UNIVERSITÀ DEGLI STUDI DI PALERMO

Dottorato di Ricerca in Scienze Fisiche

Dipartimento di Fisica e Chimica - DiFC

SSD FIS/07

Statistical validation of investment decisions and transactions in financial markets

PHD CANDIDATE

Federico Musciotto

COORDINATOR

PROF. Gioacchino Massimo Palma

TUTOR

PROF. Rosario Nunzio Mantegna

CICLO XXX

2018

Contents

1 Introduction

My PhD research has focused on the empirical analysis of financial complex systems, with a specific insight in the detection of their relevant structural features through the adoption of tools and methodologies borrowed from the fields of Physics and Statistics. The aim of this dissertation is to provide both an exhaustive overview of the methods and the tecniques adopted and a complete description of the achieved results. This chapter is intended as a short guide to the epistemological context within which my research takes place.

In the last few decades humankind has witnessed many groundbreaking revolutions in several aspects of everyday life. These innovations, both technological and cultural, are reshaping the way man thinks of himself and his surroundings and of the impact that he has in shaping the world he lives in. Many brand new ideas in medical sciences, social sciences, economics, biology and many more fields are radically changing our lifestyle, but more importantly they are modifying our way of conceptualizing reality. Despite the high heterogeneity of impulses that are expanding human knowledge in a large spectrum of different directions, there is a common denominator which is shared by many of them. Indeed, one of the greatest revolutions that has been carried out in many disciplines since the second half of the XX century is the shift from the focus on the single components that contribute to a phenomenon, to an interest in the collective properties of the phenomenon itself. This awareness lies at the basis of the broad discipline which is now put under the umbrella term of science of complexity. As it usually happens when such a big shift in human perspective occurs, this change was not an abrupt process, breaking out all of a sudden: its seeds had been slowly spread during the previous decades, or even centuries, before the full awareness of their implications reached the critical point required to build a new paradigm. What follows is just a very short and not exhaustive description of the most significant steps in the roadmap that has brought humankind to a more refined idea of complexity. It is not intended as a detailed chronology of complexity-related developments during human history, but simply as an introduction to some of the main concepts that will be used in this dissertation.

One crucial moment in the development of modern mathematics, which is most likely to be found in any textbook of Network Science, is the event that brought the prolific Swiss mathematician Leonhard Euler to lay the foundations of graph theory almost three centuries ago. In 1735, while he was visiting the Prussian city of Königsberg, Euler found the solution to a puzzle which was popular at that time: was it possible to cross all the seven bridges that connected the four different parts in which the river Pregel split the city without crossing any of them twice? The answer is no, and in order to prove it Euler build a graph with four vertices (the pieces of land) and seven links (the bridges), building a rigorous proof based on the properties of the degree sequence of the graph [1], [2], as it is shown in Figure 1.1. Although it is impossible to say whether Euler was aware of the impact that his idea would have had in the following centuries, it is evident that the full possibilities of graph theory were not explored until last century. Indeed, one of the most powerful implications of network science, which is based on the mathematical principles of graph theory, is that no matter which kind of agents and relations are involved in a given phenomenon, visualizing the set of all interactions as a graph allows to analyze the structure of the system, underline its weaknesses and predict its evolution.

An other relevant building block of the science of complexity is the enunciation of the laws of statistical mechanics, in the late XIX century. The bases of this field were mainly built by Ludwig Boltzmann, an Austrian physicist that dedicated his life to expanding the embryonic ideas of Daniel Bernoulli and James Maxwell on the motion of molecules. The principles of statistical mechanics added a fundamental layer of awareness to scientific knowledge, showing that analytically solving the equations of motion of each single component of a set of interacting particles is not always convenient (nor possible). Indeed, when the number of these particles grow significantly, writing down their equation of motions becomes practically impossible. And even if it were possible, limits in computational power and time make it a huge and much complicated task.

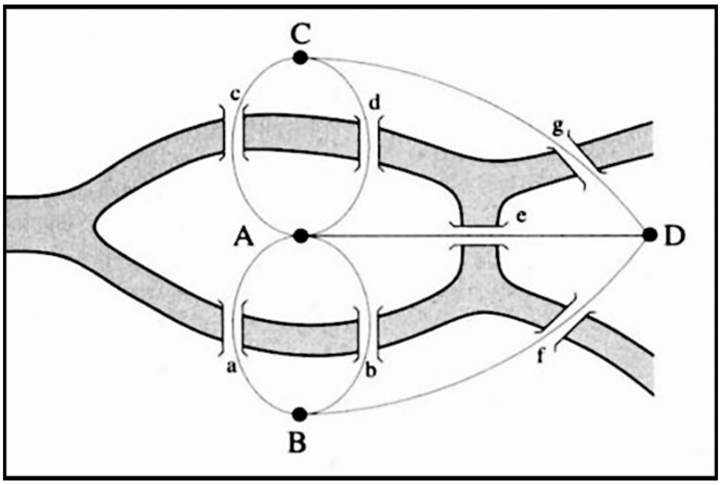


Figure 1.1: Schematic representation as a graph of the city and the bridges of Königsberg as they looked like at the time of Euler.

The answer of statistical mechanics is that sometimes it is just more efficient to describe a process looking at its statistical properties, which can be computed by generalizing, simplifying and aggregating the behavior of the single parts.

Going further in time, the beginning of XX century saw Jacob L. Moreno, an Austrian-American psychiatrist and social scientist, lay the foundations of group sociometry. One of Moreno's main contributions to social science may be found in his pioneristic work on social groups and the focus on the role played within them by human interactions. In fact, among his other merits, Moreno's is largerly recognized as the founder of social network analysis [3], which knew a great expansion from the sixth decade of XX century and contributed to the great interest that has been converging toward network science for the last decades. The relevance and innovation of his work may be effectively summarized in the words that he himself addressed to Sigmund Freud after a lecture of the latter, and that sound as a working manifesto of his beliefs: "Well, Dr. Freud, I start where you leave off. You meet people in the artificial setting of your office. I meet them on the street and in their homes, in their natural surroundings. You analyze their dreams. I give them the courage to dream again. You analyze and tear them apart. I let them act out their conflicting roles and help them to put the parts back together again."

Despite their heterogeneous background with respect to discipline and historical period, all these developments intuitively add their contribution to shape a modern definition of the concept of complexity. Indeed, all these episodes share a critique towards the idea of reductionism, i.e. the principle of focusing on the properties on the single object without looking at its interactions with its surroundings, that dominated western science for a long time. This conceptual revolution has been greatly enhanced also by two emerging factors: the exploding increase in computational power and the constant growth of data availability. Indeed, as Moore's law, inspired by Ref. [4], famously claims, the number of transistors in a dense integrated circuit doubles approximately every two years. This exponential growth has unleashed a surprising potential in data processing, with the diffusion of increasingly small processors able to perform in a breeze calculations that decades ago would have required years of manual work. Combined with this, the considerable developments in the production of cheap, space-saving storage supports have brought to the increase of the amounts of available data. Data analysis is nowadays an element with increasingly high impact on our lives: it has created new professional figures, represents a high source of income for many big corporations, thus reshaping our economical landscape, and poses new ethical dilemmas about privacy and freedom in modern society [5]. Moreover, the growth of data availability has dramatically enlarged the number, the category and the size of systems that can be analyzed by scientific research.

Along with the opportunities that unleashed, this technological revolution has brought science to face new complexity-related challenges. Indeed, dealing with huge amount of data can often lead to confusing or unreadable information, due to the size and the heterogeneity of the systems under analysis. In such cases, it is required to filter out redundant noise and detect only relevant, significant information. A working example of the implications of these challenges can be found in the field of econophysics, which is aimed to the analysis of finance/economy related systems through the tools of Physics and Statistics [6]. Indeed, the technological revolution of the last decades has continuosly enlarged the order of magnitudes of available data, making increasingly difficult the detection of stable, significant patterns. Moreover, economics well represents the idea of a world dominated by strong, untouchable dogmas that would greatly benefit from a scientifical, complexity-driven revolution [7]. To this extent, an increasingly large set of tools has been developed in order to deal with these challenges. This thesis follows this line of research.

This dissertation is organized as follows. Chapter 2 provides both a theoretical background and a review of the methods that are extensively used in the rest of this work, with a particular stress on statistically validated networks and their applications [8]. Chapter 3 reports the result obtained in the investigation of the trading strategies of investors at the Nordic Stock Exchange (NSE), venue of Helsinki. Specifically, the first part of the chapter is devoted to the description of a method capable to detect clusters of investors characterized by similar trading activity. Moreover, a comparison with the information contained in hierarchical trees is made. The second part is about the analysis of the dynamics of these clusters, with a focus on the features of the ecology of investors that emerges from the investigation, following a line of research proposed in [9]. These results help in obtaining a better understanding of characteristics and evolution of the different types of investors that coexist in the dynamics of a stock market. Chapter 4 presents an analysis of the dynamics of investors approaching the stock market for the first time during the unfolding of the *dot com* financial bubble, within the Helsinki venue of the NSE. The first part of this analysis is about the demographical characterization of the investors which were buying the Nokia asset for the first time during the bubble. This characterization is obtained through the tools of statistical validation. The second part of the chapter introduces an agent based model designed in order to reproduce the dynamics of the underlying process. In Chapter 5 the tools of statistical validation are applied to a database of financial transactions with the aim of analysing the impact of high frequency trading on the structure of the network of investors. Specifically, the approach has been designed in order to investigate whether high frequency trading affects the random interaction of traders supposed by considering the anonimity of traders in the stock market. Finally, in Chapter 6 the results shown in the previous chapters are discussed, and the conclusions are drawn.

2 Theory and methods

The outbreak of the concept of complexity in modern science has led to the definition and implementation of a large set of disciplines and methods, conceived to deal with its challenging implications. In what follows, a more rigorous definition of complexity is provided. Afterwards, short reviews of network science and agent based modeling, which are closely related to the work presented in this dissertation, are presented. Finally, a detailed overview of the statistical methods and filtering techniques applied in the following chapters is reported.

2.1 Complexity

Providing a rigorous and widely accepted definition of complexity is not an easy task, due to its multidisciplinary nature. However, a simple yet hopefully exhaustive characterization of it can be obtained by putting together its most common traits. A system is considered complex if it is made up of many heterogeneous agents whose interactions generate a highly nonlinear dynamics, which is characterized by the presence of feedback among agents and emergent phenomena. In the following list all these ingredients are analyzed in detail.

- **Size.** Complex systems are usually made up of a great number of single components. The order of magnitude may vary from a discipline to another, but in order to legitimely speak of complexity usually at least dozens of agents are required. Empirical examples of the orders of magnitude that can be found in different disciplines are:
    - Biology: one can range from the interactions between different areas of brain (10^2) to the dynamics of the cells that constitutes it (10^9).

  - Ecology: the number of species on Earth is close to ten millions, although most studies focus only on subsets of this quantity.
  - Technology: the amount of sites in the World Wide Web which are known is already huge (10^9), although a comprehensive map of its structure is still missing.
  - Society: the analysis of human behavior can range from actual social interactions ($10 - 10^2$) to state-level scale ($10^6 - 10^9$).

- **Heterogeneity.** Counting the number of interacting agents within a system is not enough to account for its complexity. To fully understand its dynamics one should look at how differently these agents act. Indeed, heterogeneity is one of the most characterizing features of complexity. This is the case in ecology, with different species playing different roles (e.g. preys vs predators), in economics, with different class of investors adopting different trading strategies, in society, with all the roles that it is possible to have in a social system, and so on.

- **Nonlinearity.** The key principle of linearity is that a small stimulus generates a small reaction. However, in nature many examples that contradict this simple statement are easily found. This happens with the all-or-none law in biology, that claims that there is a defined threshold that regulates the reaction of nerves to stimulus: if the stimulus is below the threshold no reaction is triggered, otherwise the entire impulse is discharged [10]. Other examples are found in society, with small rumours having a huge impact on elections, in economics, with big financial shocks that can be triggered by moderate inputs, and so on.

- **Feedback.** In a complex system, the actions of a class of agents usually have impact on the others, enforcing relationships of self-regulation and mutual adjustment. To better understand it, one can think of a system of preys and predators: the wealth of one class is directly related to the state of the other, in a constant interplay between the two.

- **Emergent phenomena.** In a system which has the features described so far, the dynamics is not likely to be a plain, predictable consequence of the aggregation of its components. In this context, an emergent phenomenon is a large scale, collective behavior of a system which is not directly explainable from its single constituents [11]. Examples of emergent phenomenon are social segregation in cities, viral epidemics, technological failure cascades, political turmoils and revolutions,... Actually, the first track of this awareness dates back to the eighth book of Aristotles Metaphysics, which brought to the famous sentence "the whole is more than the sum of the parts".

2.2 Network science

Network science is a discipline that deals with the analysis of complex systems through the adoption of the formalism of networks, whose mathematical foundation is borrowed from graph theory [12], [13]. In order to deal with this ambiguity in terminology, in what follows the term *graph* will refer to the mathematical object defined by graph theory, while *network* will be used to indicate the graph structure of empirical systems found in society or in nature. Although the origins of network science are rooted in social network analysis, as witnessed by Moreno's experience, the very last decades have seen an explosion of the number of fields in which it has been adopted. A key principle that lies behind this universality is holism. According to the holistic approach, in order to understand how a system works it should be analysed as a whole, and not just as the aggregation of its single parts. This is evident in the approach of network science, since it focuses on the complete structure of a system by looking at the interactions between its components, without taking into account their specific nature. This generality allows networks to represent many different systems, like social networks [14], the cells [15] or the areas of the brain [16], technological systems [17], financial entities [18] and many more other examples. In the following paragraphs a brief review of the most relevant concepts related to network science is provided.

