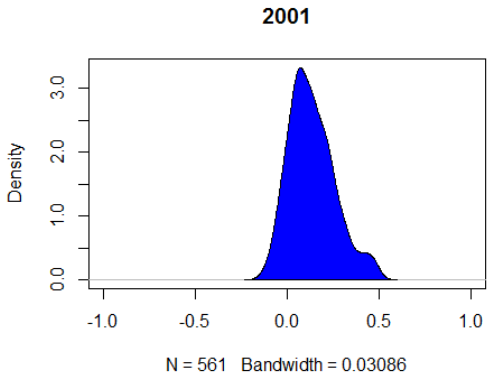
The sixteen periods were grouped into four different groups: 2000 and 2001, 2002 to 2006, 2007 to 2011, and 2012 to 2015. In the earlier periods, most stocks were majorly uncorrelated, with the distributions all stacked around the value zero. This can be seen in Figure 5.1, as the correlation distribuition for these two years is narrow and tall around 0 and quickly flattens going left and right. Nevertheless, this phenomenon aggravated itself after 2002 and until 2006. Note that, for these five graphs, it is hard to find instances in $[-1; -0.3]$ and $[0.3; 1]$. After 2007, the picture changed considerably, with the number of stocks with considerable positive correlation steadily increasing until 2011. For these years, graphs became more flat, and wider to the right. Afterwards, and until the last period (2015), strong positive correlations still were observed, though with a smaller magnitude, when compared with the period 2007-2011, with graphs become more narrow than the previous set.

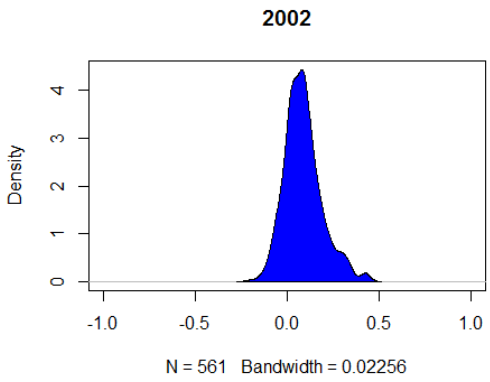
Given the above mentioned analysis, we will apply for the positive graphs, the threshold values presented in Table 6.1. The main goal of adjusting the values of θ_{pos} was to be able to have edge density values rounding 20%, following the example of Shirokikh et al. (2013). In Shirokikh et al. (2013), the edge density of the study was around 1%. However, since our case study involves much smaller graphs, with no more than 46 nodes in each period, a larger edge density is required. Note that their graphs included over 3400 nodes (stocks).

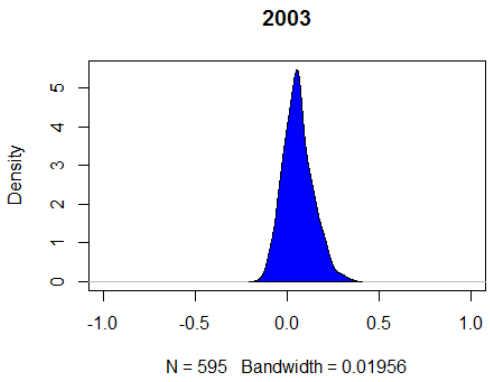
Due to the scarcity of negative correlations in this market, the threshold θ_{neg} was set to -5% to construct the negative graph. The application of this threshold led to smaller edge densities, with a maximum observed value of 12.4% (2014) and a minimum of 3.4% (2010).

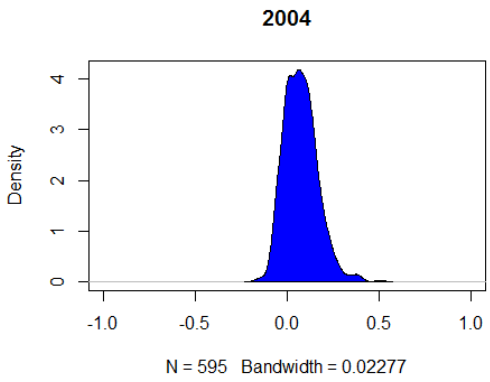
The existence of so many negative correlation values close to zero leads to believe that the analysis of the negative graph will not lead to robust conclusions, as it is hard to consider that edges represent strong correlation for such small threshold values.