2.2.1 Concepts and definitions

A topological graph G is a mathematical object defined by an ordered pair (V, E), with V being the set of N vertices (nodes) and $E \subseteq V \times V$ the set of edges (links) that connect couples of vertices. A graph G can also be written in form of its $N \times N$ adjacency matrix A which is defined as follows: $A_{ij} = 1$ if vertices i, j are connected by a link, 0 otherwise. A self loop is a link that connects a vertex with itself are not allowed. A simple graph is a graph with no self loops, so its adjacency matrix has zero values in the diagonal. A directed graph is a graph with directed links, thus $(i, j) \neq (j, i)$. This implies that the adjacency matrix of a directed graph is not symmetrical in general. A graph can be *weighted*, $G = (V, E, w)$, with $w : E \to \mathbb{R}$ a function that assigns to each link a real value. In a network, weights can refer to the frequency or the strength of interaction between the pair of vertices, to their distance,. . .

When the structure of a graph is known, different measures capable to describe its topological properties can be computed. Among these, the most popular are centrality measures. A centrality measure is an estimator of the importance of a vertex or a link with respect to the whole graph [12]. One of the simplest and most common examples of such a measure is degree centrality. The *degree* of a vertex i is defined as the number of vertices connected to i through links. It is quite straightforward to understand why degree is a centrality measure. Indeed, the higher is the number of neighbors of a vertex, the more likely it is for it to play a central role in the network structure. The degree sequence of a network is also connected with more abstract concepts like hierarchy [19], community structure [20] and the dynamics of processes spreading on the network itself [21]. Other examples of centrality measures, which rely on different criteria to estimate centrality, are closeness, betweenness, eigenvalue centrality,. . .

In a graph, a *path* between a pair of vertices i, j is a sequence of consecutive links that allows to go from i to j. For any pair i, j, more than a path can exist in general. The number of links in a path represents its *length*. The path of minimum length between two vertices i, j is called the *shortest path* between i and j. The *diameter* of a graph is the shortest path of maximum length among all pairs of vertices of the graph.

2.2.2 From randomness to scalefreeness

Since the end of World War II, network science has witnessed a great development, boosted by both the advancements in graph theory and the increase in data availability. One the major outcomes of this line of research is the observation that a huge variety of systems, despite their completely different nature, share some peculiar structural properties. Probably the most famous example of this phenomenon is the extent of the diffusion of scale-free networks. This paragraph follows the process that led to this achievement.

In the central part of last century, two major actors on the scene of graph theory were Alfréd Rényi and Paul Erdős, two Hungarian mathematicians. Among their other contributions, their 1959 work on random graphs [22], is largely recognized as a milestone in the development of the discipline. In this paper they developed a model for a random graph conceived as a set of vertices whose links are assigned randomly according to a fixed probability. Their main goal was to build a class of graphs that reproduce the feature of real world networks. Nevertheless, although their model is still very popular in network science, they failed on this particular point. The reasons for this failure emerge clearly by comparing the features of real world networks with those of Erdős-Rényi model:

- **Degree sequence.** It can be easily proved [22] that the degree sequence of an Erdős-Rényi graph follows a Poisson distribution. This implies that the degree sequence has a finite variance and is distributed sharply around an average value. As it will be extensively shown in the following paragraphs, most real networks do not have this property.

- **Clustering coefficient.** The clustering coefficient of a graph is a measure of how much the neighbors of a given vertex tend to be connected among themselves. Indeed, the *local clustering coefficient* for a vertex i is defined as $C_i = \frac{2L_i}{k_i(k_i-1)}$, where L_i is the number of links between the neighbors of i and $\frac{k_i(k_i-1)}{2}$ it the number of all possible pairs of neighbors. Many real world networks have high values of clustering coefficient, which implies that in these systems interactions are locally correlated. On the other hand Erdős-Rényi graphs, whose links are placed in a totally random way, fail to reproduce this feature.

Another step towards the faithful description of real world networks was made by small world graphs, defined by Watts and Strogatz in 1998 [23]. The starting point of their work is the analysis of regular lattices, which can be represented as graphs in which every vertex is connected to a fixed number of neighbors. It is intuitive to understand why a regular lattice has high clustering coefficients, equals for all its vertices. Indeed, a fixed fraction of the neighbors of each vertex is always interconnected. However, since long range interactions are missing, the diameter of such a graph is large, contrary to what is observed in real networks. Thus, their proposal is to introduce a random rewiring of the links, occurring with a probability p. For $p = 1$ an Erdős-Rényi graph is obtained. For intermediate values, a graph with small diameter and high clustering coefficient emerges.

Nevertheless, small world graphs still leave unresolved the problem of degree distribution. Indeed, many networks naturally arising in nature and in society have a degree sequence whose distribution belongs to the same family, i.e. power laws. The general structure of a power law distribution is

$$p(k) = ck^{-\gamma}, \tag{2.1}$$

with c being a proportional constant and γ the exponent. The first one to detect a power law in a complex system was the Italian economist Pareto in his observations on the distribution of wealth in Italy in the XIX century [24], as he noted that income was not distributed with a typical scale around an average value. Indeed, few individuals earned great portions of wealth while the majority of population earned small amounts. Since then, power law distributions in networks have been found in biology [25], in social systems [14] and almost everywhere. A network with a power law distributed degree sequence is called *scale-free*, and it is characterized by the presence of few hubs, with most of the vertices connected directly to them, and the absence of a typical scale. Indeed, for power law distributions with exponent $\gamma \leq 3$ all moments of order greater than one are not finite, and it is not possible to compute their variance. This implies that there is not a typical scale in which the degree is more likely to be contained.

The universal presence of scale-free networks has arosen much interest in the phenomenon lying behind their formation. A successful modelization of scale-free graphs was presented by Barabasi et al. in 1999 [26]. In this paper the author showed that a scale-free graph with exponent $\gamma = 3$ emerges when adopting preferential attachment. In fact, in their model a set of vertices is added to a starting graph; each time that a vertex is added, m additional links that connect it with the old vertices are put. The new links are added with probabilities that are proportional to the degree of existing vertices. Thus, this work proves the existence of a link between preferential attachment and scale-free networks: the tendency to connect to the most influential vertices when entering a network produces a power law distribution. An alternative explanation, that deals with a different class of processes, is provided in [27], according to which power laws naturally arise from processes in which the dynamics follows a progressive reduction of the sample space. Additional examples are presented in Ref. [28], which provides a review of processes that produce a power law distribution, such as phase transitions and critical phenomena. Despite the variety of its conclusions, this line of research has helped to spread more light on the concepts of randomness and disorder in nature. Indeed, although many processes naturally arising in nature are generally conceived as truly random, the presence of scale-free networks reveals that this is not the case. Indeed, all these processes imply the existence of a precise direction in the evolution of systems that end up with a power law distribution. This means that in nature and in society interactions between elements do not spread out randomly, but instead follow well defined self-regulating laws.

2.2.3 Applications

Following its increasing popularity, network science has experienced the development of a considerable amount of applications and research areas. What follows is a short review of the most significant, mainly related to the subject of the present dissertation.

Community detection

The diffusion of *scalefreeness* has underlined that the degree sequence of real world networks often lacks a typical scale. However, a dishomogeneity in the distribution of links may be observed also locally. This dishomogeneity is associated with the presence of communities, which are observed at a mesoscopic scale. Indeed, although a unique, universally accepted definition does not exist, communities (or modules) are usually referred to as subsets of the vertices of a network which are more closely connected within themselves than they are with all the others. The community structure may be related to different aspects, like segregation/homophily or functional diversity, but it is one key aspect in understanding the global structure of the network.

Community detection has witnessed the development of a large number of different techniques, borrowed from different fields and aimed to consider different aspects of the problem [29].

- **Clustering.** The first methods developed for community detection were just an extension of methods adopted in traditional clustering, such as graph partitioning, hierarchical and partitional clustering.

- **Modularity optimization.** Modularity first appeared as a quality function for clustering algorithms in [30]. Simply put, it measures the fitness of a partition on a given network in comparison to a null model of random allocation of links. Thus, the point is to find the partitions with the highest scores of modularity. This task is not trivial, due to the ragged landscape that modularity usually has [31] and the computational problems related to a full exploration of the space of partitions. This complexity has led to the introduction of many different methods.

- **Random walks** The idea behind this class of methods is that, in the presence of a defined community structure, a random walker that goes through the links of a network would not spend his time homogeneously on all the nodes, but it would be temporary trapped in the different modules. Infomap [32], which is one of the most popular community detection algorithms, belongs to this class.

Multiplex networks

In many complex systems the interaction between agents is not limited to only one kind of action, but can be expressed through different channels. Examples of this aspect are (i) ecological systems, in which animals can interact through fight, reproduction or cooperation, (ii) social networks, in which people are exposed to different types of interaction (work, leisure time, family,. . .). One possibility is to collapse all different layers of interaction to an unique one, but this is not always the most rigorous and efficient possibility. For this reason a specific formalism has been developed to deal with these multilayered networks, also called *multiplex* [33].

Bipartite networks

The usual representation of networks, with a set of vertices in which anyone can virtually interact with anyone else, is in many cases limiting and misleading. Indeed, it often happens that the structural connections within a system naturally arise in a bipartite form. Bipartite graphs are designed to account for these systems. In them, vertices belong to two distinct sets, with no links allowed between vertices of the same set. Examples are the actor-movie IMDB network, in which the two sets are made up of actors and movies, with links representing the presence of the actor in the cast of the movie, or the network of scientific collaboration, in which the first set contains authors and the second papers. Classical "unipartite" networks can always be obtained by a bipartite one through a procedure called projection. The projection of one of the two sets is obtained by putting a link between all pairs of vertices that share at least a common neighbor in the other set. Although applying the tools designed for "unipartite" networks on projections is often reliable [34], bipartite networks contains additional layers of information. For example, looking only at the projection on actors of the IMDB network hides part of the information on the activity of single actors. An alternative to projection which takes into account the level of co-occurrent activity of the projected set is provided by the formalism of statistically validated networks (SVN) [8], which plays a central role in the present dissertation.

2.3 Agent based models

When dealing with a complex system, its evolution can be history dependent. Indeed, it is not possible to repeat the dynamics behind the social links that an individual builds during his lifetime, or to observe in a closed environment the interactions that animals usually develop on large space and time scales. In this context, building a toy model that describes computationally these interactions through autonomous agents that act according to a simplified sets of rules is often the only way to study the dynamics of such systems in depth. This opportunity is nowadays provided by agent based modeling (ABM). Indeed, according to Nigel Gilbert, "agent-based modeling is a computational method that enables a researcher to create, analyze, and experiment with models composed of agents that interact within an environment", [35]. Although ABM is generally regarded as a branch of computational social science, its connection with physics and natural sciences has been strong since its birth. Indeed, many have stressed how the Ising model [36] can be considered an early implementation of agents based modeling. In fact, the atoms that align their spin/magnetic moment according to the temperature of the system and the state of neighbors behave like basic agents of an ABM. Moreover, the formalism of Ising model has been often extended to the description of social/economical system. This was achieved by generalizing the concept of magnetization in order to represent many different processes, like opinion polarization [37] or metastability in a financial system [38]. Thus, it emerges how one of the major outcomes of ABM is to allow social/ecological sciences to adopt the paradigm of natural science. In fact, thanks to ABM it is possible to implement a system that follows a defined set of rules and create and observe multiple realizations of it.

Although an excessive simplification of interactions in ABM design may result in a loss in accuracy and predictive power, its careful adoption can help in effectively identifying the key actors of an emergent phenomenon, without being misleaded by non-relevant factors. In this sense, a well known example is provided by the model developed by Schelling in 1969, [39], which is one of the first implementations of an ABM. Schelling adopted a very simple framework to describe demographic related dynamics in a social environment. He modeled two different sets of agents, "red" and "blue", as points that lie on a bidimensional lattice; at $t = 0$ the agents are distributed homogeneously with respect to the two sets. When the simulation starts, they follow a "homophily rule". This means that, if an agent is not surrounded by a fixed fraction p of nearest neighbors of the same set, he can relocate to a random site. Running the model brought to an unexpected result. Indeed, the system ends up with segregation between the two groups even with values of p well below 0.50 (the critical threshold for the emergence of segregation is 0.37). This result is somehow surprising, since it stresses out that no active dislike or racism is required to make segregation emerge, but only a slight degree of homophily.

Since these first examples, ABMs have become more complex and sophisticated. This is reflected in the design of agents that, despite the great heterogeneity of models and developed techniques, shares some common traits:

- **Perception.** The agents have the ability to detect the changing properties of the environment and the behavior of neighbors.

- **Performance.** They can perform a set of actions, like moving, communicating, collecting resources, being born/dying,...

- **Memory.** They store information on their previous actions and states.

- **Policy.** They are assigned a set of rules that, taking into account their memory on previous states of the system, makes them follow a strategy compatible with their goals.