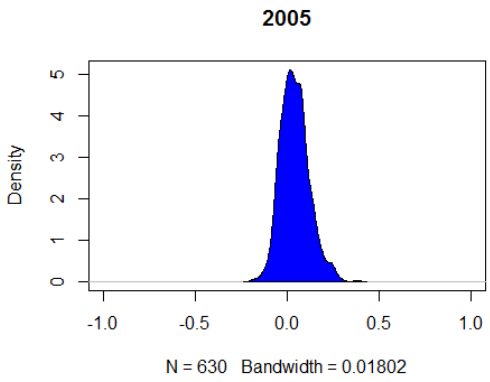

2000
Density
0
1
2
3
4
-1.0
-0.5
0.0
0.5
1.0
N = 465 Bandwidth = 0.0241

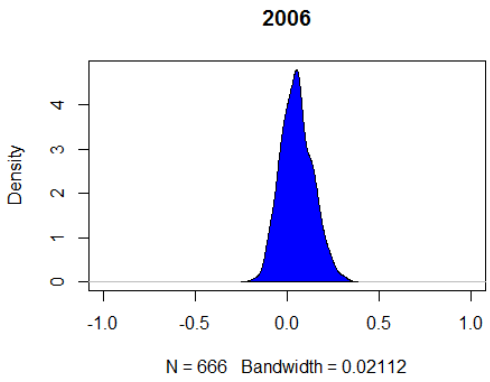

2001
Density
0.0
1.0
2.0
3.0
-1.0
-0.5
0.0
0.5
1.0
N = 561 Bandwidth = 0.03086

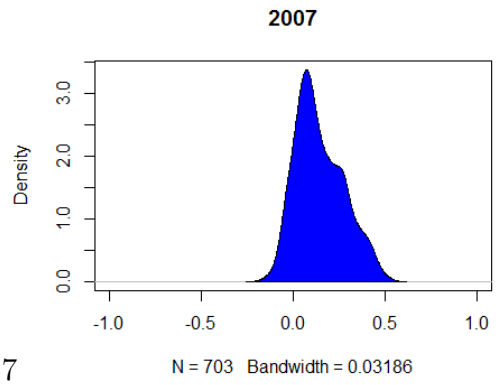

2002
Density
0
1
2
3
4
-1.0
-0.5
0.0
0.5
1.0
N = 561 Bandwidth = 0.02256


2003
Density
0
1
2
3
4
5
-1.0
-0.5
0.0
0.5
1.0
N = 595 Bandwidth = 0.01956


2004
Density
0
1
2
3
4
-1.0
-0.5
0.0
0.5
1.0
N = 595 Bandwidth = 0.02277


2005
Density
0
1
2
3
4
5
-1.0
-0.5
0.0
0.5
1.0
N = 630 Bandwidth = 0.01802


2006
Density
0
1
2
3
4
-1.0
-0.5
0.0
0.5
1.0
N = 666 Bandwidth = 0.02112


2007
Density
0.0
1.0
2.0
3.0
-1.0
-0.5
0.0
0.5
1.0
N = 703 Bandwidth = 0.03186

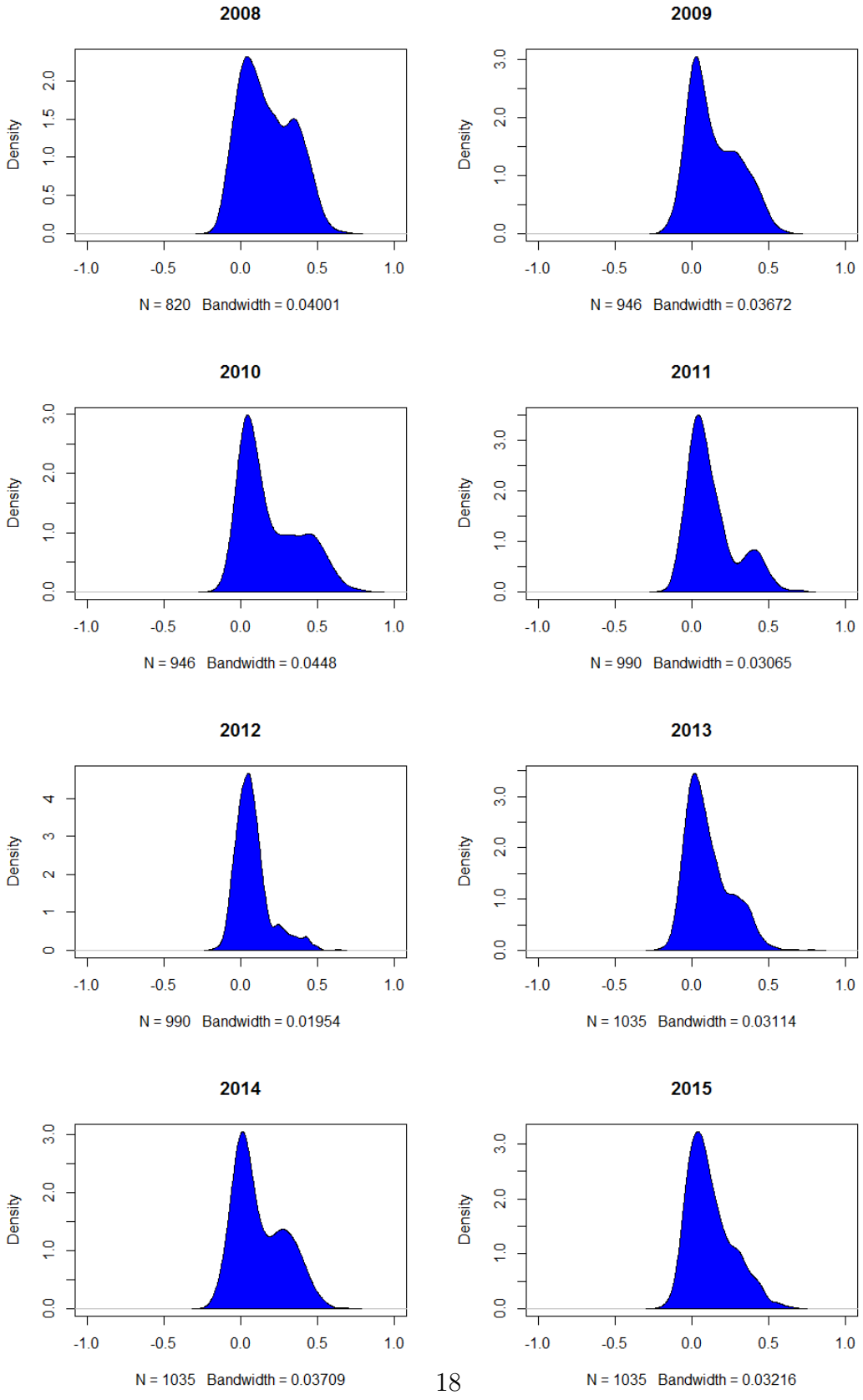


Figure 5.1: Distribution of correlation coefficients for years 2000 to 2015

Year	**Pairs of nodes**	**Deterioration in correlation using Algorithm from Table 5.2 with $\alpha =$**			
		0.5%	**1.25%**	**2.5%**	**5%**
2000	465	73.8%	73.8%	20.9%	72%
2001	561	73.3%	76.5%	73.6%	74.2%
2002	561	64.5%	64.9%	68.1%	64.9%
2003	595	61.7%	60.3%	59.5%	57%
2004	595	65.4%	63.9%	61.8%	61.7%
2005	630	55.9%	58.4%	59%	55.7%
2006	666	62.6%	62.6%	63.5%	62.6%
2007	703	76.8%	77.2%	76.1%	77.7%
2008	820	77.6%	77.6%	73.8%	77.2%
2009	946	46.5%	46.5%	44.6%	44.3%
2010	946	49.5%	74.2%	47%	75.6%
2011	990	42%	43.6%	43%	20.1%
2012	990	40.5%	49%	59.2%	61.8%
2013	1035	45.7%	43.6%	44.3%	46.3%
2014	1035	6.1%	6.2%	5.8%	7.3%
2015	1035	9.2%	10.1%	10.3%	10.3%

Table 5.3: Summary of the implications caused by data pre-processing algorithm (Table 5.2)

Year	**Original mean absolute correlation values**	**Mean absolute correlation values for $\alpha =$**			
		0.5%	**1.25%**	**2.5%**	**5%**
2000	**11.9%**	8.6%	8.6%	10.6%	9.4%
2001	**14.4%**	10.9%	10.6%	10.9%	10.9%
2002	**10.2%**	8.2%	8%	8.1%	8.5%
2003	**7.8%**	6.8%	6.6%	6.5%	6.8%
2004	**8.9%**	6.9%	7.1%	7.5%	7.5%
2005	**6.7%**	5.9%	5.8%	5.7%	5.8%
2006	**7.9%**	5.9%	6.3%	6.1%	6.1%
2007	**15.3%**	10.8%	11.6%	11%	10.6%
2008	**18.9%**	14.3%	14.4%	14.4%	14.5%
2009	**16.1%**	15.2%	14.9%	14.9%	14.9%
2010	**19.8%**	17.9%	15.7%	17.5%	15.1%
2011	**14.4%**	13%	12.8%	13.2%	14.1%
201	**9.6%**	8.8%	8.7%	8.3%	8.5%
2013	**12.9%**	11.4%	11.7%	11.4%	11.2%
2014	**15.1%**	15.1%	15%	15.1%	15.2%
2015	**13.8%**	13.6%	13.5%	13.3%	13.3%

Table 5.4: Comparison of the mean absolute correlation values for each year for different levels of significance, after applying the algorithm from Table 5.2

Chapter 6

Empirical Study and Analysis

6.1 General Analysis of the Market Graph

The study of some graphs is helpful to reflect the global patterns of a graph. Characteristics such as edge density, node degree, and degree distribution provide a good overview of these patterns, and so, of its underlying data set (Shirokikh et al., 2013).

In fact, by knowing the edge density of a graph, *i.e.*, the percentage of existing edges of the total possible ones, one can intuitively understand how connected is the graph, controlling the degree of connectivity, so that only the most significant correlations are studied. As stated in Chapter 5, edge density is kept close to 20% on the positive graphs, and to -5% on the negative graphs to analyse the most meaningful correlations, following the work of Shirokikh et al. (2013).

The degree distribution analysis allows to understand the overall connectivity of the graph. More specifically, the value of the β coefficient, corresponding to the power-law regression, gives relevant information regarding the number of connected components in the graph, and of the existence of cliques and clusters, as previously addressed in Section 2.2.

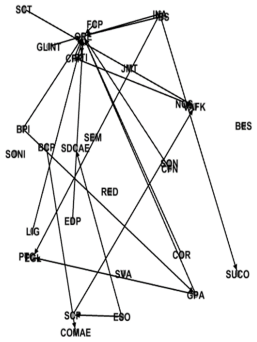
It is important to bear in mind that the composition of the stock market and edge structure of the graph change over the years. For this reason, we will study how many nodes appear, disappear and are kept from one year to the next. This analysis allows to assess graph stability and to what extent meaningful correlations between stock prices remain over the years.

The nodes with the highest degrees correspond to the stocks with the largest number of meaningful correlations with other stocks. Those correspond to companies having stock price evolution similar to a large number of other companies stock price evolution. Therefore, they can be used to explain the market behaviour. Furthermore, among these stocks one can find market makers.

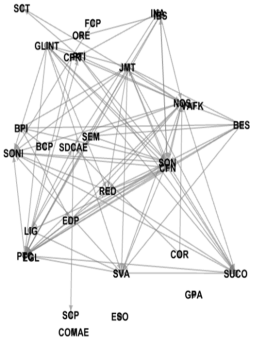
Figure 6.1 shows the market graphs of the PSI Geral index. For all the years in the time horizon under study (2000-2015). These are the graphs to be used in the

following subsections.