All these features contribute to make agents heterogeneous entities that (i) follow different individual strategies, (ii) are strongly reactive to environmental changes and to the interaction with neighbors and (iii) are able to learn (both individually and as whole species). This heterogeneous characterization makes ABMs a flexible and suitable tool that allows to model complex systems. An interesting example is provided by El Farol bar problem [40]. In this ABM, the goal of all agents is to go to the El Farol bar on Thursday in order to enjoy the "Irish music night". However, if the bar is too crowded, no one will enjoy it. Thus, agents are left with the elaboration of a strategy that let them go to the bar only when they expect that few others are going. After the original introduction this problem was formalized in terms of the so-called *minority game* [41]. One first evidence is that if all agents shared the same expectation for the night attendance a paradoxical result would be obtained. Indeed, if everyone is expecting the bar to be crowded no one will go, leaving it empty; otherwise, if all expect the others to stay at home, everyone will go to the bar, making it too crowded. Thus, the best way to face this problem is to make the agents elaborate different strategies. In [41] this is achieved by providing all agents with a memory on the outcome of previous nights and making them elaborate randomly a fixed number of strategies that produce a decision for the incoming night from the history of past ones. Then all the agents rank their own strategies on the basis of the successful decisions that they produce over time. The observation of this ABM shows that for some values of the model parameters some agents perform better than they would if taking random decisions night by night. Thus this model, despite its simplicity, allows to observe agents acting through sub-optimal heuristic rules of bounded rationality, which is something that happens also in other social systems like financial ones [9]. Indeed, the stock market can be seen as an example of minority game, in which it is convenient to sell when everyone else is buying, and viceversa. To this extent, the El Farol bar problem is an empirical proof that contrasts with the classical economic paradigm of the Efficient Market Hypothesis [42], that implies that all investors behave in an optimal rational way by efficiently exploiting all available information.

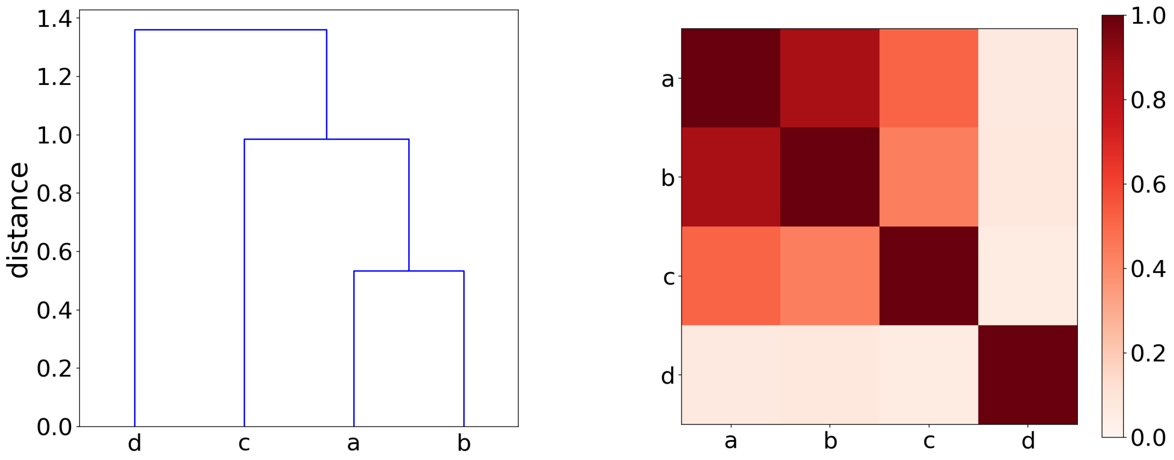
Chapter 4 of this dissertation will introduce a financial ABM, designed to describe the trading decisions of investors buying an asset for the first time during a bubble.

2.4 Statistical methods

The abundance of data and the increasing power of the available computational tools have greatly affected the scale of systems that it is possible to analyze with the approach of complexity. In this framework, the adoption of statistical methods has become crucial in order to detect stable, significant trends out of the huge amount of available information. Econophysics is one of the fields in which this line of research has helped in the detection of patterns that show great universality and validity across different datasets and systems, the so-called *stylized facts* [43]. This section is a short review of some of such methods.

2.4.1 Hierarchical clustering

Cluster analysis is a branch of statistical data analysis aimed to the detection of groups of elements (*clusters*) that partition the whole system. The criterion used in the detection of clusters is similarity. This means that elements within the same cluster should be very similar while elements belonging to different clusters should be distinct. Cluster analysis is characterized by a huge variety of different branches, tecniques and algorithms; this section is dedicated to hierarchical clustering. Hierarchical clustering is a branch of cluster analysis whose aim is to detect groups of elements with a nested hierarchical



(a)

(b)

Figure 2.1: Figure 2.1a shows the dendrogram of a simple system of 4 variables. The corresponding correlation matrix is plotted in Figure 2.1b. In the tree, the first elements to be connected are the most similar, a and b, at a distance of about 0.5. Then c and lastly d, which is the most dissimilar element, are added.

.

structure [44]. Hierarchical clustering algorithms can belong to two different categories: divisive and agglomerative. Divisive algorithms, also called *top down*, start from a partition in which all elements are grouped together and procede iteratively to split them in different clusters. On the other hand, agglomerative algorithms (*bottom up*) proceed the other way around. At the beginning all element are assigned to a different cluster; through iteration they are then aggregated according to their distance pattern.

Hierarchical clustering is associated with *dendrograms*. A dendrogram is a tree-like diagram that shows the hierarchical relationships between elements, together with the levels of distance at which they are merged in the same cluster. An example of dendrogram obtained from a simple system is shown in Figure 2.1.

Similarity measures

A fundamental building block of clustering analysis is the concept of (dis)similarity. Indeed, once the similarity pattern of a set of elements is obtained, in most cases the original data sample is not required anymore when applying the clustering algorithm. A similarity measure is a monotonic estimator of the extent to which two elements are alike. Thus, the more they are similar, the higher are the values taken by similarity.

Dissimilarity measures, instead, are monotonic estimators of the difference between pairs of elements. The term *proximity* is often used to refer to both similarity and dissimilarity. A similarity measure s must have the following properties:

- **Symmetry**
  $s(x, y) = s(y, x) \; \forall \; x, y$
- $s(x, y) = 1$ only if $x = y$

However, in most cases clustering algorithms are more efficient if a different class of proximity measures is used [45]. This class is refered to as *distances* and it is a subset of dissimilarity measures (although this notation is not always respected, as in some cases the term distance is used simply as a synonym for dissimilarity). The reason for which distance is preferable to similarity is the presence of an additional property that holds for the former, the so-called triangle inequality. Indeed, the properties that define a distance d are

- **Symmetry**
  $d(x, y) = d(y, x) \; \forall \; x, y$
- **Positivity**
  (a) $d(x, y) \geq 0 \; \forall \; x, y$
  (b) $d(x, y) = 0$ only if $x = y$
- **Triangle inequality**
  $d(x, z) \leq d(x, y) + d(y, z) \; \forall \; x, y, z$

A similarity measure can be easily converted to a distance through a suitable transformation rule.

Agglomerative algorithms

The set of agglomerative techniques is the most popular in hierarchical clustering. The general scheme of an agglomerative algorithm is summed up in Algorithm 1.

Thus, agglomerative algorithms require the introduction a criterion that regulates the computation of distances between clusters. Three of such criteria are the most used:

Algorithm 1 Agglomerative algorithms

```
Compute the distance matrix
All elements are assigned to different clusters
while More than a cluster exists do
    Merge the pair of clusters with minimum distance
    Update the distance matrix taking into account the new formed cluster
end while
```

- **Single linkage.** According to single linkage, the distance between two clusters is obtained by taking the minimum distance between all possible pairs belonging to the two clusters. This method has an analogy with graph theory. In fact, if all the data points are seen as vertices of a graph, applying single linkage is equivalent to inserting links, ordered according to increasing distance, until the system is all connected in a single component. The connected components that emerge during this procedure are the clusters detected at the level of distance of the last link inserted. Single linkage is quite sensitive to noise and outliers.

- **Complete linkage.** When using complete linkage, distance between clusters is computed as the maximum distance between all the pairs belonging to the two clusters. Looking at it through the perspective of graph theory, a cluster is highlighted by the complete linkage approach when all its elements are connected one with the other, i.e. when they form a clique. Complete linkage has the tendency to split larger clusters.

- **Average linkage.** The average linkage is an approach that lies between single and complete linkages. The distance between two clusters is computed as the average of all the distances between pairs that belong to the two clusters.

2.4.2 Minimum spanning tree

As shown in the previous section, hierarchical clustering provides the partitioning of a system into hierarchically nested clusters. Although the information contained in the associated dendrogram is highly informative on the structure of the system, there is no clear indication on its level of statistical significance. Indeed, the output of clustering algorithms can be affected by noise, whose origin in general can belong to a heterogeneous set of factors (mistakes in the transcription of data, random noise inherently linked to the observed process,...). This problem can be solved by cutting the dendrogram at a distance threshold at which the emerging partition is statistically significant. However, there is not a self consistent and universally accepted method to obtain this threshold, and several different methods are proposed [29], [46]. In this context, the problem can be seen from another perspective. In fact, if all information but the most essential is filtered out, the most unreliable connections between elements are dropped and what remains can be expected to be a robust, although minimal, description of the system. This is the approach adopted with minimum spanning trees (MST).

A *tree* is a fully connected graph that does not contain loops. The absence of loops implies that it is not possible to find paths that start and end in the same vertex without crossing any link twice. If visualized as graphs, the dendograms produced by hierarchical clustering algorithms are trees. Thus, trees have well defined topological constraints; in particular, a tree can be seen as the most compact way of connecting N vertices, since it has only $N-1$ links. With minimum spanning trees, this compactness is linked with its hierarchical structure. Indeed, there is a direct relationship between the construction procedure of the MST and the single linkage hierarchical clustering procedure. One possible implementation of MST is obtained by applying Prim's algorithm, which is described in Algorithm 2. The MST approach is an effective tool to detect the essential backbone of a set of multivariate variables, with a particular focus on financial systems, [47]. Indeed, when building a MST only the $N-1$ links associated with the lowest values of distance that maintain the graph as a tree are considered, discarding all the rest. The same formalism has been extended to different topological constraints [48], and bootstrap has been proposed as a tool to verify the reliability of links in MSTs [49].

2.4.3 Statistically validated networks

As seen in section 2.2.3, bipartite networks allow to represent systems whose structural functionality is linked to the interaction between two different types of elements. Their adoption is particolarly useful when the focus is on the detection of clusters of elements

Algorithm 2 Prim's algorithm

```
Initialize $V$, list of vertices
Initialize $C$, list of vertices priority
Assign INFINITY to all element of $C$
select a vertex $v$ randomly and set its priority to 0
while $V$ is not empty do
  Select $u$, vertex with minimum priority
  $V$ discard $u$
  for $p$ in $V$ do
    if distance $(p, u) < C[p]$ then
      $C[p]$=distance $(p, u)$
      p is chosen as neighbor of $u$
    end if
  end for
end while
```

which tend to be linked to the same elements of the other set. As an example, in biology the analysis of the relationships between genes and diseases can help in a better understanding of the latter. However, due to the large amount of interactions in these systems, these clusters rarely emerge clearly from the components of the projected network. Thus, a method capable to extract only the connections that share a statistically significant amount of information is required. However, this result cannot be achieved simply by fixing a threshold on the minimum number of shared neighbors. Indeed, since heterogeneity in the activity of each element is a characterizing feature for these systems, such a threshold could be too strict for elements with few interactions or too permissive for elements with many. A solution to this problem has been proposed with the formalism of statistically validated networks (SVN) [8]. The SVN approach is based on the formulation of a null hypothesis that properly takes into account the heterogeneity of the system. In fact, if the degree distribution of set B is homogeneous, one can model the random allocation of links between elements of set A and B as a series of random draws with replacement. Since the random draw problem is well known in statistics, the prediction of the corresponding null hypothesis can be expressed analytically. Indeed, for each pair (i, j) of vertices of set A, one can compute the probability of having X common neighbors after a draw with replacement through

$$H(X|N, d_i, d_j) = \frac{\binom{d_i}{X}\binom{N-d_i}{d_j-X}}{\binom{N}{d_j}}, \tag{2.2}$$

where N is the number of elements of set B, $d_{i(j)}$ is the degree of vertex $i(j)$ and H is the hypergeometric distribution. From this probability it is possible to compute a p-value by using the cumulative distribution,

$$p = 1 - \sum_{a=0}^{X-1} \frac{\binom{d_i}{a}\binom{N-d_i}{d_j-i}}{\binom{N}{d_j}}. \tag{2.3}$$

The p-value in Eq. 2.3 is the probability that i and j share X or more neighbors after a random draw. Thus, a small p-value indicates that i and j are having more co-occurrences than those expected by the null hypothesis. In this context, the adoption of a significance threshold α allows to highlight the links between the elements whose co-occurences not statistically significant. However, since the test is repeated for each link of the projected network, the significance threshold must be corrected for multiple comparisons, in order to keep the rate of false positives low. The strictest correction possible is the Bonferroni one [50], which assumes that all performed tests are independent. When applying this correction, the resulting threshold is

$$t_b = \alpha/N_t, \tag{2.4}$$

where N_t is the total number of tests. Despite its high precision, in most cases the Bonferroni correction is not statistically accurate, because it increases the number of false negatives. A less rigid correction is the control on the False Discovery Rate (FDR) [51]. With FDR, the rate of true positives is taken into account when fixing the threshold. Indeed, the threshold increases linearly with the number of rejected hypothesis. Since the starting point for the FDR correction is the Bonferroni threshold, the validated graph obtained through the former is always included in the latter. Although in the current context the statistical test of random draws has been presented following its application to bipartite networks, it can be applied also to detect the over-expression of attributes inside communities [52], or to study the evolution in time of a set of clusters

[53]. Moreover, when looking at the structure of communities of a projected networks, SVNs have been proven to be highly precise [54]. Indeed, although in some cases the accuracy is low due to the excessive severeness of the multiple hypothesis test correction, the detected communities are robust against noise and highly informative on the real structure of the system.

3 Long term ecology of investors in a financial market

Financial markets are essential economic institutions that shape our economic world, with stock markets being one of the most important ones. A stock market is an infrastructure where equities, bonds and other kinds of financial products are issued and traded among investors. A stock market can be considered as a paradigmatic example of complex systems. Indeed, its dynamics is the result of the actions of a large set of agents, that (i) belong to different classes of investors, (ii) may have different purposes and (iii) follow different strategies. This huge degree of heterogeneity affects the process of price formation, and makes practically impossible to study its dynamics by considering analytically all its components. This calls for an approach that exploits the knowledge of physics and statistics in order to spread light on its regularities and fundamental laws [43]. In this context, a frequently explored (but still open) problem is the characterization and classification of investors on the basis of their trading activity. This is also an interesting field with respect to the change of perspective that the world of economists has witnessed in the last decades. Indeed, it is now recognized that investors who differ with respect to their institutional role, biographical details, economic wealth and available information will adopt rather different trading strategies on the stock market. This observation, despite its apparent triviality, hides powerful implications, in a field which has been dominated for a long time by the Efficient Market Hypothesis (EMH) [55]. This theory, proposed for the first time by E.F. Fama in his groundbreaking PhD thesis, affirms that the price of an asset instantly reflects all the information available about its intrinsic value. This simple statement implies a key concept: the stock market works as a fair game. This means that according to EMH there is no place for arbitrage nor other ways of beating the market, and that all investors behave in a fully rational way. Although the effectiveness of this implication had been doubted even before its first enunciation [56], EMH has been partially overcome only quite recently, with the development of the Adaptive Market Hypothesis (AMH) [9], [55]. AMH recollects and organically integrates many of the arguments that have been opposed to EMH. It starts from Herbert Simon's observation [57] that the full rationality required by neoclassical economists is hardly reachable in reality. Indeed, real investors often adopt strategies which are typically only sub-optimal, and follow a bounded rationality, limited by the amount of information that they are able to consistently process and the interplay of emotivity and behavioral issues during trading decisions. Once the paradigm of a fully rational behavior is dropped, an evolutionary approach can be adopted in order to describe the dynamics of different strategies within different groups of investors [9]. Moreover, the survival or extinction of these strategies is linked to their success, i.e. the amount of profit that they are able to provide when adopted. In this context, econophysics can play an useful role in the characterization of investors behavior in a statistically significant way, in order to extract evidence of different strategies from the trading decisions adopted by all active agents. The present chapter of this dissertation goes in this direction, and it is organized as follows: i) Section 3.1 provides a short overview of the existing literature about the ecology of different trading strategies; ii) Section 3.2 provides a description of the features of the database used in this analysis; iii) Section 3.3 introduces to the formalism of categorical variables adopted in order to apply the approach of statistically validated network; iv) Section 3.4 describes the detection of clusters from the statistically validated networks of investors, and compare these clusters with those contained in the hierarchical trees obtained from the same system; v) Section 3.5 exploits the approach of SVNs to characterize the long term ecology of investors trading the Nokia stock at the Helsinki venue over a time span of 15 years.

3.1 Background

In this wide area, the literature about the investigation of investors behavior is rich and detailed. What follows should be considered only a partial review of some of these works, useful for the purposes of introducing the main subject of this chapter but without claims of exhaustiveness. It is now broadly accepted that trading strategies can be divided into two main groups, fundamentalist and chartist [58], which are characterized by the different philosophies that lie behind their trading decisions. Fundamentalist investors base their strategy on the analysis of the economic indicators (called *fundamentals*) of quoted companies, together with other macroeconomic estimators that refer to the condition of overall economy. Fundamentalist traders usually base their analysis on variables such as the capitalization, the number of employees or the production rate in order to understand whether the value of a company is being under or overestimated on the market, and act accordingly. On the other hand, a chartist investor does not try to estimate the value of the assets he/she is interested in trading, but relies exclusively on the recent behavior of their price series in the stock market. There are many different kinds of instruments that can be used to extract trading signals from the time series of prices, which go under the umbrella term of *technical analysis*. A further distinction can be made between contrarian and momentum investors, [59], [60], with both strategies being well represented among the tools of technical analysis. The two categories differ with respect to the timing of their trading decisions. Indeed, momentum traders buy assets whose price is growing and sell those whose price is selling, while contrarian act the other way around in an attempt to beat the market. Going further, a crucial role in shaping the strategy of investors is played by available information, making the distinction between informed and uninformed investors [61] particularly meaningful. Other relevant works include the analysis of Barber et al., [62], on the Taiwanese stock market. The authors found out that household investors, due to their aggressive attitude when placing orders, obtain significant losses from their trading decisions, while institutional investors have significantly higher profit. The same database presented in this chapter has also been investigated by Grinblatt and Keloharju in a series of studies [63, 64] on the trading characteristics of individual and institutional investors, and on behavioral aspects of individual trading. Lastly, Tumminello et al. in [65] investigated the same dataset adopting the approach of statistically validated network to obtain clusters of investors with similar trading activity. This formalism will be extended in this chapter, bringing to the detection of different collective strategies shared by large sets of investors.

3.2 Database

The database used for this work is maintained by the Euroclear Finland (previously Nordic Central Securities Depository Finland) which is the central register of shareholdings for Finnish stocks and financial assets in the Finnish Central Securities Depository (FCSD). The register contains the shareholdings in stocks of all Finnish investors and of all foreign investors asking to exercise their vote right, both retail and institutional. The database records official ownership of companies and financial assets on a daily basis according to the Finnish Book Entry System. The records include transactions, i.e. operations that are executed in stock exchanges and change the ownership of the assets. The database has associated a certain amount of metadata; specifically, it classifies investors into six main categories: a) non-financial corporations, b) financial and insurance corporations, c) general governmental organizations, d) non-profit institutions, e) households, and f) foreign organizations. The database is collected since January 1st, 1995.

The database covers all the stocks traded at the Helsinki venue of the Nordic Stock Exchange. For legal reasons, the database treats Finnish domestic investors (or foreign investors asking to exercise their vote right) in a different way from foreign investors. In fact, while the database contains very detailed information about the Finnish domestic investors, foreign investors can choose to use nominee registration. In this last case, the investor's book entry account provider, for example a bank, aggregates all the transactions from all of its accounts. This implies that a single nominee register coded identity contains the holdings of several foreign investors. [1]

[1] If an institution can trade both for itself and also on behalf of nominee registered investors, in the current analysis its trading activity is split in two distinct IDs, one regarding its activity as a Finnish investor and one regarding its activity on behalf of nominee registered investors (labeled as NR).

3.3 Nominal variables characterizing the trading activity

The high degree of heterogeneity that characterizes the activity of investors calls for methods of analysis that are robust with respect to this aspect. This necessity provides the motivation to analyze the trading activity of individual investors in terms of categorical variables. Specifically, the adopted nominal variables have been introduced in Ref. [65]. These variables are defined as follows: for each investor i, each stock k, and each trading day t, the volume sold $V_s(i,k,t)$ and the volume purchased $V_b(i,k,t)$ of stock k by the investor at day t are computed. For each stock and for each day, this information is then converted into a nominal variable with 3 states: primarily buying b, primarily selling s, buying and selling bs such that all positions will be essentially closed before the market closes. The nominal variables are obtained by considering the ratio

$$r(i,k,t) = \frac{V_b(i,k,t) - V_s(i,k,t)}{V_b(i,k,t) + V_s(i,k,t)}. \qquad (3.1)$$

For each stock k, an investor is assigned a primarily buying state b when $r(i,k,t) > \theta$, a primarily selling state s when $r(i,k,t) < -\theta$, and a buying and selling state bs when $-\theta \leq r(i,k,t) \leq \theta$ with $V_b(i,k,t) > 0$ and $V_s(i,k,t) > 0$. Throughout this chapter the threshold value is fixed to $\theta = 0.01$.

3.4 Trading profiles of investors

The first step of the current analysis concerns the detection of clusters of investors on the basis of their trading profiles. Specifically, the investigation is performed on the investment decisions related to 23 of the 25 stocks that compose the OMXH25 market index in 2003[2]. The total number of investors is $105,005$, and the trading activity of different investors is highly heterogeneous. The information about the complete set of investors is summarized in Table 3.1.

In the following analyses done with similarity measures, a threshold was put on the minimum activity of investors. The motivation for this choice is the need to be able to estimate a similarity measure which is minimizing the discretization role associated

[2] The original plan was to investigate all the 25 stocks that were used to compute the index but finding the time history of stock price for two of them was impossible to us.

with the presence of a very limited number of attributes. Specifically, in the similarity based tests only the investors who have traded one of the OMXH25 stocks at least 5 times during 2003 are considered. The summary statistics of these investors is given in Table 3.2

3.4.1 Clustering of trading profiles by correlation

The degree of similarity in the trading profile of investors is evaluated by constructing for each selected investor and for each stock of the OMXH25 a vector of trading actions. This vector has a number of components equal to three times the number of trading days of 2003 (which is 253x3=759). The first 253 components carry the information whether the investor was buying (b state) on a specific day, the second 253 components carry the information whether the investor was selling (s state) on a specific day, and the third 253 components carry the information whether the investor was buying, selling and closing the position (bs state) on a specific day. The final vector is therefore a binary vector of 1s and 0s. To investigate these vectors a similarity measure which is robust with respect to the asymmetric presence of the two attributes (1s are rarely present whereas 0s are highly observed in most of the cases) is needed. Thus, the Jaccard coefficient is used as similarity measure because it is known to be robust for asymmetric binary vectors [66]. The Jaccard coefficient is defined as the ratio between the size of the intersection on the size of the union of two sets. In the case of two binary vectors i and j, this definition leads to

$$J(i,j) = \frac{M_{11}}{M_{11} + M_{01} + M_{10}}, \tag{3.2}$$

where M_{ab} is the number of components in which the first vector has value a and the other has value b.

Fig. 3.1 show the hierarchical tree obtained with the average linkage algorithm with a dissimilarity measure defined as $d_{i,j} = \sqrt{2(1 - J_{i,j})}$. Nokia has 7824 investors trading at least five times during 2003 and therefore the complete visualization of a hierarchical tree of so many elements is not simple in a limited space. Fig. 3.2 plots a region of the hierarchical tree involving 1419 investors. This level of details allows to appreciate the hierarchical structure of the similarity of investment profiles.

A similar behavior is observed also for the other stocks. Thus, the use of categorical

Table 3.1: Summary statistics of the number of investors making at least one transaction for each of the 23 of the OMXH25 stocks during 2003. The different stocks are labeled with their tick symbol. The single investors are grouped in six category: Corporations (non financial corporations), Financial, Financial NR (i.e. those financial institutions that are trading for nominee registered not using their right to vote), Foreign entities, Governmental organizations, Households, and Non-profit organizations. The last row provides the total number of distinct Finnish investors trading the corresponding stock. The last column provides the total number of investors trading at least one stock of the OMXH25. The total of the column is not the sum of each category row because the same investor can trade different stocks.

Category	AMEAS	ELIAV	ELQAV	FUM1V	HUH1V	KCI1V	KESBV	KONBS	MEO1V	NDA1V	NOK1V	NOR1V
Corporations	624	1161	684	816	737	339	812	766	810	1082	3216	403
Financial	107	121	70	133	116	82	95	102	133	136	221	80
Financial NR	18	18	16	17	18	16	15	18	18	16	35	15
Foreign Org.	39	56	36	68	57	21	32	62	49	97	422	19
Governmental	60	52	38	59	54	42	42	53	65	65	88	43
Households	3439	14742	5201	10372	5045	1652	7345	6018	6808	13374	37140	2996
Non-profit	132	108	55	182	173	67	196	207	167	207	270	85
Total	4419	16258	6100	11647	6200	2219	8537	7226	8050	14977	41392	3641
Category	OUT1V	POH1V	POS1V	RTRKS	SAMAS	STERV	TIE1V	TLS1V	UNR1V	UPM1V	WRT1V	Total
Corporations	1426	528	701	528	924	977	886	1216	323	1874	672	6110
Financial	128	58	110	112	120	118	133	116	82	181	82	300
Financial NR	18	16	16	18	18	19	18	20	16	20	17	22
Foreign Org.	85	32	28	38	63	55	50	113	34	164	41	772
Governmental	67	31	47	51	62	76	58	58	48	87	35	114
Households	11831	5843	5004	3821	6600	8301	6290	16003	1635	21033	7285	96900
Non-profit	237	64	55	80	159	152	121	96	92	303	104	787
Total	13792	6572	5961	4648	7946	9698	7556	17622	2230	23662	8236	105005

Table 3.2: Summary statistics of the number of investors making at least five transactions for each of the 23 of the OMXH25 stocks during 2003.The different stocks are labeled with their tick symbol provided at the top of the column. The single investors are grouped in six category: Corporations (non financial corporations), Financial, Financial NR (i.e. those financial institutions that are trading for nominee registered not using their right to vote), Foreign entities, Governmental organizations, Households, and Non-profit organizations. The last row provides the total number of distinct Finnish investors trading the corresponding stock. The last column provides the total number of investors trading at least one stock of the OMXH25. The total of the column is not the sum of each category row because the same investor can trade different stocks.