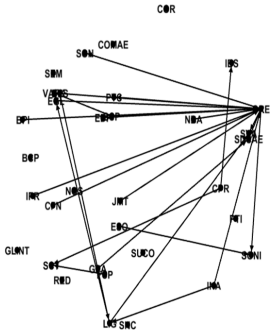
Negative Graph - 2000



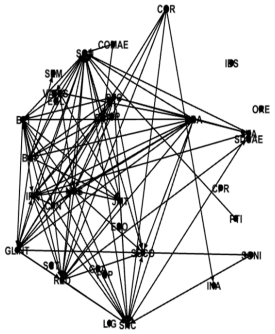
Positive Graph - 2000



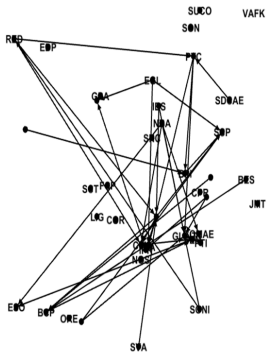
Negative Graph - 2001



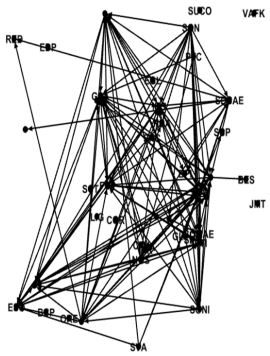
Positive Graph - 2001



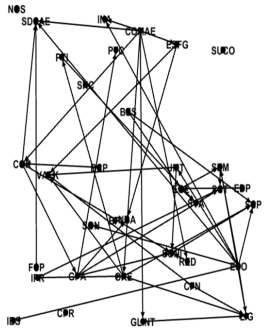
Negative Graph - 2002



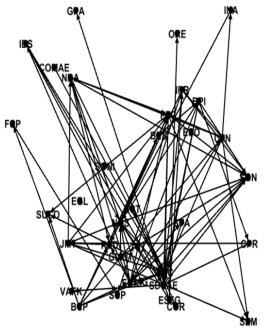
Positive Graph - 2002



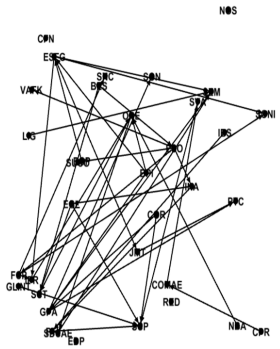
Negative Graph - 2003



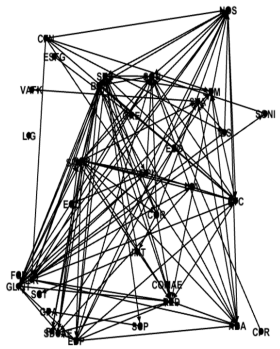
Positive Graph - 2003



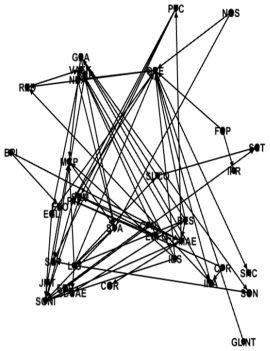
Negative Graph - 2004



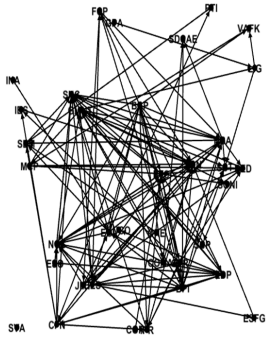
Positive Graph - 2004



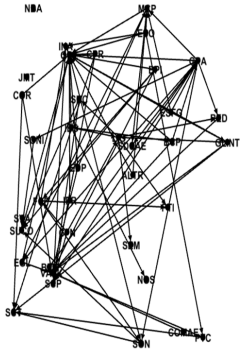
Negative Graph - 2005



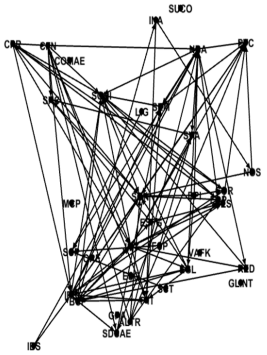
Positive Graph - 2005



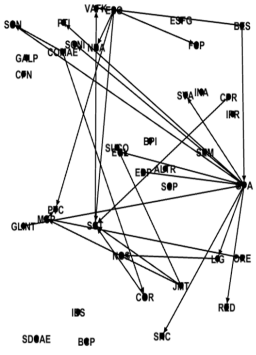
Negative Graph - 2006



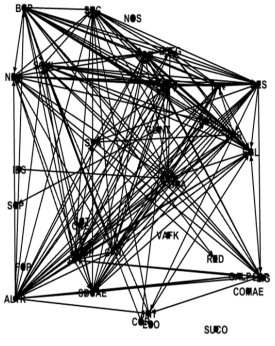
Positive Graph - 2006



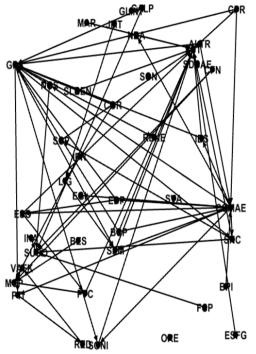
Negative Graph - 2007



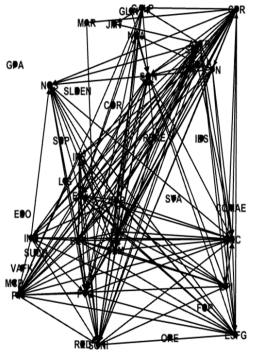
Positive Graph - 2007



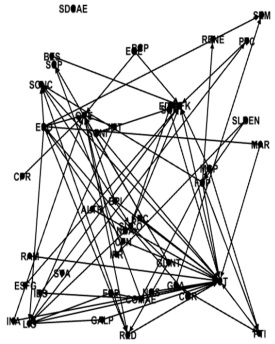
Negative Graph - 2008



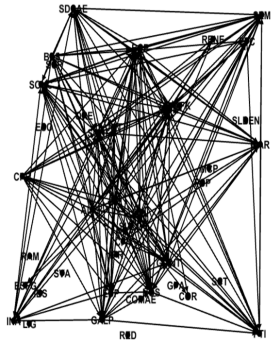
Positive Graph - 2008



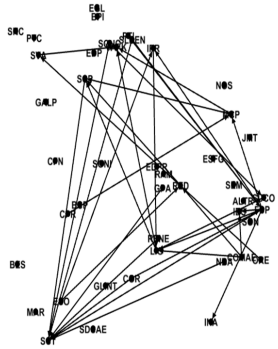
Negative Graph – 2009



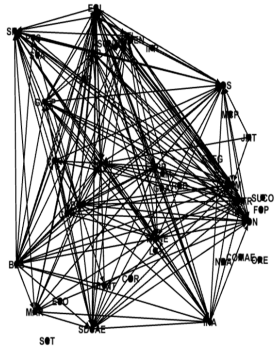
Positive Graph - 2009



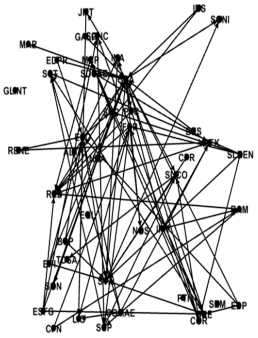
Negative Graph - 2010



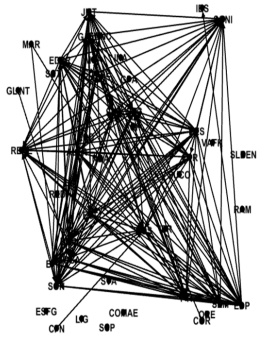
Positive Graph – 2010



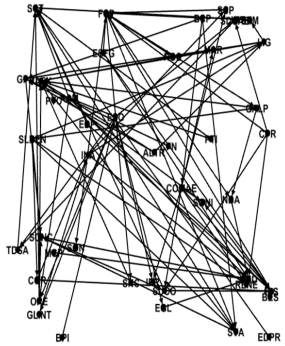
Negative Graph - 2011



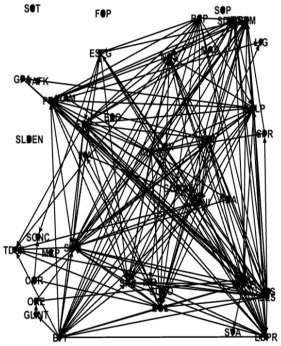
Positive Graph - 2011



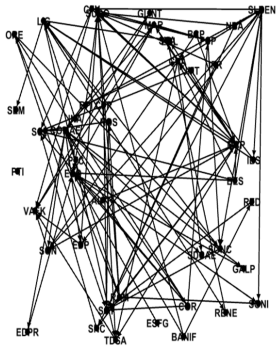
Negative Graph - 2012



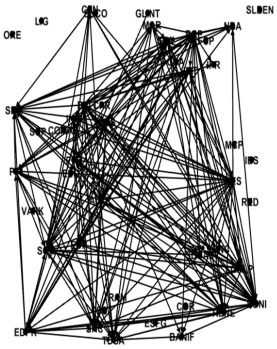
Positive Graph - 2012



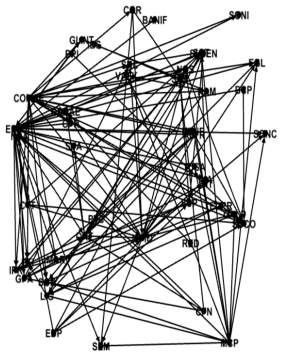
Negative Graph - 2013



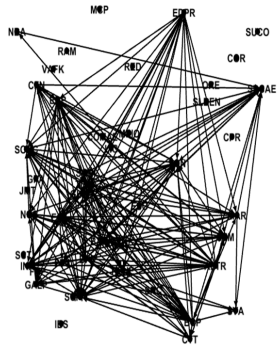
Positive Graph - 2013



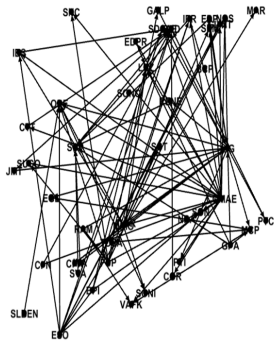
Negative Graph - 2014



Positive Graph - 2014



Negative Graph - 2015



Positive Graph - 2015

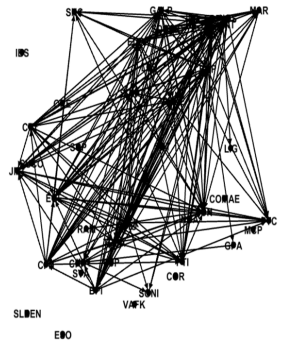


Figure 6.1: Market graph representation of negative and positive graphs between 2000 and 2015.

Period	No. of nodes	Positive Graph					
		Threshold	No. of edges	Edge density (%)	Max. node degree	Power-law parameters	
						β	R^2
2000	31	20%	84	18.1%	16	1.214	54.2%
2001	34	25%	89	15.9%	18	0.1817	4%
2002	34	15%	119	21.2%	18	0.8662	40.9%
2003	35	15%	80	13.4%	14	0.8347	58.5%
2004	35	15%	103	17.3%	15	0.7543	32.1%
2005	36	10%	122	19.4%	16	0.2356	7%
2006	37	15%	89	13.4%	13	0.3765	5.5%
2007	38	25%	164	23.3%	21	0.2391	1.7%
2008	41	35%	160	19.5%	21	0.2632	4.2%
2009	44	30%	186	19.7%	23	-0.06096	0.2%
2010	44	40%	181	19.1%	22	0.3076	3.4%
2011	45	25%	190	19.2%	21	0.2643	2.6%
2012	45	15%	172	17.4%	20	0.1656	3.9%
2013	46	25%	191	18.5%	25	0.2249	3.1%
2014	46	30%	189	18.3%	23	0.6287	12.1%
2015	46	25%	198	19.1%	24	0.1763	2.7%

Period	No. of nodes	Negative Graph					
		Threshold	No. of edges	Edge density (%)	Max. node degree	Power-law parameters	
						β	R^2
2000	31	-5%	22	4.7%	12	0.9679	97.5%
2001	34	-5%	24	4.3%	14	0.7208	78.2%
2002	34	-5%	38	6.8%	8	0.7878	79.4%
2003	35	-5%	47	7.9%	8	0.6302	64.1%
2004	35	-5%	40	6.7%	6	0.6869	71.2%
2005	36	-5%	64	10.2%	12	0.9197	69.5%
2006	37	-5%	74	11.1%	17	0.6924	60.4%
2007	38	-5%	28	4%	11	0.6932	87.1%
2008	41	-5%	61	7.4%	14	0.6113	44.8%
2009	44	-5%	66	7%	19	0.5843	52.7%
2010	44	-5%	32	3.4%	9	0.8306	89.6%
2011	45	-5%	80	8.1%	21	0.6651	60.4%
2012	45	-5%	87	8.8%	12	0.7855	48.4%
2013	46	-5%	103	10%	13	0.6321	41.1%
2014	46	-5%	128	12.4%	24	0.893	66.7%
2015	46	-5%	88	8.5%	20	0.8041	65.3%

Table 6.1: Global characteristics of the market graphs for years 2000 to 2015.

6.1.1 Positive Graph Analysis

We will start this analysis by interpreting the β coefficient of the power-law regression. As already explained in Section 2.3, this coefficient can be used to understand the size of the largest connected component. All graphs, except the one for the year 2000, have a β below 1, which implies the existence of a single connected component aggregating all nodes having at least one edge. Regarding the year 2000, the graph seems to have two separate components, as a result of β being above 1.

When the connected components of the Positive Graph are compared with such deductions, one can verify that these are almost always similar. More precisely, in a total of 16 instances, there are only three mismatches: in 2000, and in where there is only one connected component. 2004 and 2007, where, in both cases, there is a large component and a second one having only one edge connecting two nodes. Table 6.2 shows the number of connected components and respective size.

		Number of nodes	
Year		Positive Graph	Negative Graph
2000	1^{st} connected component	25	18
	2^{nd} connected component	—	2
	3^{rd} connected component	—	2
2001	1^{st} connected component	24	20
	2^{nd} connected component	—	4
2002		31	26
2003		31	30
2004	1^{st} connected component	30	31
	2^{nd} connected component	2	—
2005		35	36
2006		29	35
2007	1^{st} connected component	25	26
	2^{nd} connected component	2	—
2008		25	36
2009		26	40
2010		24	23
2011		27	40
2012		36	42
2013		27	45
2014		26	45
2015		28	43

Table 6.2: Connected components for the positive and negative graphs.

Some conclusions may be drawn about the graphs stability by analyzing Table 6.3, which shows the graph dimension regarding the number of edges, as well as the

various alterations in these number of nodes for two consecutive years. Until 2006 there were some remarkable changes to the edge structure of the graphs. In fact, during this 7 year period, there were four different moments where less than 50% of the edges transitioned from one year to the next (2000 to 2001, 2002 to 2003, 2004 to 2005 and 2005 to 2006). Actually, from 2005 to 2006 only 21% of the edges were kept.

This higher degree of alteration observed from 2000 to 2006 shows that in this period price correlations varied substantially. In fact, this variation brought about the need for changing the threshold (see Table 6.1).

	Positive Graph				Negative Graph			
	No. of Edges	Kept	New	Erased	No. of Edges	Kept	New	Erased
2000	84	—	—	—	22	—	—	—
2001	89	35	54	49	24	5	19	17
2002	119	62	57	27	38	6	57	27
2003	80	48	32	71	47	6	32	71
2004	103	42	61	38	40	5	61	38
2005	122	48	74	55	64	4	74	55
2006	89	26	63	96	74	10	63	96
2007	164	60	104	29	28	9	104	29
2008	160	112	48	52	61	8	48	52
2009	186	102	84	58	66	9	84	58
2010	181	135	46	51	32	10	46	51
2011	190	137	53	44	80	7	53	44
2012	172	137	35	53	87	13	35	53
2013	191	125	66	47	103	13	66	47
2014	189	102	87	89	128	23	87	89
2015	198	131	67	58	88	30	67	58

Table 6.3: Alterations to the graph edges from year N-1 to year N.

From 2006 onwards, graphs got more stable from year to year, with approximately 70% of the edges passing on from graph to graph. The only exception is the year 2014 where this ratio dropped down to 53%, a figure that still shows some stability. Table 6.4 shows the company that has the largest number of connections, i.e., largest degree node, for each year of the time horizon under consideration.

As it can be seen, the sector of the highest degree stocks changes over time. Most recently, since 2012, these stocks are exclusively of the Banking Sector. This makes sense as it was during this period that Portugal was under a Financial Assistance Program. As a consequence, during this time frame, the Portuguese Stock Market strongly devaluated as well as did the banks, which have been under the spotlight with constant news questioning their stability.

From 2007 until 2011, two stocks must be highlighted due to being frequently the ones with the most relevant correlations: ALTR and SONI, that represent groups of companies of the industrial sector, which was affected by the real estate crisis that happened during this period.

From 2000 to 2006, we must highlight the companies from the Telecom sector, as being the highest degree nodes in most of those years. In fact, if we extend the analysis to the second highest degree node, we can find companies of this sector in years 2000 to 2004. This is easily explained as during this period the Telecom sector was in expansion.

Finally, the degree of the highest degree node varies substantially from graph to graph, as shown in Table 6.4. Nevertheless, the last two years (2014 and 2015) must be highlighted due to the high number of edges registered on the highest degree stock: BCP. It was in 2014 that BES, another bank, was delisted, which had a large impact in the market, and so, it is normal that another bank, in this case BCP, was the company with most connections in that year.

6.1.2 Negative Graph Analysis

In the negative graph, all β values are below 1 (Table 6.1), indicating that the graphs will, most probably, be composed of one large connected component, aggregating all the nodes that have at least one edge. Recall that the β coefficient can be used to understand the size of the largest connected component. This can easily be seen to be the case for the Negative Graphs after 2001, as illustrated in Figure 6.1. This means that for these 14 years, there was only one connected component, leading again to the interpretation that the stocks with the stronger negative correlations are all somehow connected between themselves.

This observation can be also inferred from the following two characteristics of the graphs. First, the number of edges on these graphs is considerably small: 11 of the 16 graphs have a number of edges less than or equal to 80, which is the smallest edge dimension one can find in the positive graphs. At the same time, the threshold of -5% is not restrictive enough, which does not give confidence to the notion of edge, since weak and strong correlations are represented on the same way.

When analysing Table 6.3 we get the opposite interpretation to the one obtained for the Positive Graph. In this case, graphs are very unstable from year to year, with only 10% to 30% of the edges remaining from one year to the next. These percentages are widely explained by the fact that the number of edges is very small, which means that alterations on a small set of edges will have major impacts.

In the Negative Graphs and due to the more unstable structure, the highest degree node changes more frequently. Nevertheless, the highest degree stock is always in one of the following sectors: Investment Management, Sports, and Information Technology (see Table 6.5). Actually, the Investment Management sector appears in nine out of sixteen years as the most negatively correlated. This has to do with the

fact that companies such as ORE, SDC, and GPA manage less liquid assets, which are not directly traded in the financial markets such as Real Estate and Companies Strategic Assets. Regarding the other two sectors, Sports and Information Technology, the explanation for having the highest degree node in the negative graph is simpler, as these are industries that depend less on the financial market than banking, which leads to a larger probability of having different behaviours from most of the remaining companies in the market. Finding such stocks s a powerful resource mainly during period of crisis, as they are less impacted by the devaluation of market or even increase their value.

6.1.3 Final Remarks

The Portuguese stock market has very few listed companies, specially, when compared with the major stock markets used in the studies by Huang et al. (2009) (US Stock Market) and by Shirokikh et al. (2013) (Chinese Stock Market). At the same time, there is a high concentration of correlation values around the value 0 (Chapter 5), not being useful for this study.

As a result, low threshold values had to be used to construct the graphs; otherwise the resulting dimension would not deliver meaningful analyses thus no relevant conclusions would be drawn. This goal was achieved for the Positive Graph, where there was an interesting number of nodes to analyse, even if could not remain constant throughout the whole 16 years period. However, this was not the case for the negative graph, where the threshold was too low and graphs were considerably sparse.

Given the small sized negative graphs obtained, even for very low threshold values, not much can be concluded, even if no reliability issues are considered.

Year	Ticker	Degree	Name	Sector
2000	NOS	16	NOS SGPS	Telecom
2001	SON	18	SONAE SGPS	Investment Management
2002	NOS	18	NOS SGPS	Telecom
2003	NOS	14	NOS SGPS	Telecom
2004	BCP	15	BANCO COMR. PORTUGUES 'R'	Banking
2005	SNC	16	SONAE COM LIMITED DATA	Texlecom
2006	IPR NBA	13	IMPRESA SGPS NOVABASE	Media Information Technology
2007	ALTR	21	ALTRI SGPS	Construction
2008	CPR SON SONI	21	CIMENTOS DE PORTL.SGPS SONAE SGPS SONAE INDUSTRIA SGPS	Industrial Services Investment Management Industrial Services
2009	ALTR	23	ALTRI SGPS	Construction
2010	SONI	22	SONAE INDUSTRIA SGPS	Industrial Services
2011	ALTR	21	ALTRI SGPS	Construction
2012	BES	20	BANCO ESPIRITO SANTO SUSP - 03/03/15	Banking
2013	BPI	25	BANCO BPI	Banking
2014	BCP	23	BANCO COMR. PORTUGUES 'R'	Banking
2015	BCP	24	BANCO COMR. PORTUGUES 'R'	Banking

Table 6.4: Highest degree stocks in the positive graph.

Year	Ticker	Degree	Name	Sector
2000	ORE	12	OREY ANTUNES	Investment Management
2001	ORE	14	OREY ANTUNES	Investment Management
2002	SDCAE	8	SDC INVESTIMENTOS	Investment Management
2003	ESO	8	ESTORIL SOL 'B'	Sports
2004	ESFG	6	ESPIRITO SANTO FGP. (LIS) SUSP - 10/07/14	Investment Management
2005	ORE	12	OREY ANTUNES	Investment Management
2006	ORE	17	OREY ANTUNES	Investment Management
2007	GPA	11	IMMOBL.CON. GRAO-PARA	Investment Management
2008	COMAE GPA SCT	14	COMPTA IMMOBL.CON. GRAO-PARA TOYOTA CAETANO	Information Technology Investment Management Industrial Services
2009	SCT	19	TOYOTA CAETANO	Industrial Services
2010	FCP	9	FUTEBOL CLUBE DO PORTO	Sports
2011	GPA	21	IMMOBL.CON. GRAO-PARA	Investment Management
2012	FCP	12	FUTEBOL CLUBE DO PORTO	Sports
2013	COMAE MCP	13	COMPTA MEDIA CAPITAL	Information Technology Media
2014	ESO	24	ESTORIL SOL 'B'	Sports
2015	COMAE	20	COMPTA	Information Technology

Table 6.5: Highest degree stocks in the negative graph.

6.2 Analysis of Cliques and Clusters in the Graph

The discovery of node clusters in the market graph allows for the identification of stocks with correlated price variations. Cliques, Quasi-Cliques and K-Core Decomposition are used to cluster nodes. In this section, we will analyse the application of the aforementioned methodologies.