Category	AMEAS	ELIAV	ELQAV	FUM1V	HUH1V	KCI1V	KESBV	KONBS	MEO1V	NDA1V	NOK1V	NOR1V
Corporations	44	83	98	65	67	29	58	81	105	144	1041	40
Financial	41	57	40	61	43	33	32	53	54	54	117	44
Financial NR	14	13	12	14	13	12	12	15	15	13	33	10
Foreign Org.	6	5	9	8	5	5	2	7	8	6	68	6
Governmental	11	21	8	23	12	13	9	11	25	31	60	15
Households	71	194	391	214	110	44	173	227	442	468	6477	110
Non-profit	11	12	13	9	10	7	6	14	17	17	54	13
Total	198	385	571	394	260	143	292	408	666	733	7850	238
Category	OUT1V	POH1V	POS1V	RTRKS	SAMAS	STERV	TIE1V	TLS1V	UNR1V	UPM1V	WRT1V	Total
Corporations	146	58	70	50	93	198	116	123	25	347	50	1310
Financial	51	24	47	49	54	65	55	46	29	80	32	150
Financial NR	16	13	12	13	14	15	14	14	12	15	13	20
Foreign Org.	6	5	6	5	7	11	8	13	4	14	6	85
Governmental	31	9	12	21	27	34	18	23	7	43	7	69
Households	416	186	352	91	340	781	454	671	23	1271	167	8320
Non-profit	20	8	14	10	18	24	14	16	5	29	4	78
Total	686	303	513	239	553	1128	679	906	105	1799	279	10032

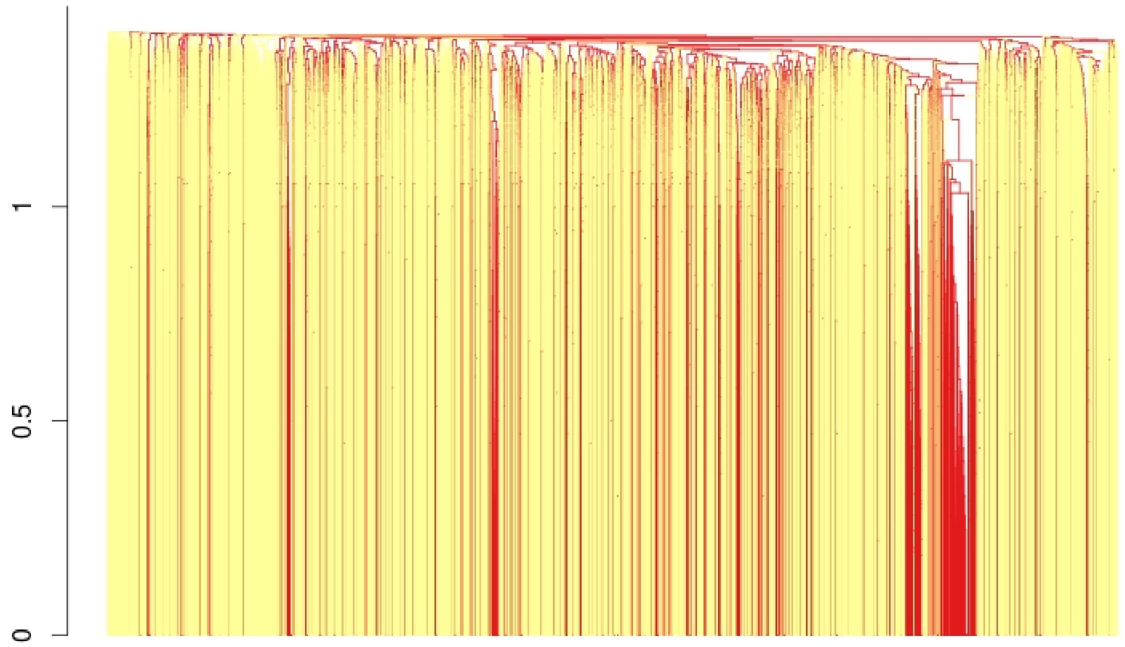


Figure 3.1: Average linkage hierarchical tree of the trading profile similarity of investors trading Nokia in 2003. The selected investors have performed at least 5 transactions of the Nokia stock during 2003. The presence of large clusters comprising hundreds of investors is clearly detected (see the region of the hierarchical tree highlighted in red). The investors labeled with red lines belong to clusters obtained by setting a cutting threshold equal to $d = 1.09$ (see Sect. 3.4.3 for details).

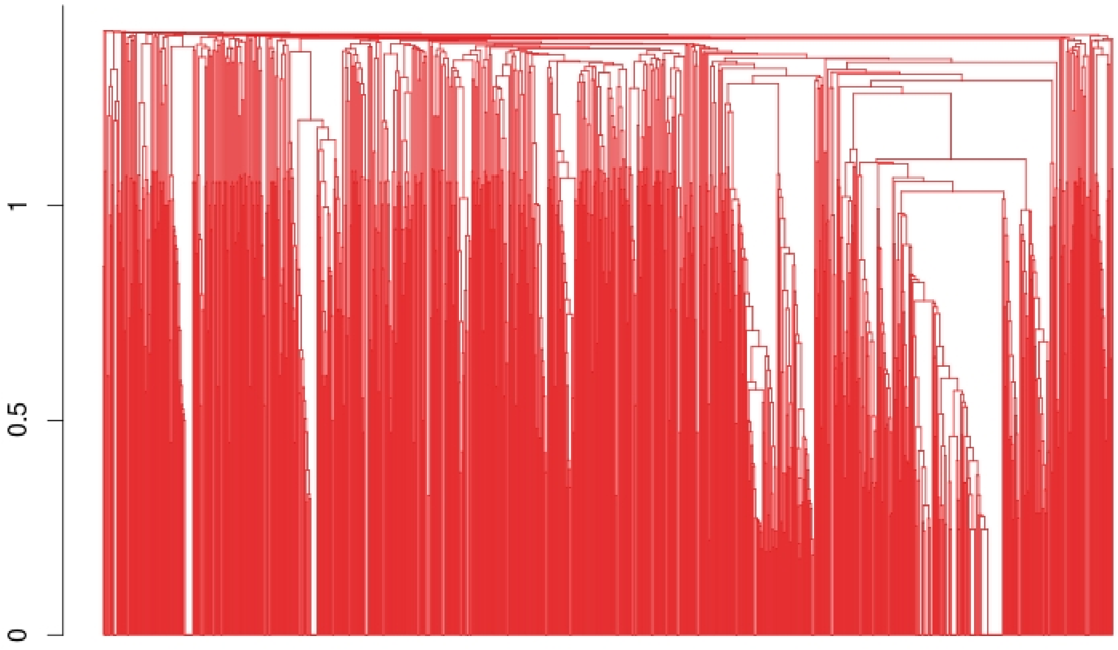


Figure 3.2: Regions of the average linkage hierarchical tree of Fig. 3.1 that are selected by using the threshold $d = 1.09$ (see Sect. 3.4.3 for details). All remaining investors are removed from the present tree. The two big clusters seen in the right part of the figures are the two clusters highlighted in red in the right part of Fig. 3.1.

variables together with the choice of the Jaccard coefficient allow to detect the hierarchical clustering structure of the trading profile of the different investors.

3.4.2 Over-expressed trading profiles

In a previous study [65] a different approach based on statistically validated networks [8] was used to detect clusters of investors characterized by a similar trading profile when trading the Nokia stock over a time period longer than 5 years. In this chapter the idea is to compare the clusters obtained through the approach of Ref. [65] with the information obtained from the hierarchical tree of the same system.

The method of the statistical validation of the co-occurrence of categorical variables is the same as in Ref. [65] and works as follows. The bipartite network used as a starting point for the statistical validation is made by a set of investors and a set of days, with a link put between investor i and day t if i was active on t. When testing the null hypothesis of random co-occurrence of activity presented in Section 2.4.3, the size of the urn from which investors i and j draw their days of activity is not chosen equal to the total number of days N, but to the length of the intersection of the corresponding activity periods, N_T. This choice has been made in order to make the null hypothesis more appropriate to the trading dynamics. Indeed, one of the aspects of the high heterogeneity that characterizes this system is that the activity of investors is often sharply localized only in defined time windows which are a subset of the trading year. Thus, using N_T instead of N fits in a better way the description of this system. Moreover, since the hypergeometric distribution for the same value X of co-occurrence assumes higher values if the size of the urn N is larger, this choice makes the test more severe. Thus, if N_A (N_B) is the number of days when investor i (j) is in the state A (B) and $N_{A,B}$ is the number of days in which there is a co-occurrence of state A for investor i and state B for investor j the p-value for the pair of investors in the states A and B is

$$p(N_{A,B}) = 1 - \sum_{X=0}^{N_{A,B}-1} H(X|N_T, N_A, N_B). \tag{3.3}$$

The nine possible combinations of the three trading states between investor i and j are (i_b,j_b), (i_b,j_s), (i_b,j_{bs}), (i_s,j_b), (i_s,j_s), (i_s,j_{bs}), (i_{bs},j_b), (i_{bs},j_s) and (i_{bs},j_{bs}).

As presented in Section 2.4.3, both Bonferroni and the control of False Discovery Rate have been used as multiple test corrections. In this context, the Bonferroni correction for each stock k is fixed to $0.01 * 2/(9N_k(N_k - 1))$, where N_k is the total number of investors active for at least 5 different days on the stock k.

Table 3.3 reports the number of investors characterized by at least one validated co-occurrence for each investigated stock when adopting the Bonferroni correction. The largest number of investors characterized by co-occurrence in the trading profile is detected for the Nokia stock. This is not surprising because the Nokia stock was in 2003 the most traded (as it can be verified in Table 3.1) and liquid stock of the OMXH25 index. The results can also depend on the power of the statistical test that decreases when the number of tested hypotheses increases when adopting the Bonferroni correction, as it will be investigated more in detail in Section 3.5.1.

Table 3.4 instead reports the results for the multiple hypothesis test correction based on the FDR correction. As expected the number of investors showing statistically validated co-occurrences is increasing and the accuracy of the test is improved although the level of precision might slightly decrease. The sets of investors with statistically validated co-occurrences of Table 3.3 are of course always included in the corresponding entry of Table 3.4. Fig. 3.3 shows the statistically validated network of Nokia investors obtained with the Bonferroni correction. The network consists of 576 investors, the majority of whom are Households although also investors of the other categories are present. Specifically, in the network there are 142 Non financial corporations (Violet node), 18 Financial and 1 Financial NR (Grey node), 5 Foreign organizations (Yellow node), 16 Governmental (Black node), 378 Households (Blue node) and 16 Non-profit (Red node) investors. Direct inspection of the network shows that there is a large connected component of 346 investors and that all validations of co-occurrences concern i_b,j_b (Blue link), i_s,j_s (Red link), and the simultaneous validation of i_b,j_b and i_s,j_s (Black link).

As summarized in Tables 3.3 and 3.4 the largest detected statistically validated networks are those that contain Nokia investors. For this stock the Bonferroni (i.e. the statistically validated networks obtained with the Bonferroni correction) and FDR (i.e. the statistically validated networks obtained with the Benjamini-Hochberg correction) networks have 576 and 1518 nodes respectively. In the case of the other OMXH25 stocks

Table 3.3: Summary statistics of the number of investors having at least one statistical validation of co-occurrence for each of the 23 of the OMXH25 stocks during 2003.The multiple hypothesis test correction is the Bonferroni correction. The label of stocks and investors' categories is as in Table 3.1. The last row provides the total number of distinct Finnish investors having co-occurrences of some daily trading status (b, s, and bs). The last column provides the total number of investors having co-occurrences in at least one stock of the OMXH25. The total of the column is not the sum of each category row because the same investor can have co-occurrences for different stocks.

Category	AMEAS	ELIAV	ELQAV	FUM1V	HUH1V	KCI1V	KESBV	KONBS	MEO1V	NDA1V	NOK1V	NOR1V
Corporations	0	1	3	2	0	0	0	1	1	1	142	3
Financial	5	2	3	8	4	0	2	3	9	5	18	3
Financial NR	1	0	1	1	0	0	0	1	1	2	1	0
Foreign Org.	1	0	3	1	0	0	0	1	1	1	5	0
Governmental	0	1	0	1	0	0	0	0	4	1	16	8
Households	2	0	0	2	0	0	0	6	0	2	378	1
Non-profit	0	0	0	0	0	0	0	0	0	0	16	2
Total	9	4	10	15	4	0	2	12	16	12	576	17

Category	OUT1V	POH1V	POS1V	RTRKS	SAMAS	STERV	TIE1V	TLS1V	UNR1V	UPM1V	WRT1V	Total
Corporations	2	0	0	2	1	8	1	1	1	11	0	154
Financial	4	0	2	5	4	12	3	4	1	10	0	31
Financial NR	0	1	0	0	0	0	0	1	0	0	0	3
Foreign Org.	0	1	0	0	0	0	0	1	0	0	0	6
Governmental	4	3	0	6	2	12	0	1	0	10	0	18
Households	2	2	0	0	2	6	3	0	0	4	2	389
Non-profit	3	3	0	1	0	5	0	2	0	4	0	16
Total	15	10	2	14	9	43	7	10	2	39	2	617

Table 3.4: Summary statistics of the number of investors having at least one statistical validation of co-occurrence for each of the 23 of the OMXH25 stocks during 2003.The multiple hypothesis test correction is the FDR correction. The label of stocks and investors' categories is as in Table 3.1. The last row provides the total number of distinct Finnish investors having co-occurrences of some daily trading status (b, s, and bs). The last column provides the total number of investors having co-occurrences in at least one stock of the OMXH25. The total of the column is not the sum of each category row because the same investor can have co-occurrences for different stocks.

Category	AMEAS	ELIAV	ELQAV	FUM1V	HUH1V	KCI1V	KESBV	KONBS	MEO1V	NDA1V	NOK1V	NOR1V
Corporations	0	1	3	3	2	0	0	1	2	1	298	4
Financial	5	4	3	8	5	0	2	3	10	5	40	6
Financial NR	1	0	1	1	1	0	0	1	1	2	2	1
Foreign Org.	1	0	3	1	1	0	0	1	1	1	9	1
Governmental	0	1	0	2	0	0	0	0	5	3	39	9
Households	2	0	2	4	0	0	0	6	0	5	1098	3
Non-profit	0	0	0	0	0	0	0	0	0	0	32	5
Total	9	6	12	19	9	0	2	12	19	17	1518	29

Category	OUT1V	POH1V	POS1V	RTRKS	SAMAS	STERV	TIE1V	TLS1V	UNR1V	UPM1V	WRT1V	Total
Corporations	5	0	0	5	3	11	3	5	1	14	0	310
Financial	7	0	2	7	5	15	3	4	1	13	0	48
Financial NR	0	1	0	0	0	0	0	1	0	0	0	4
Foreign Org.	2	1	0	0	0	0	0	1	0	3	0	10
Governmental	8	5	0	7	6	17	2	1	0	12	0	41
Households	13	2	0	0	2	16	6	1	0	13	2	1116
Non-profit	5	3	0	3	0	8	0	3	0	6	0	34
Total	40	12	2	22	16	67	14	16	2	61	2	1563

the Bonferroni and FDR networks are always populated but with a much smaller number of nodes. In fact the second largest statistically validated network is the one of Stora Enso comprising 67 investors, obtained with the FDR correction (see Table 3.4).

This last statistically validated network is shown in Fig. 3.4. In this case the majority of investors are Governmental organizations, with the different categories represented in the following way: 11 Non financial corporations (Violet node), 15 Financial (Grey node), 17 Governmental (Black node), 16 Households (Blue node) and 8 Non-profit (Red node) investors. As in the case of the previous example, most of the validations of co-occurrences concern i_b,j_b (Blue link), i_s,j_s (Red link), and the simultaneous validation of i_b,j_b and i_s,j_s (Black link). The largest connected component has only 21 nodes, so it does not cover the majority of nodes.

Most of the remaining OMXH25 stocks have statistically validated networks of the type observed for Stora stock, i.e. a network of small disjoint clusters. As in Ref. [65] the clusters of investors with similar trading profile were obtained by applying a widely used community detection algorithm, the Infomap one [32], to the weighted version of the SVNs. The weight of each link is the number of co-occurrences validated between the two investors (for example in the case when the co-occurrences i_b,j_b and i_s,j_s are validated, the weight of the link is set to two). In most cases the components observed in the networks are not further partitioned by the algorithm. However, in the case of presence of a highly populated large connected component (as for Nokia) the large component is partitioned into smaller clusters.

By using the detected partitions it is possible to visualize the trading patterns of investors belonging to the different clusters. Fig. 3.5 shows the trading profile of Nokia and Stora investors associated with the clusters of the SVNs shown in Fig. 3.3 and 3.4. In the figure a red spot indicates a buy action, a green spot a sell action, and a white spot a buy/sell action, while a black spot represents absence of trading for the specific investor and trading day. Fig. 3.5 clearly shows the presence of different trading profiles among the different clusters, while each cluster is rather homogeneous.

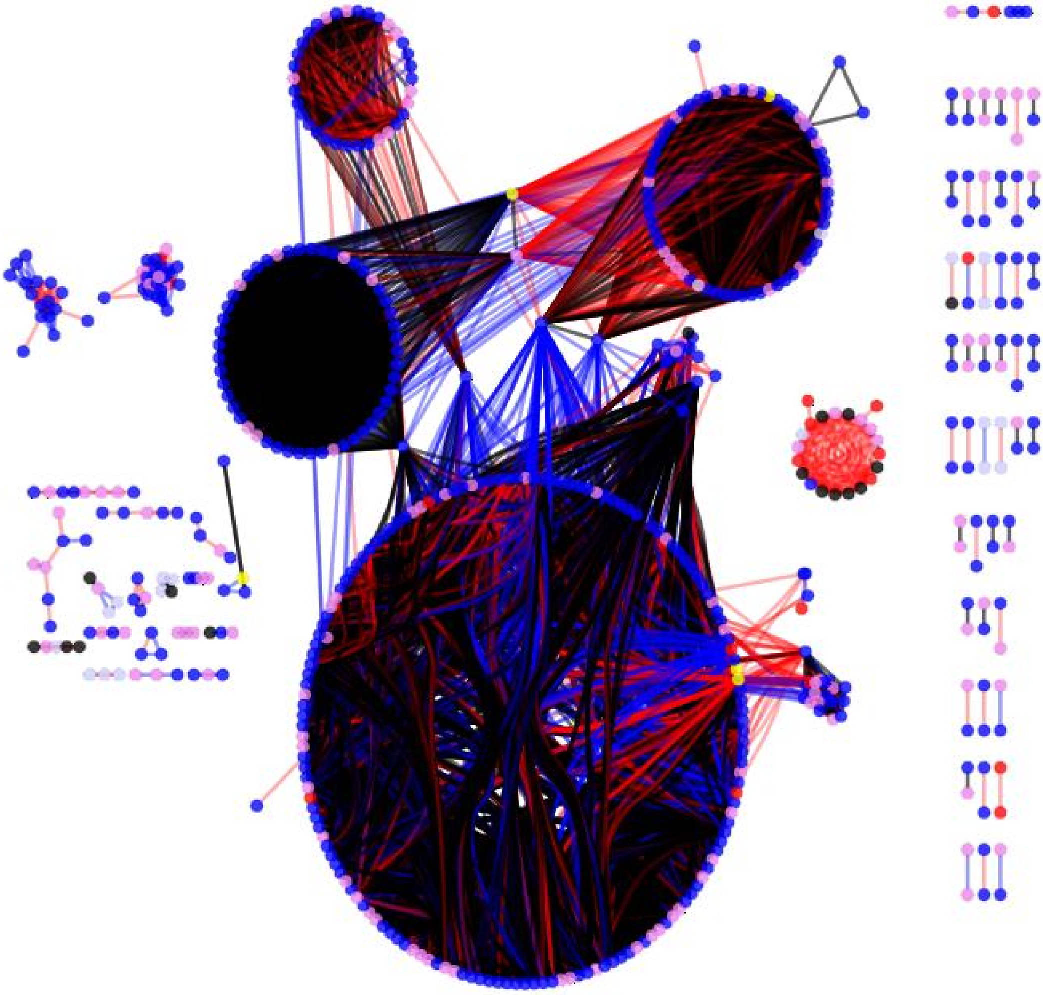


Figure 3.3: Network of 576 individual investors having statistically validated co-occurrences of trading decisions about the Nokia stock in 2003. The multiple hypothesis test correction is the Bonferroni correction. The category of investor is labeled as follows: Corporations (Violet), Financial (Grey), Foreign organizations (Yellow), Governmental (Black), Households (Blue) and Non-profit (Red). The type of statistically validated co-occurrence is labeled as follows: i_b,j_b (1075 Blue links), i_s,j_s (2468 Red links), and the simultaneous validation of i_b,j_b and i_s,j_s (10353 Black links).

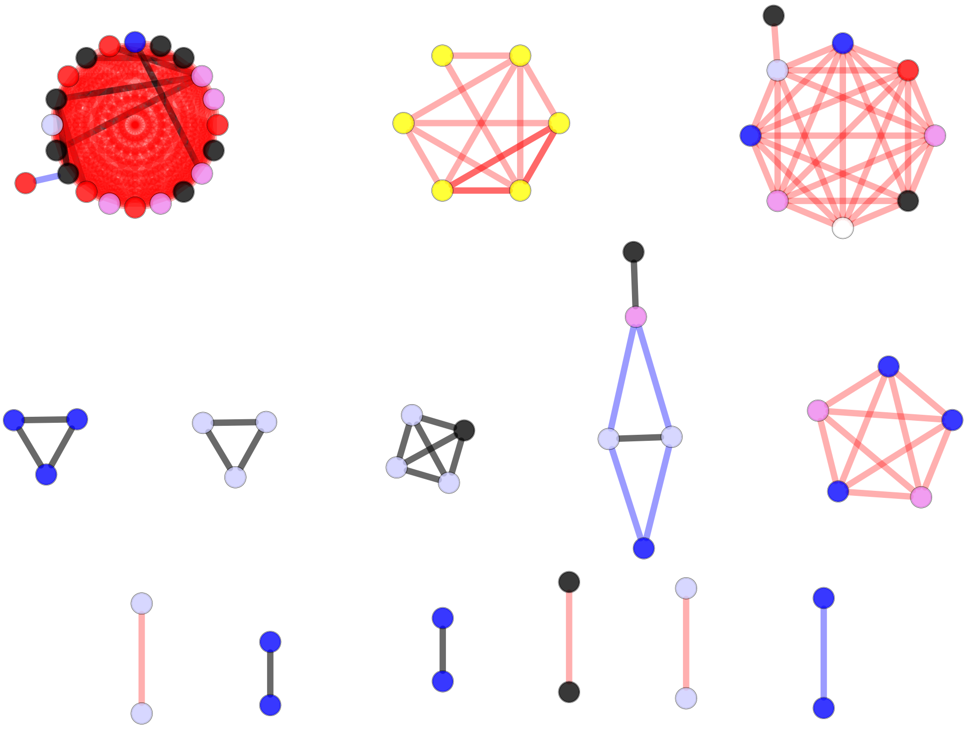


Figure 3.4: Network of 67 individual investors having statistically validated co-occurrences of trading decisions about the Stora Enso stock in 2003. The multiple hypothesis test correction is the FDR correction. Links are coded as follows: i_b,j_b (6 Blue links), i_s,j_s (210 Red links), and the simultaneous validation of i_b,j_b and i_s,j_s (26 Black links)

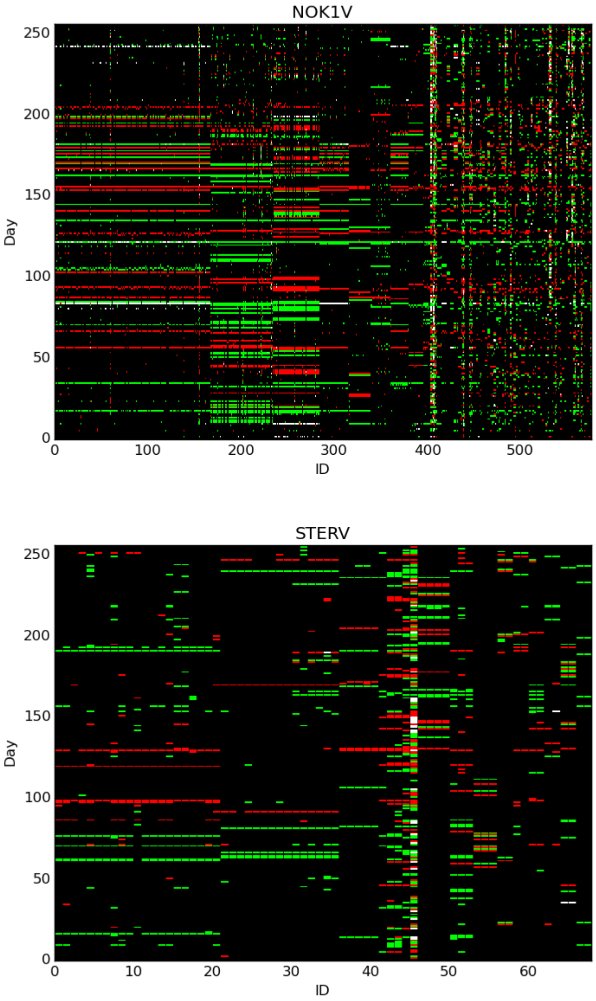


Figure 3.5: Color code representation of the trading profile of investors trading the Nokia (left) and the Stora (right) stocks. In the horizontal axis we order different investors whereas the vertical axis is time (in number of trading days). The left panel shows trading profiles of the 576 Nokia investors whose trading co-occurrences were statistically validated with the Bonferroni correction whereas the right panel shows to the 67 Stora investors validated with the FDR correction. A red spot indicates a buy action, a green spot a sell action and a white spot a buy/sell action. Black spots indicate absence of trading. The investors are ordered according to the membership of the clusters detected with Infomap algorithm.

3.4.3 Comparison of partition methods

In the previous sections it has been shown that both a correlation analysis and a statistical validation procedure provide information about (i) the clusters of different investors (in the case of the statistical validation approach jointly used with the community detection procedure), and (ii) about the hierarchical structure of the trading profiles. This section is dedicated to investigating whether the two types of information overlap or are rather providing complementary types of information; in case of overlap, the aim is to estimate its extent.

One difficulty in the comparison of the two sets of information is that whereas a partition is provided in the case of statistically validated networks, the hierarchical clustering just provides a hierarchical structure that needs to be processed to obtain a partition. A basic way to obtain a partition from a hierarchical tree is to cut it at a given value of the dissimilarity (or similarity) measure. The corresponding clusters are the groups of elements that are connected at distances lower that the cutting threshold. The method is simple and effective but its drawback is that there is no simple and widely accepted optimal way to select the threshold level.