Firstly, we will look at the Maximum Cliques and interpret the results obtained in the context of financial markets. Afterwards, Quasi-Cliques and K-Core Decomposition are also analysed, mainly by comparing the cluster size of these approaches with that of the maximum clique size.

Regarding the analysis done, the financial interpretation of the three methodologies is generally similar, mainly due to the similar size of the resulting graph,

6.2.1 Maximum Cliques

The characteristics of the connected components found (Section 6.1) already indicate what to expect regarding the size of the maximum cliques, as the graphs are composed of a large connected component it is expected that they have large maximum cliques.

As shown in Table 6.6, maximum cliques are considerably big during two periods: from 2000 to 2002 and from 2007 to 2015. In these two periods, the edge density of the maximum clique is higher than 30% of the edge density of the original graph, and sometimes even above 50% (see the last column of Table 6.6).

Maximum cliques are sparser in the 2003-06 period. This can be explained by the fact that this timeframe had smaller edge density, as stock correlations are very close to 0 during these years. Note that for these years the thresholds used to construct the graphs were very low (10% to 15%).

The major conclusion is that, as time evolved, the maximum cliques got larger, meaning that the clusters of stocks with similar price variations tend to increase in size, as the years go by. This conclusion is in line with that of Boginski et al. (2003) that have found out and that modern stock markets have large groups of instruments that are correlated with each other.

Table 6.7 shows the number of stocks of each sector that are in the maximum clique of each graph. There are some sectors with permanent presence them being Banking, Investment Management, and Telecom. Additionally, as graph sizes expand from 2000 to 2015, it also brought the presence of companies from the Industrial Services and Construction sectors in the maximum cliques.

As explained in Chapter 3, cliques can be too restrictive as a clustering process, as, in some cases, they exclude nodes that are only missing a small number of edges to be eligible for the clique. This excessive restriction generates information losses, as a node that misses one or two edges in order to be in a clique is still a relevant element for the cluster. In order to flexibilize the clustering analysis, we will compare

the sizes of the clique induced graphs with the ones induced by Quasi-Cliques and K-Core Decomposition, in order to validate the premises above.

	Maximum Clique		Graph		ED_C/ED_G ratio
Year	Clique Size	Edge Density (ED_C)	Graph Size	Edge Density (ED_G)	
2000	9	7.7%	31	18.1%	43%
2001	10	8.0%	34	15.9%	51%
2002	10	8.0%	34	21.2%	38%
2003	6	2.5%	35	13.4%	19%
2004	5	1.7%	35	17.3%	10%
2005	6	2.4%	36	19.4%	12%
2006	4	0.9%	37	13.4%	7%
2007	14	12.9%	38	23.3%	55%
2008	11	6.7%	41	19.5%	34%
2009	13	8.2%	44	19.7%	42%
2010	16	12.7%	44	19.1%	66%
2011	18	15.5%	45	19.2%	81%
2012	15	10.6%	45	17.4%	61%
2013	14	8.8%	46	18.5%	48%
2014	13	7.5%	46	18.3%	41%
2015	14	8.8%	46	19.1%	46%

Table 6.6: Summary of maximum cliques for each graph.

6.2.2 Quasi-Cliques

As defined in Subsection 3.2.1, γ denotes the percentage of the number of edges of the maximum clique that is necessary to form the quasi clique. We will analyse γ values from 50% to 95% in order to compare the size of the quasi-cliques with that of the maximum clique for each graph.

As it can be seen from Table 6.8 and Figure 6.2, graph sizes, as expected, increase as γ decreases in an almost linear relation. This fact is aligned with the existence of a predominant big connected component, as there are always new nodes on the verge of being included into the quasi-clique as requirements relax.

These graphs can be divided into two different groups. The first one, from 2000 to 2006, where smaller relaxations have less impacts, as clusters increase slowly. From 2007 to 2015, increases are steadier. The explanation for this is that graphs in the first case are sparser than those of the second case, making it harder to generate cohesive clusters.

Year	2000	2001	2002	2003	2004	2005	2006	2007
Construction							1	2
Banking		1	3		2	2	1	2
Sports								
Industrial Services	1		1				1	4
Media	1	2	1			1	1	2
Information Technology	2	3	1	1				
Public Utilities								
Investment Management	1	1	1	1				2
Energy				1	1			
Consumer Goods	2							
Health								
Telecom	2	3	3	3	2	3		2

Year	2008	2009	2010	2011	2012	2013	2014	2015
Construction	2	2	2	2	2	2	3	3
Banking	1	1	3	3	2	3	3	1
Sports								
Industrial Services	2	4	3	4	3	3	2	1
Media							1	1
Information Technology		1						
Public Utilities								1
Investment Management	2	3	2	1	1	1	2	1
Energy	1		3	4	4	3	2	4
Consumer Goods	1		1	1				1
Health								
Telecom	2	2	2	3	3	2		1

Table 6.7: Sector distribuition of maximum clique nodes

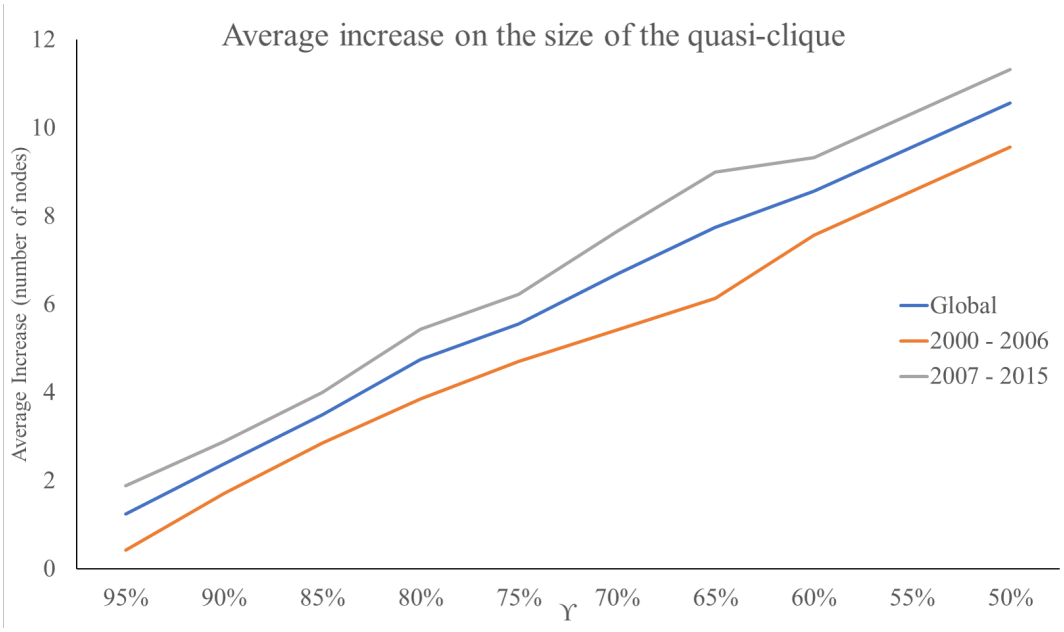


Figure 6.2: Increases of the quasi-clique size, comparing with maximum clique size.

One of the remarks done to this methodology is that working with ratios may lead to misleading conclusions, as it may aggregate highly cohesive nodes with some nodes that are very sparse (Seidman, 1983). Edge distribution was analysed for the quasi-cliques with γ greater than or equal to 80%.

Table 6.9 shows that from 2000 to 2006, quasi-cliques size increases slowly, and so, there are not many differences between the maximum and minimum number of edges connecting to a node. From 2007 onwards, quasi-cliques size increase quicker, leading to larger differences in the number of edges connecting nodes in the quasi-cliques. More specifically, it is possible to identify relevant differences between the maximum clique and the quasi-clique for $\gamma = 95\%$. For $\gamma = 80\%$, differences in the number of edges per node are drastic. For example, take into consideration the graph for the year 2015, where the most connected node has 20 edges and the least connected node only has 9 edges.

Summing up, the application of γ values close to 1 is beneficial to gather additional nodes into the cluster, i.e., includes highly connected nodes that would not be in the maximum case. For this case, this analysis showed to be coherent as it did not generate great unbalance in the number of relevant correlations of each stock. This means that, by using quasi-cliques, we can group more stocks into the same group than by considering the maximum clique, and so, have a broader set of stocks that explains the generalized behaviour of the market.

Year	Maximum Clique Size ($\omega(G)$)	γ									
		95%	90%	85%	80%	75%	70%	65%	60%	55%	50%
2000	9	10	10	11	12	12	13	13	13	16	17
2001	10	10	10	10	12	12	14	15	16	17	17
2002	10	10	13	14	15	16	16	16	18	19	19
2003	6	6	6	8	8	10	11	12	12	14	14
2004	5	5	9	10	12	13	13	13	16	16	18
2005	6	8	8	10	11	12	13	13	16	16	18
2006	4	4	6	7	7	8	8	11	12	12	14
2007	14	16	17	18	19	19	21	22	22	24	24
2008	11	14	16	17	19	20	21	22	22	24	25
2009	13	16	18	19	20	21	22	24	24	24	24
2010	16	18	19	20	21	22	23	24	24	24	24
2011	18	19	20	20	20	20	23	23	23	23	27
2012	15	16	16	18	19	19	19	22	22	24	25
2013	14	16	16	19	19	22	23	24	24	26	27
2014	13	14	14	14	19	19	22	24	25	26	26
2015	14	16	18	19	21	22	23	24	26	26	28

Table 6.8: Quasi-clique size for different γ values.

6.2.3 K-Core Decomposition

Recall that a K-Core is a sub-graph composed of a set of nodes which have at least k edges connecting to the remaining nodes of the sub-graph. The degeneracy of the K-Core is the maximum possible value of k for which the referred property is satisfied (Shirokikh et al., 2013).

The size of the K-Core is at least as large as the maximum clique size, thus the K-Core is a relaxation of the notion of clique.