For this reason the choice of the distance threshold has been guided by the detection of the degree of maximal overlapping between the partition of the SVN and the ones obtained from hierarchical trees. Thus, first a hierarchical clustering procedure is selected and the hierarchical tree of the 7824 investors active on Nokia stock in 2003 is obtained. In particular, single, average and complete linkage [45] have been applied. For each hierarchical tree, a partition is obtained by fixing a distance threshold and extracting clusters. The single cluster is made up of all the investors whose relative distance is less then the threshold. The investors linked at higher distances are considered as isolated nodes. The obtained hierarchical tree partition is then shrunk to the 576 investors that are also present in the Bonferroni network. At this point the two partitions of 576 investors can be compared by computing the Adjusted Rand Index (ARI) [67], a measure which is widely used as an estimator of the similarity of two partitions. It is a normalized indicator and assumes a value equal to one when the two partitions are identical and a value close to zero when the two partitions are randomly assigned. Fig. 3.6 shows the Adjusted Rand Index between the partition obtained through the SVN

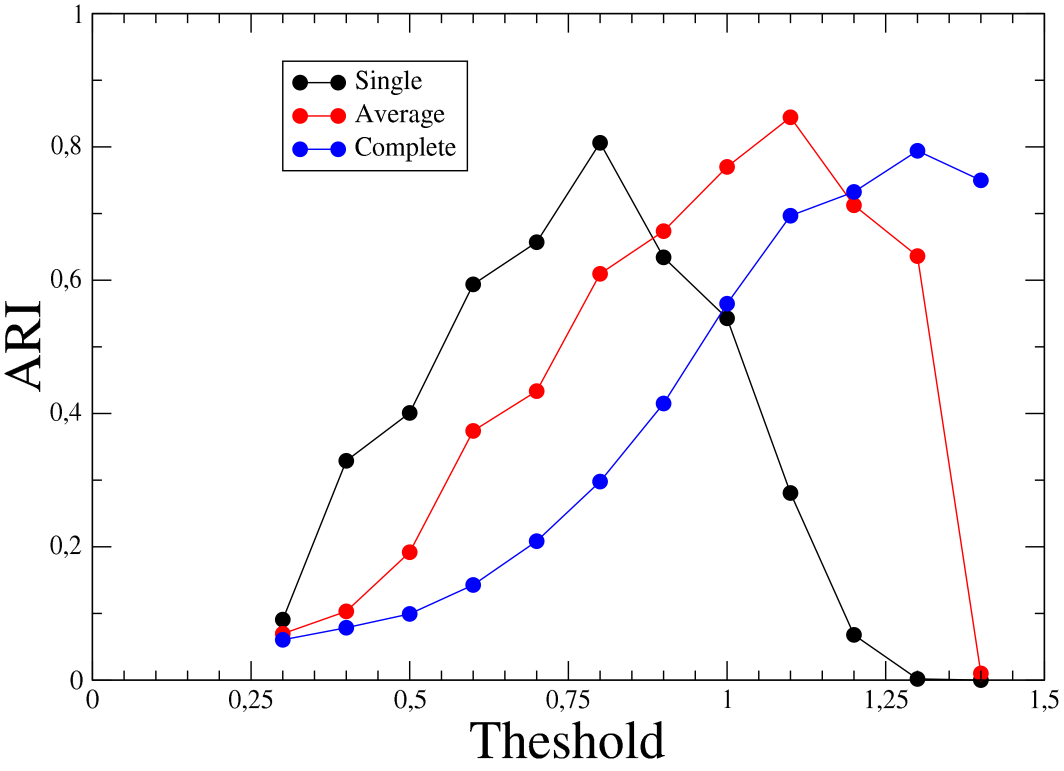


Figure 3.6: Adjusted Rand Index (vertical axis) between the partition of Nokia investors obtained from the Bonferroni statistically validated network partitioned by the Infomap algorithm with the partition of the same investors obtained by performing hierarchical clustering algorithms on the dissimilarity measure of Nokia investors at different values of the cutting threshold (horizontal axis). The comparison is performed for the single (black circles), average (red circles) and complete linkage (blue circles).

approach (Bonferroni correction) and the partition obtained from hierarchical clustering algorithms at different values of the cutting threshold (horizontal axis). The comparison is performed for the single (black circles), average (red circles), and complete linkage (blue circles). In all three cases a bell shaped curve of the ARI as a function of the threshold is observed. Therefore in all three cases a single maximum exists for a specific value of the threshold. Although the figure shows the values of the ARI computed by varying the threshold in steps of 0.1, in the proximity of the maximum the calculations have been made by increasing the threshold by steps of 0.01. With this resolution the highest value of the ARI are observed for threshold distances equal to 0.80, 1.09 and 1.31 for the single, average and complete hierarchical tree respectively, with the highest values being 0.806, 0.923, and 0.795. These are quite high values and therefore for the optimal threshold values the partitions of the hierarchical trees are pretty similar to the partitions obtained by applying Infomap on the Bonferroni SVN. The highest similarity is observed for the average linkage hierarchical clustering.

To provide a visual comparison of the similarity of the two partitions, Fig. 3.7 shows the color code representation of the trading profile of Nokia investors grouped both as defined by the Bonferroni SVN and as defined by the average clustering hierarchical tree when a threshold obtained by maximizing the ARI is used. In the specific case the threshold value is set to 1.09 and the ARI is 0.923. The figure shows the high overlap quantified by the high value of the ARI index. It also shows that the hierarchical tree approach extends the information available to a larger number of investors. In fact when a threshold of value 1.09 is used 1419 Nokia investors are detected in 319 clusters of at least 2 investors. The number of investors of the hierarchical tree partition is significantly larger than the number of 576 Nokia investors observed by the Bonferroni statistically validated network. There is empirical evidence that the highest values of ARI are typically observed for the average clustering. The use of the single linkage is increasing the probability of observing large clusters whereas the complete linkage selects clusters or relatively smaller size. The average linkage provide an intermediate behavior.

Fig. 3.8 shows the probability density function of the size (in number of investors) of clusters observed for the Bonferroni partition, and for the three partitions obtained

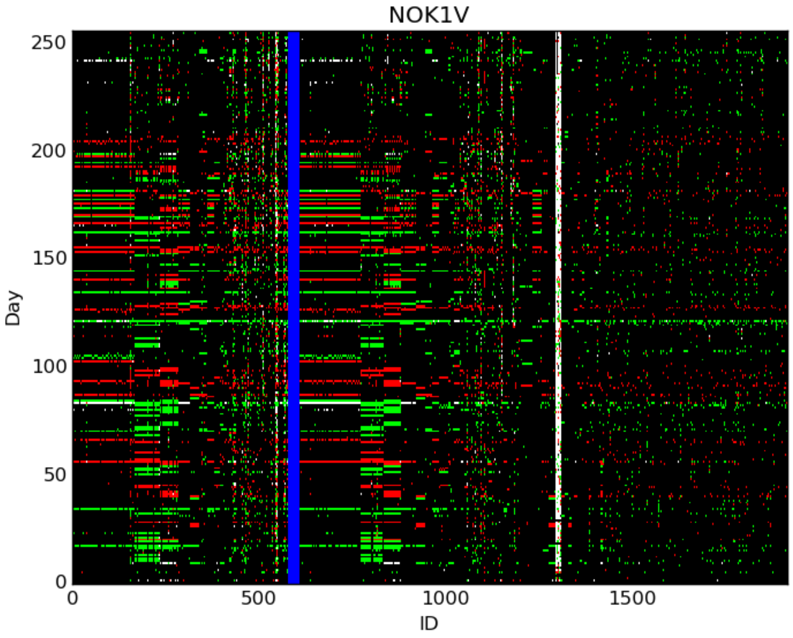


Figure 3.7: Color code representation of the trading profile of investors trading the Nokia. The left part of the figure (limited by a blu vertical bar) shows trading profiles of the clusters of 576 Nokia investors in the Bonferroni statistically validated network (same plot as in the top panel of Fig. 3.5) while the right part are the clusters obtained from the hierarchical tree of the average clustering by using the 1.09 threshold. The overlapping clusters are ordered with the ordering of Fig. 3.6.

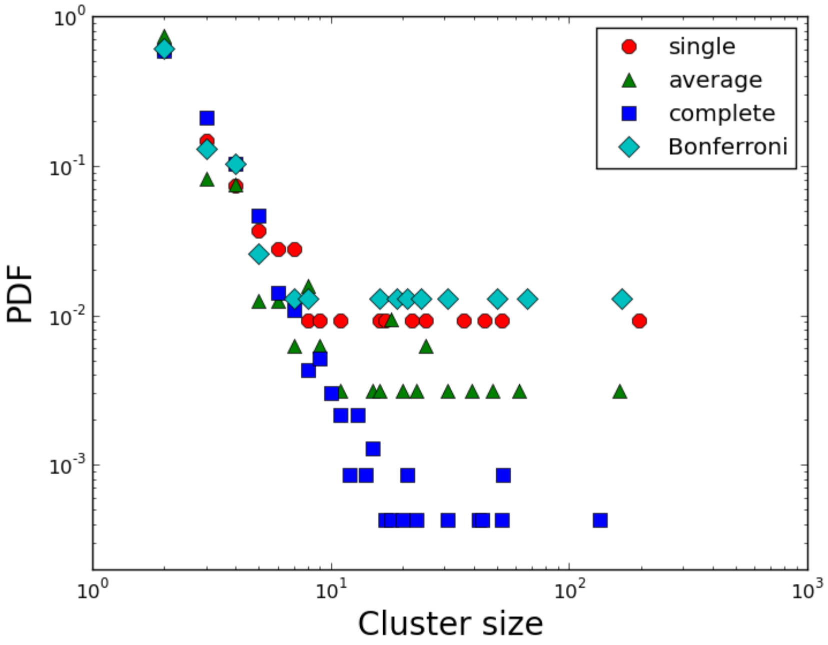


Figure 3.8: Probability density function of the cluster size of the partitions obtained with the Bonferroni statistically validated network partitioned by the Infomap algorithm (light blue diamond), single linkage (red circle), average linkage (green triangle), and complete linkage (blue square). The stock is Nokia.

with the hierarchical clustering algorithms. All the three hierarchical clustering methods reproduce well the behavior observed for the Bonferroni partition. The number of partitions and the number of investors present in the partitions are higher for the complete linkage, intermediate for the average linkage and lower for the single linkage. In the case of the complete linkage partition, the broader covering of investors is probably obtained at a cost of a lower precision of the partitioning. In the present system and in the present example the average linkage provides the best choice for partitioning a number of investors larger than the Bonferroni partition which is maintaining approximately the same precision as the Bonferroni approach.

The observation that the Bonferroni SVN is providing highly precise but not necessarily accurate clusters is not surprising. In fact it is known that the Bonferroni correction is too restrictive. This choice ensures a high level of precision (because the number of false positive is minimal) but on the other hand it might be associated with a high number of false negatives and therefore be characterized by a relatively low level of statistical accuracy. The cluster detection based on the correlation measure is therefore providing results which might be a bit less precise but more accurate. The problem with the correlation approach is that the thresholding process is not supported by any theoretical indication.

The observation of a significant overlap of the clusters obtained with the two distinct approaches suggests the effectiveness of a new methodology using SVNs partitioned with an efficient community detection algorithm together with a hierarchical clustering procedure. Within this approach the optimal threshold to be used in the hierarchical clustering procedure can be determined by using the precise information obtained with the Bonferroni approach. Finally, when the threshold is determined, one obtains the partition from the hierarchical tree.

3.5 Time evolution of clusters

The second part of our analysis is devoted to the study of the dynamics of investors active on the Helsinki venue of the Nordic Stock Exchange (NSE) over a time span of 15 years, from 1995 to 2009. For this purpose, the set of investors has been limited to those active on the stock Nokia, which in the previous section was proven to be the most liquid stock of the Helsinki venue. The starting point of the analysis is the set of clusters detected through the approach of SVNs year by year. The aim of this investigation is to map the ecology of investment profiles over the years, in order to detect whether different sets of investors adopting different strategies coexist in the NSE over long time spans. Such an evidence would represent a significant empirical confirmation of the claims of the adaptive market hypothesis.

Table 3.5 reports the number of Nokia investors making at least one transaction during the reported calendar year. The information is reported for all investors and category by category, while Table 3.6 reports the same information for the investors active in at least five different days. In addition, Table 3.6 reports also the number of links of the projected network of investors. A link is present between two investors when they are active in a period which is overlapping for at least one day, even if they do not have days of co-occurrence in trading decisions. The density d of the edges in the network of investors is given by $d = \frac{E}{n(n-1)/2}$, where E is the number of links and n is the number of vertices. The average density of edges for the network of investors is ranging from a minimum value 0.042 observed in 2006 to a maximum value 0.090 observed in 2003.

In what follows the dynamics of investors is analyzed starting from the clusters obtained from the FDR statistically validated networks. The motivation for this choice is related to the power of the statistical test used to obtain the SVNs and it is explained in the next section.

3.5.1 Power of test

Table 3.6 shows that the sizes of the projected networks of investors differ sharply within the considered period. This heterogeneity calls for a test on the power of the method of statistical validation. In fact, when comparing results obtained from networks of different sizes, one should verify that the power of the statistical test is not affected by the number of tested links. One first proof that this is not the case is given by the comparison between the ratios of validated links on total links for the different years. This ratio is shown by the orange lines of figure 3.9, both for the Bonferroni and the FDR correction. Although 2002 is the year in which the highest number of links is observed, the highest value of this ratio occurs in 2005. In order to track this information in a more rigorous way a suitable test on the power of the statistical validation has been designed. The test works as follows: year by year the statistical validation is applied on samples of links that are drawn from the projected network of investors. When drawing the samples the proportion of links between investors of the different categories is maintained. Figure 3.9 plots the ratios of validated links on total links both for the samples and the whole system. The left panel is related to the Bonferroni correction, while the right panel shows the results for the FDR correction. The samples contain 1,000,000 links. For each year, the sampling procedure was repeated 10 times. The figure shows that, when using the FDR correction, the power of the test remains constant on systems of different size. This suggests that the difference in the size of SVNs observed at different years are due to the underlying dynamics of financial markets and investors activity and not to different power of the methodology. However, this is not true when using the Bonferroni correction. In this case the larger the system is, the less powerful the test performs.