Table 6.10 shows that this methodology is in fact a less restrictive tool to cluster nodes. K-Core Decomposition shows to produce larger sub-graphs for 7 of the 16 instances: 2002 to 2006, 2009 and 2015. For the latter two cases, 2009 and 2015, the differences are not to significant, since only a small number of additional nodes is considered. However, for the period form 2003 to 2006, K-Cores aggregate many more from nodes than the maximum clique. It is important to remember that during this period, graphs are sparser and the maximum cliques are very small. When we look to the degeneracies of those K-Cores (5 to 7), it is noticeable that these are larger than the number of edges per node of the maximum clique (3 to 5). This fact leads to the conclusion that the number of nodes in the K-Core is more stable throughout the years, when compared with the maximum clique size, as it is less prone to variations in the number of edges.

	Number of edges per node											
	$\omega(G)$ -1	gamma = 95%			gamma = 90%			gamma = 85%			gamma = 80%	
		Nodes	No. of nodes		Nodes	No. of nodes		Nodes	No. of nodes		Nodes	No. of nodes	
Year			Min	Max		Min	Max		Min	Max		Min	Max
2000	8	10	7	9	10	7	9	11	6	10	12	5	11
2001	9	10	9	9	10	9	9	10	9	9	12	6	11
2002	9	10	9	9	13	9	12	14	8	13	15	7	14
2003	5	6	5	5	6	5	5	8	5	7	8	5	7
2004	4	5	4	4	9	7	8	10	6	9	12	7	11
2005	5	8	6	7	8	6	7	10	6	9	11	6	10
2006	3	4	3	3	6	4	5	7	4	6	7	4	6
2007	13	16	12	15	17	10	16	18	10	17	19	9	18
2008	10	14	11	13	16	11	15	17	10	16	19	9	18
2009	12	16	12	15	18	11	17	19	11	18	20	9	19
2010	15	18	12	17	19	11	18	20	8	19	21	6	20
2011	17	19	15	18	20	12	19	20	12	19	20	12	19
2012	14	16	11	15	16	11	15	18	10	17	19	7	18
2013	13	16	13	15	16	13	15	19	10	18	19	10	18
2014	12	14	9	13	14	9	13	14	9	13	19	11	18
2015	13	16	13	15	18	12	17	19	11	18	21	9	20

Table 6.9: γ-quasi-clique maximum and minimum number of edges per node for γ and number of nodes of the maximum clique $\omega(G)$.

From the analysis above, one can conclude that this method has better clustering results than the maximum clique, as it allows to aggregate more nodes on sparser graphs. At the same time, the fact that it is less sensitive to variation in the graphs provides a certain degree of stability to the clustering analysis.

	K-Core		Maximum Clique	
Year	Degeneracy (number of edges)	K-Core Size (number of nodes)	Edges (n-1)	Clique Size (n)
2000	8	9	8	9
2001	9	10	9	10
2002	9	13	9	10
2003	5	10	5	6
2004	7	9	4	5
2005	6	13	5	6
2006	5	13	3	4
2007	13	14	13	14
2008	10	11	10	11
2009	13	15	12	13
2010	15	16	15	16
2011	17	18	17	18
2012	14	15	14	15
2013	13	14	13	14
2014	12	13	12	13
2015	13	15	13	14

Table 6.10: Comparison of the k-cores and maximum clique sizes.

6.3 Analysis of Graph Topological Stability

The analysis of the graph topological stability has the goal of assessing if graphs are robust to changes in the number of nodes. The vertex attack method will be used, as proposed by Huang et al. (2009), with the objective of analyzing the reductions in the number of edges of the largest component of each graph by removing some nodes, as well as the edges incident to them. In this sense, the smaller the reductions are, the more robust is the graph.

Huang et al. (2009) studied the effects of node removal by taking into consideration different thresholds θ used to construct the graph (as on Chapter 5 of this dissertation). They were able to conclude that graphs are more prone to selective node removals as threshold values increase, as edges are sparser and the graphs smaller. Our approach differs from that of Huang et al. (2009) in the sense that we will be removing a specific number of nodes ranging from 1 to 5. The reasons behind this choice are two: firstly our graphs are small, varying between 31 and 46 nodes, while the one used by Huang et al. (2009) had 1080 nodes. Secondly, Huang et al. (2009) use the graph threshold θ as a variable throughout their work. Our dissertation is based on the work of Shirokikh et al. (2013), thus we use a predefined threshold value to construct the graphs for all scope of analysis.

As addressed in Chapter 4, graphs can be more or less volatile to random or selective attacks. Huang et al. (2009) proved that, as the threshold value increases, the graph becomes more exposed to the selective removal of the highest degree nodes, since they become smaller, and so, such nodes become increasingly more relevant.

This section is divided into two different streams of analysis: first, we analyse the effects of the selective removal of the highest degree nodes. Then, the impacts of selective and random node removal are compared using the RR index, which evaluates the average difference in the number of edges of the largest component size after those removals.

6.3.1 Selective Node Removal

The selective node removal represents the effects of specific stocks. In here, the stocks considered are the ones associated with the highest degree nodes, i.e., the stocks with the most meaningful correlations. It is intended to analyse how the graphs decrease in size with such removal, knowing that the smaller the graphs, the more exposed the market exposed is such stocks. In more detail, this work will assess the consequences of iteratively removing the highest degree nodes of the graphs. As expected, the impacts are larger on the sparsest graphs, i.e., on those with less edges. It was these specific graphs that the removal of such nodes results in larger reductions on the number of edges of the largest component. Yhe graphs from 2000 to 2006, which are the sparsest ones, except for the graph for 2005, lost between 58% and 70% of their edges. From 2007 onwards, the percentage of edges

lost is smaller; however it is still highm between 48% to 56% of the edges.

More specifically, the graphs for the years 2000, 2001 and 2003, which are the 3 less dense graphs, felt the most impact after the first selective removal – 17% to 20% of the edges were removed. For the remaining years, 11% to 15% of the edges disappeared after the first removal. It should be noticed that the percentage of edges lost for successive edge removals almost follows a linear pattern (see Table 6.11).

Summing up, mainly due to its small dimension, the Portuguese Stock Market is highly exposed to highest degree stocks. The simulation of their delisting prooved to have a large impact on many other stocks, shown by the massive disappearance of edges.

Year	Number of nodes	Number of edges	$R_{\|CO\|}$ after selective node removal				
			Number of removed edges				
			1	2	3	4	5
2000	31	84	81%	64%	49%	38%	30%
2001	34	89	80%	65%	53%	42%	31%
2002	34	119	85%	71%	60%	49%	39%
2003	35	80	83%	68%	55%	45%	38%
2004	35	103	85%	73%	61%	51%	42%
2005	36	122	87%	75%	66%	57%	48%
2006	37	89	85%	72%	60%	49%	40%
2007	38	164	87%	75%	65%	55%	46%
2008	41	160	87%	74%	63%	53%	43%
2009	44	186	88%	76%	66%	56%	47%
2010	44	181	88%	77%	66%	57%	48%
2011	45	190	89%	79%	69%	61%	52%
2012	45	172	88%	78%	68%	59%	51%
2013	46	191	87%	75%	64%	53%	44%
2014	46	189	88%	77%	67%	57%	49%
2015	46	198	88%	77%	67%	58%	49%

Table 6.11: Summary of the reductions occurred in the largest component size by removing the highest degree nodes. The measurement unit used is the $R_{|CO|}$ which is the percentage of edges of the largest component that are kept after node removal.

6.3.2 Comparison between Selective and Random Node Removal

The comparison between the effects caused by the selective and by the random node removals will be performed via RR index. Recall that the RR index is the

average difference between the effects of stochastic and selective removal methods, taking into consideration the number of edges in the largest component after such removals. The value of the index transposes to which kind of attacks is the graph more susceptible. If RR is close to 0, then the graph is similarly affected by both attacks, otherwise, it responds differently to random and selective removals. If the index is positive, then selective attacks have larger effects in the graph than the random ones, while if the index is positive, the reverse is true.

We have created 100 instances in which we have randomly and iterativelly removed five nodes and then computed the $R_{|CO|}$ difference to the corresponding selective node removals. Afterwards, we have computed the average RR index, as being the arithmetic average of these differences. To produce the conclusions for this analysis we will take into consideration the average RR index value for each graph, as well as breaking down the analysis to the impact of each iterative removal.

The major conclusion of this study is that all graphs are more sensitive to selective attacks, as the RR index is always positive, in line with the conclusion of Huang et al. (2009) (Figure 6.3). This conclusion is very plausible in this specific case as these graphs are more exposed to a niche of high degree stocks, due to its smaller size and consequent relevance of these specific instruments.

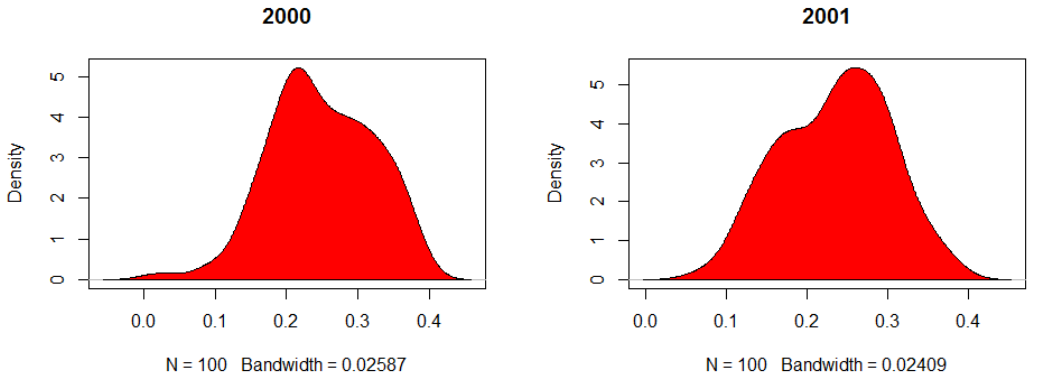
As the years go by, the RR index decreases. From 2000 to 2006, its average value of this spans between 18 p.p. to 25 p.p., meaning that selective attacks remove, on average, these percentage of edges. This means that, during this period, the highest degree stocks play a relevant part on the stability of the graph, which makes sense, as these are the sparser graphs for this time frame. When we look to the separate impact of each iterative removal, we see that for the removal of the one node ($f = 1$), the selective node attack removes many more edges than the random one, being followed by the second and third removals, where edge reductions also tend to be larger for the selective removal.

From 2007 until 2015, the graphs are denser, and, as a result, the RR index decreases to values rounding 10 p.p., translating that selective attacks still have more impact than random ones, but on a smaller degree. When we separate the impacts of each additional removal, we see that the pattern is differences from the previous period, with smaller and more constant differentials between selective and random node removals.

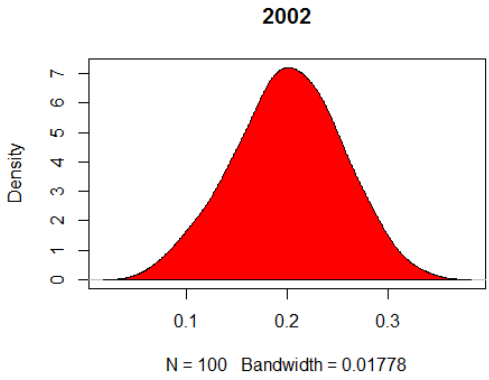
Overall, the fact that the Portuguese market is composed of few stocks exposes itself to every kind of alterations. Though the RR index shows that there is a clear difference in the impacts created by selective or random removals, the values are not sufficiently large to discard the effects of random attacks. In conclusion, this means that in general the Portuguese stocks are very exposed to each other, and any kind of delisting would have impacts on several other stocks. Nevertheless it is also clear that there is a pattern shift during the sixteen year period: in the first years, the market was more exposed to the set of companies with most meaningful correlations, while from 2007 onwards, this difference tended to become smaller.

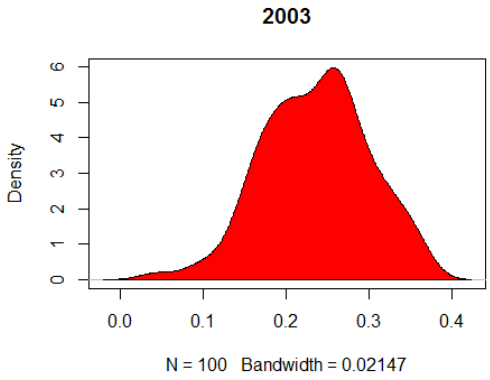
RR Index for the removal of f nodes						
Year	Average RR Index	f=1	f=2	f=3	f=4	f=5
2000	25 p.p.	11 p.p.	20 p.p.	28 p.p.	31 p.p.	33 p.p.
2001	22 p.p.	12 p.p.	20 p.p.	23 p.p.	27 p.p.	30 p.p.
2002	21 p.p.	9 p.p.	16 p.p.	21 p.p.	27 p.p.	31 p.p.
2003	24 p.p.	11 p.p.	20 p.p.	27 p.p.	31 p.p.	33 p.p.
2004	18 p.p.	8 p.p.	15 p.p.	20 p.p.	23 p.p.	26 p.p.
2005	17 p.p.	8 p.p.	13 p.p.	18 p.p.	22 p.p.	25 p.p.
2006	19 p.p.	8 p.p.	15 p.p.	21 p.p.	25 p.p.	28 p.p.
2007	11 p.p.	4 p.p.	8 p.p.	11 p.p.	14 p.p.	16 p.p.
2008	14 p.p.	6 p.p.	10 p.p.	15 p.p.	18 p.p.	20 p.p.
2009	11 p.p.	5 p.p.	8 p.p.	11 p.p.	14 p.p.	17 p.p.
2010	10 p.p.	4 p.p.	7 p.p.	10 p.p.	12 p.p.	14 p.p.
2011	9 p.p.	3 p.p.	6 p.p.	10 p.p.	13 p.p.	15 p.p.
2012	15 p.p.	6 p.p.	11 p.p.	16 p.p.	20 p.p.	23 p.p.
2013	14 p.p.	5 p.p.	10 p.p.	14 p.p.	18 p.p.	21 p.p.
2014	11 p.p.	4 p.p.	8 p.p.	12 p.p.	14 p.p.	16 p.p.
2015	11 p.p.	5 p.p.	8 p.p.	11 p.p.	14 p.p.	17 p.p.

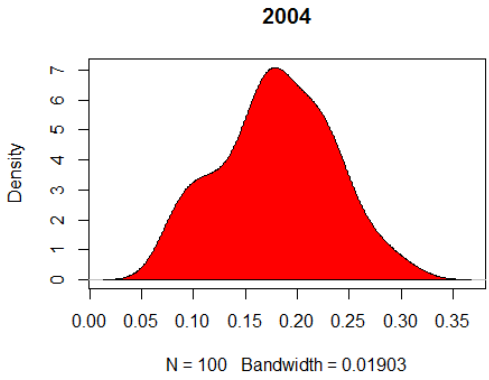
Table 6.12: Comparison of the differences caused on the largest component size ($R_{|CO|}$) caused by selective and random node removals.

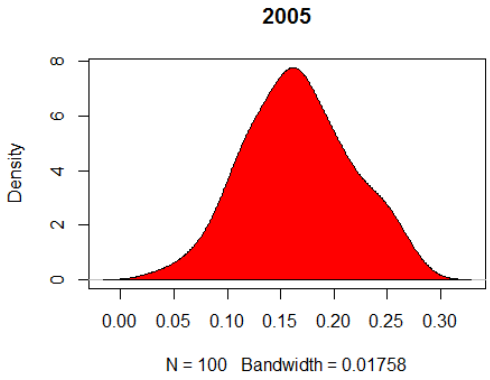

2000
Density
0.0 0.1 0.2 0.3 0.4
N = 100 Bandwidth = 0.02587

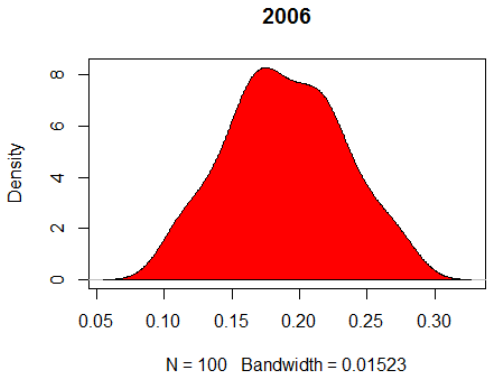
2001
Density
0.0 0.1 0.2 0.3 0.4
N = 100 Bandwidth = 0.02409

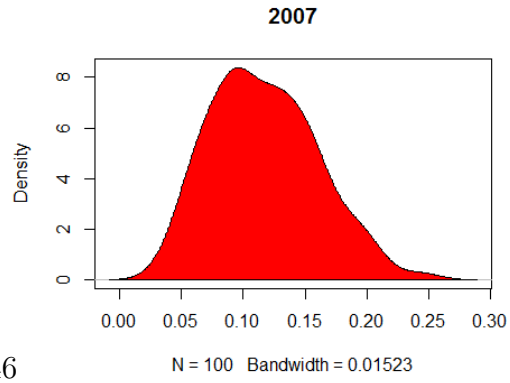

2002
Density
0.1 0.2 0.3
N = 100 Bandwidth = 0.01778


2003
Density
0.0 0.1 0.2 0.3 0.4
N = 100 Bandwidth = 0.02147


2004
Density
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35
N = 100 Bandwidth = 0.01903


2005
Density
0.00 0.05 0.10 0.15 0.20 0.25 0.30
N = 100 Bandwidth = 0.01758


2006
Density
0.05 0.10 0.15 0.20 0.25 0.30
N = 100 Bandwidth = 0.01523


2007
Density
0.00 0.05 0.10 0.15 0.20 0.25 0.30
N = 100 Bandwidth = 0.01523

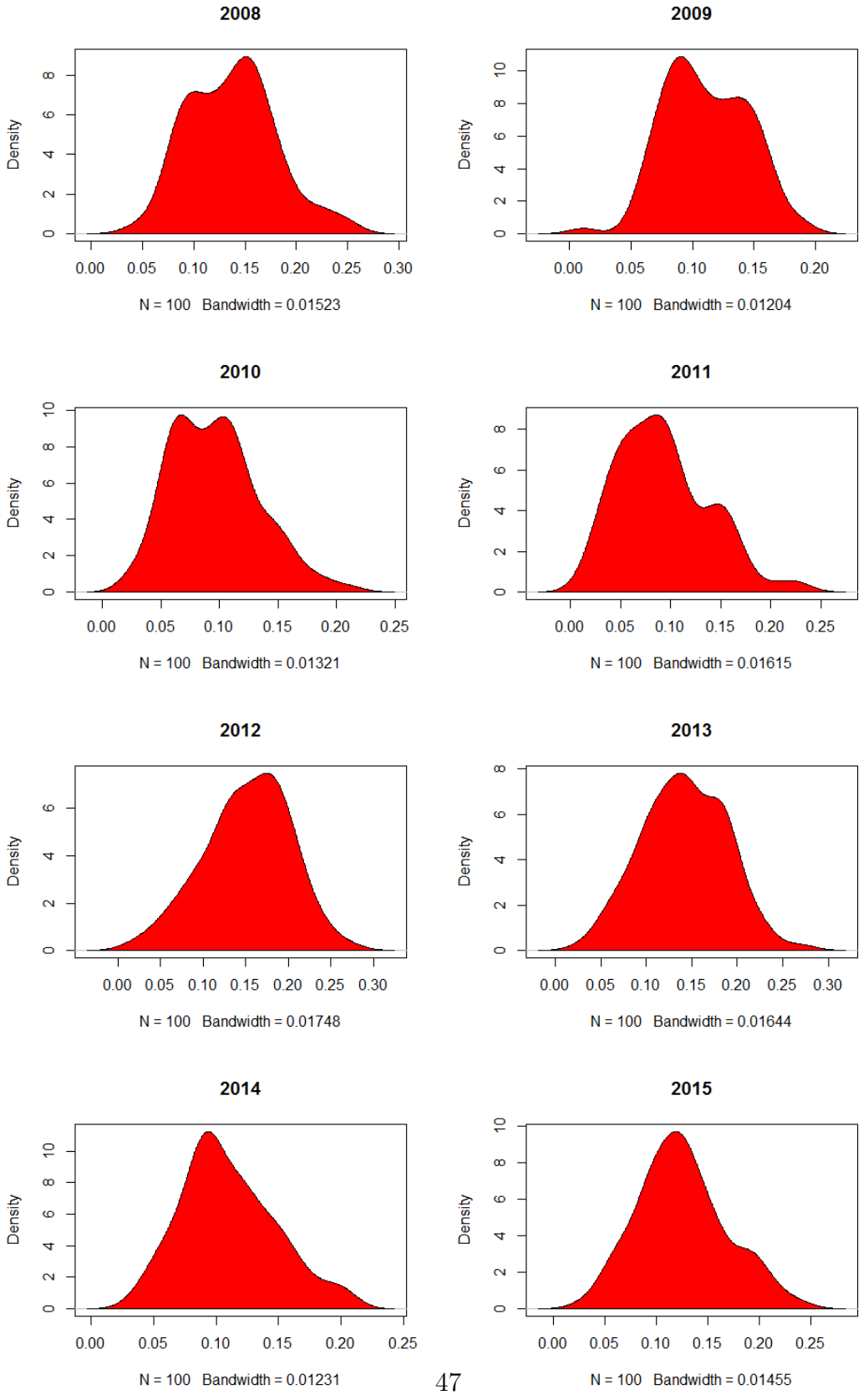


Figure 6.3: Distribution of RR index for years 2000 to 2015

Chapter 7

Conclusions

The main objective of this dissertation was to study the degree of association between the price variations of PSI Geral stocks. For that, the Market Graph was used, based on the works of Boginski et al. (2003); Shirokikh et al. (2013). More specifically, it was aimed to study stock price correlation patterns, as well as to search for groups of stocks with similar price behaviours. Additionally, it was also part of the scope of this project to study the effects of company delistings in the stability of the graph and this way assess the vulnerability of the market regarding such events. To achieve these goals, stock price correlations were calculated using the Spearman Rank Correlation Coefficient. Afterwards, the most meaningful correlations were represented by edges that connected nodes representing the corresponding stocks.

We can divide the time frame into two different periods: from 2000 to 2006 and from 2007 to 2015.

In the first period, the correlation behaviour was more dynamic, with edges changing very often from one year to the next. From 2007 onwards, an increasing stability was noticeable and strong correlations very often remain for longer periods. Regarding the study of companies with more relevant correlations, we see that until 2006, there was a strong predominance of Telecom companies. After, from 2007 to 2011, these companies were from the Industrial Sector and finally, from 2012 to 2015, from the Banking Sector.

The study of maximum cliques allowed us to understand that the dimension of such cohesive clusters is quite linked with the abundance of strong correlations in the graph. This means that the period with less correlations, namely, from the year 2003 to 2006, is also the one with the smallest cohese clusters of nodes. As time evolved, maximum cliques got larger, meaning that the clusters of stocks with strong price correlations tended to increase in size, as the years went by. When we look to the stocks present in the maximum clique, we note that there are some industries with permanent presence, such as Banking, Investment Management and Telecom, and others that gained their space in the most recent years, as is the case of Industrial Services and Construction.

The study of clique relaxations showed that both provide improvements to the maximum clique clique for clustering purposes. In the case of the quasi-clique, even the use of γ values close to 1 was beneficial as it allowed to consider some stocks that had a considerable number of strong correlations although not to all other stocks. The K-Core Decompostion showed its advantages mainly in the years when the correlations were more concentrated around 0.

Finally, the graph topological stability of the graphs was assessed, showing that the Portuguese Stock Market, due to its small dimension is quite exposed to almost all of its companies. Nevertheless, it was noticeable that, as time went by, the market became less exposed to the delisting, even if the most correlated stocks are correlated.

The main goal of this dissertation was the study of the Portuguese Stock Market. However, we must be recognized that this market is very small and thus it is difficult or even impossible to generalize the conclusions drawn. Nevertheless, many of them are in line with those of Boginski et al. (2003); Shirokikh et al. (2013); Huang et al. (2009), which have studied much larger markets.

As already mentioned the Portuguese Stock Market is very small and it is difficult to draw conclusions. Therefore, one line of future research is to develop this analysis using a broader data set. A good example would be to study Euronext, which comprises stock indexes from Portugal, the Netherlands, Belgium, the United Kingdom and France (Euronext, 2017). The addition of other elements such as commodities as petroleum or gold, interest rates and exchange rates would also be of great interest.

By constructing larger graphs, it will be possible to construct graphs with higher thresholds and smaller edge densities, this way bringing more confidence to the representation of significant correlations via edges. This would enlarge the scope to such analysis as: the study of intra and intersectoral connections, i.e., understanding how edges connect nodes from the same sector or which sectors share the most price correlations. The addition of other instruments allows the study of market exposure to the price of petroleum or to a specific exchange rate, for example.

Nevertheless, there are other options that can be explored, which involve changes to the current methodology. The first is to use weighted edges, depending on how strong is the correlation. This way, it would be possible to discard the threshold to build the graph, turning the clique construction into a maximum weighted clique problem instead of the actual maximum clique with a predefined threshold to define the edges.

Other relevant analysis would be to construct the graph with weighted nodes, for example using stocks capitalization as a reference. In this case, the selective removal of the highest degree nodes would also show how much the market value drops after the removal, in addition to analysing the counting the number of edges lost.

Finally, other clique relaxations could be used to somehow aliviate the disadvantages of the ones used. Recall that quasi-cliqes only look at the overall densitiy and

therefore may allow low connected nodes as long as their presence is compensated by highly connected ones. K-Core, on the other hand, only look at the number of edges incident to the nodes. Joining the ideas behind both could lead to a sub-graph with interesting characteristics as it would have to satisfy a global connectivity measure (density threshold) together with a local connectivity measure (node degree).

Bibliography

Aiello, W., Chung, F., and Lu, L. (2001). A random graph model for power law graphs. *Experimental Mathematics*, 10(1):53–66.

Boginski, V., Butenko, S., and Pardalos, P. M. (2003). On structural properties of the market graph. *Innovations in financial and economic networks*, pages 29–45.

Boginski, V., Butenko, S., and Pardalos, P. M. (2005). Statistical analysis of financial networks. *Computational statistics & data analysis*, 48(2):431–443.

Bollobás, B. (1978). Extremal graph theory, volume 11 of london mathematical society monographs.

Bollobás, B. and Thomason, A. (1985). Random graphs of small order. *North-Holland Mathematics Studies*, 118:47–97.

Bomze, I. M., Budinich, M., Pardalos, P. M., and Pelillo, M. (1999). The maximum clique problem. In *Handbook of combinatorial optimization*, pages 1–74. Springer.

Campbell, J. Y., Lo, A. W.-C., MacKinlay, A. C., et al. (1997). *The econometrics of financial markets*, volume 2. princeton University press Princeton, NJ.

Chiba, N. and Nishizeki, T. (1985). Arboricity and subgraph listing algorithms. *SIAM Journal on Computing*, 14(1):210–223.

Erdös, P. and Rényi, A. (1959). On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6:290–297.

Erdös, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(17-61):43.

Erdös, P. and Rényi, A. (1961). On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12(1-2):261–267.

Euronext (2017). Euronext regulated markets.

Gary, M. R. and Johnson, D. S. (1979). Computers and intractability: A guide to the theory of np-completeness.

Grané, A. and Veiga, H. (2010). Wavelet-based detection of outliers in financial time series. *Computational Statistics & Data Analysis*, 54(11):2580–2593.

Groth, D. and Skandier, T. (2005). Network+ study guide, fourth edition'. sybex. *Inc.. ISBN 0-7821-4406-3*.

Huang, W.-Q., Zhuang, X.-T., and Yao, S. (2009). A network analysis of the chinese stock market. *Physica A: Statistical Mechanics and its Applications*, 388(14):2956–2964.

Namaki, A., Shirazi, A., Raei, R., and Jafari, G. (2011). Network analysis of a financial market based on genuine correlation and threshold method. *Physica A: Statistical Mechanics and its Applications*, 390(21):3835–3841.

Pardalos, P. M. and Rebennack, S. (2010). Computational challenges with cliques, quasi-cliques and clique partitions in graphs. In *International Symposium on Experimental Algorithms*, pages 13–22. Springer.

Pardalos, P. M. and Xue, J. (1994). The maximum clique problem. *Journal of global Optimization*, 4(3):301–328.

Pattillo, J., Youssef, N., and Butenko, S. (2013). On clique relaxation models in network analysis. *European Journal of Operational Research*, 226(1):9–18.

Pena, D. (2001). Outliers, influential observations, and missing data. *A course in time series analysis*, pages 136–170.

Relvas, R. (2016). Psi 20 perdeu 44,7 mil milhões desde 2007. [Online; posted 12-April-2016].

Seidman, S. B. (1983). Network structure and minimum degree. *Social networks*, 5(3):269–287.

Shahinpour, S. and Butenko, S. (2013). Algorithms for the maximum k-club problem in graphs. *Journal of Combinatorial Optimization*, 26(3):520–554.

Shirokikh, O., Pastukhov, G., Boginski, V., and Butenko, S. (2013). Computational study of the us stock market evolution: a rank correlation-based network model. *Computational Management Science*, 10(2-3):81–103.

Watts, D. J. (1999). *Small worlds: the dynamics of networks between order and randomness*. Princeton university press.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442.