Table 3.5: Summary statistics of the number of Nokia investors of different categories making at least one transaction during the reported calendar year.

Year	All	Households	Financial	Govern.	NR	Foreign	Non profit	Companies
1995	11085	9531	95	39	18	95	188	1119
1996	9152	7553	115	42	32	86	155	1169
1997	14709	12644	151	55	29	142	196	1492
1998	23649	20778	182	68	32	298	249	2042
1999	46313	42199	231	75	27	389	504	2888
2000	87127	80394	290	111	35	702	696	4899
2001	64422	58876	250	83	45	602	374	4192
2002	53644	48697	213	86	46	524	292	3786
2003	41392	37140	221	88	35	422	270	3216
2004	56497	51509	216	84	33	461	374	3820
2005	42734	38551	194	83	37	360	363	3146
2006	39161	35493	164	61	36	333	368	2706
2007	44233	40398	180	59	37	398	407	2754
2008	48553	44551	167	51	34	399	258	3093
2009	69258	64172	179	54	28	364	317	4144

Table 3.6: Summary statistics of the number of Nokia investors of different categories making at least five transactions during the reported calendar year.

Year	All	Households	Financial	Govern.	NR	Foreign	Non profit	Companies	Edges
1995	764	390	59	19	18	9	22	247	1748672
1996	988	558	67	28	31	8	17	279	2899094
1997	1359	818	80	29	25	18	27	362	5403872
1998	2810	2047	104	42	25	29	41	522	23620445
1999	5292	4252	138	53	21	43	86	699	79849727
2000	9408	7824	157	58	30	98	97	1144	257931314
2001	9472	7903	140	64	33	91	69	1172	269916005
2002	9651	8106	120	59	30	83	67	1186	279274631
2003	7850	6477	117	60	33	68	54	1041	196930874
2004	7064	5836	111	41	28	54	53	941	148407134
2005	4733	3807	91	37	33	44	51	670	67900316
2006	3697	3014	84	23	31	36	40	469	41858993
2007	3546	2889	91	29	31	38	50	418	36950396
2008	6173	5282	94	27	30	59	48	633	104552144
2009	9348	8313	68	20	24	61	38	824	253553951

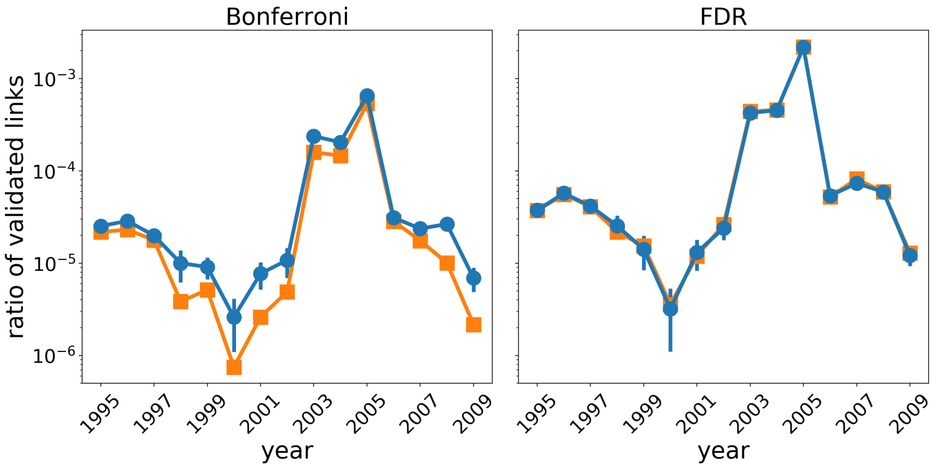


Figure 3.9: The left panel plots the ratios of validated links in the whole systems (orange line) and the average value of the same ratio on 10 samples of size 1,000,000 (blue line) with the Bonferroni correction. The error bars of the blue line plots the standard deviation on the set of 10 samples. The right panel plots the corresponding figure when using the control of the False Discover Rate as a correction for multiple tests.

This is a known issue for family wise error correction methods such as the Bonferroni one [68]. It is worth noting that the patterns shown in Fig. 3.9 are obtained by fixing the Bonferroni threshold b as $b = 0.01/N_l$, with N_l the total number of links reported in Table 3.6. The quantity N_l considers all the possible combination of trading states to be tested. This choice was motivated by the observation that, when considering all the couples of investors active in a calendar year, some of them are active in time periods that do not overlap. Thus, testing these couples is not meaningful and they are discarded when computing the Bonferroni threshold. Since this method of setting the Bonferroni threshold is slightly different from the one adopted in section 3.4.2, the results of Table 3.4 do not match with the outcome of this validation, which is reported in Table 3.7. Specifically, since the old method takes into account all possible couples of investors active in a year ($b = \frac{0.01*2}{9N_k(N_k-1)}$, with N_k the number of investors active at least 5 days) the corresponding threshold is more severe, producing a SVN with about 40% the nodes of the SVN described in Table 3.7.

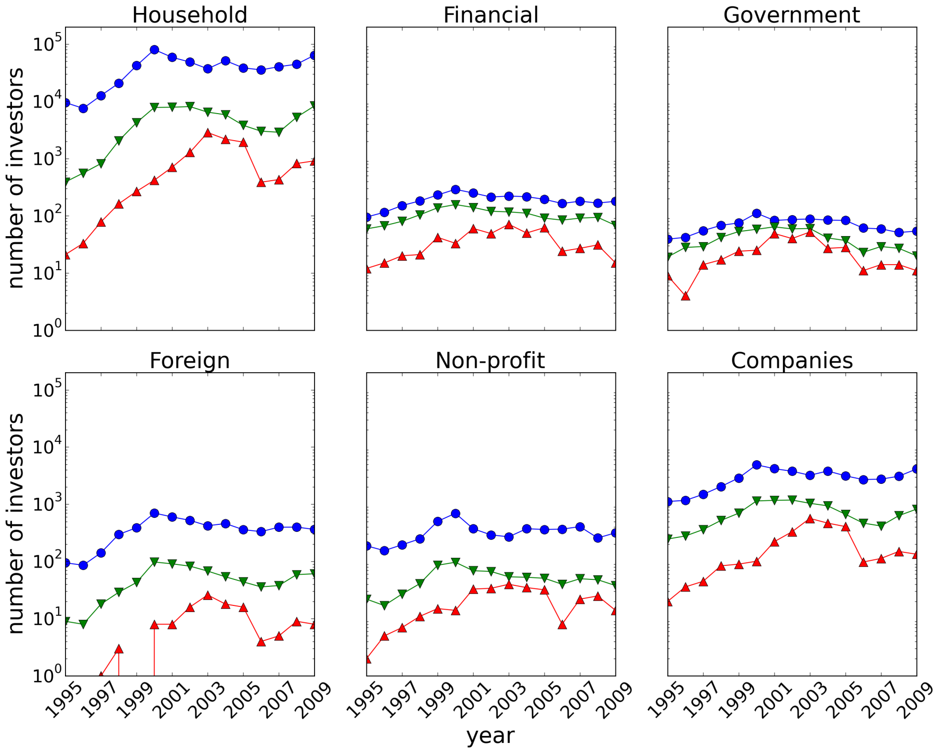


Figure 3.10: Number of investors (blue circles), number of active investors (green triangles down), number of investors included in the FDR network (red triangles up) as a function of the different calendar year. Each panel refers to a category of investors. In the top row we have households, financial institutions, and governmental organizations (from left to right), whereas in the bottom row we have foreign organizations, non-profit organizations, and companies (from left to right).

3.5.2 Statistically validated networks of investors over a 15 years time interval

In order to discriminate among the behavior of different categories of investors, Fig. 3.10 shows the number of investors active at least one day, the number of those active at least five days and the number of those that have at least one link in the FDR SVN as a function of the calendar year for each of the six categories of the database. The figure shows that households are controlling the unconditional statistics of the number of investors present in statistically validate networks. Moreover, non financial corporations show an overall profile which is similar to the one observed for households whereas the remaining categories show a much less time pronounced pattern.

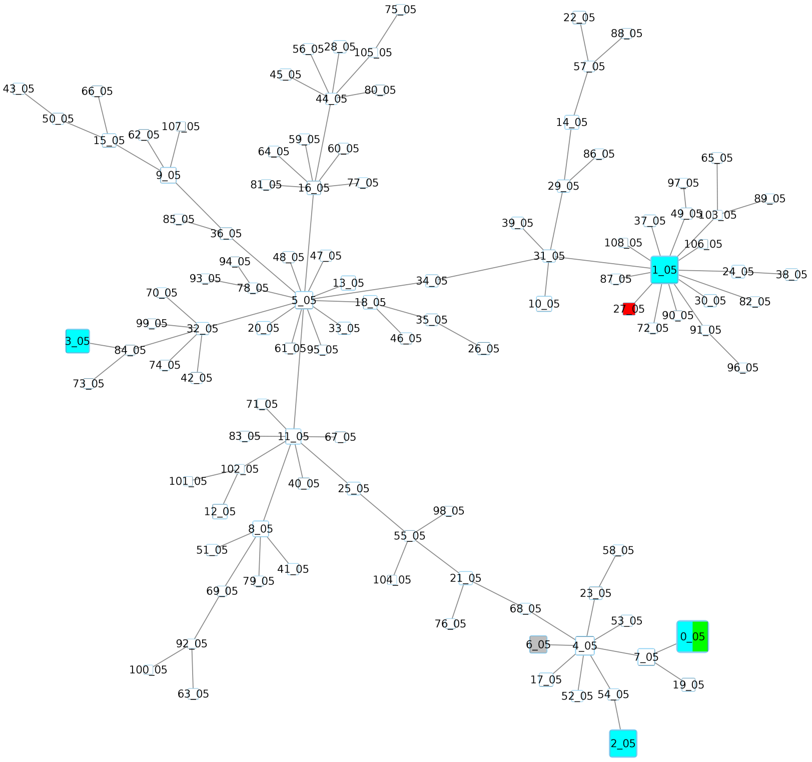


Figure 3.11: Minimum spanning tree of the similarity matrix associated with the trading activity of clusters for 2005. Each cluster is labeled by a number. Symbols in colors indicate the over-expression of one or more category of investors.

Table 3.7 reports a summary statistics of the clusters of investors detected in the FDR networks, obtained with the approach described in Section 3.4.2. The number of clusters and their size (in number of investors) is varying over time. The size of the clusters of investors observed is ranging from the minimum value of 2 to the maximal value of 425 (observed in 2005). Clusters of size bigger than 100 are observed during the period from 2002 to 2005 and in 2008.

For each cluster of investors the overall buying, selling, and buy-selling activity for each trading day of the year can be computed. With this approach a vector of approximately 750 records for each cluster is obtained (each trading day contributes to three different records, i.e. one for buying, one for selling and one for buying-selling). The similarity between each pair of clusters is evaluated by estimating the Pearson's correlation coefficient between the activity vectors of the two clusters. A simple and efficient way

Table 3.7: Summary statistics of the clusters detected in the FDR statistically validated networks with the Infomap algorithm during the reported calendar year. The column FDR nodes gives the number of investors present in the FDR network. The column Clusters gives the number of clusters detected by Infomap. The column Clusters 5 gives the number of clusters with a number of nodes higher than five. The Size biggest column gives the number of investors present in the biggest cluster. The smallest clusters have always size two.

Year	FDR nodes	Clusters	Clusters 5	Size biggest
1995	66	28	0	4
1996	100	33	3	12
1997	174	60	5	12
1998	301	81	12	26
1999	444	115	17	31
2000	602	172	23	22
2001	1082	282	34	39
2002	1760	333	75	163
2003	3618	509	186	309
2004	2803	419	121	216
2005	2505	313	109	425
2006	542	123	23	31
2007	622	136	30	54
2008	1053	206	46	101
2009	1361	277	57	92

to highlight the main similarities present between the investment activity of different clusters is through the minimum spanning tree (MST) associated with the correlation coefficient matrix of all clusters of a given calendar year [47]. Fig. 3.11 shows the MST of clusters of 2005. It has also been investigated whether each cluster presents the over-expression of the number of investors of a given category. The statistical test used to perform this kind of analysis is described in Ref. [52]. It should be noted that the result of this statistical test is not simply pointing out the categories of investors with maximal number of investors in a given cluster. In fact, the test detects whether a given category is significantly over-represented in a cluster with respect to a null hypothesis that takes into account the heterogeneous size of the different categories. When the over-expression is detected (always by taking into account the correction for multiple tests), the symbol of the cluster is labeled with a given color. For example the cluster 0_05 has an over-expression of households (cyan color) and of non financial corporations (green color). Other clusters showing over-expression of some category of investors are clusters 1_05 (households over-expression), 3_05 (households over-expression), 6_05 (governmental over-expression, label in grey color), and 27_05 (financial over-expression, label in red color). The MST shows that the investment activity of some clusters is quite dissimilar. For example clusters 0_05, 1_05, and 3_05 are located in distinct branches of the MST suggesting a high degree of dissimilarity among them (the first number of the cluster label is an arbitrary numeric label and the second number are the last two digits of the calendar year), although they are all over expressed in the same category, i.e. households. Fig. 3.12 shows the trading profile of Nokia investors belonging to the six clusters with over-expression of investors' categories detected in 2005. The horizontal axis orders distinct investors of each cluster whereas the vertical axis is time (in number of trading days). In the figure a red spot indicates a buy action, a green spot a sell action and a white spot a buy/sell action, while a black spot indicates absence of trading for the specific investor and trading day. Visual inspection of Fig. 3.12 suggests that the trading strategy of the different clusters is rather different under many aspects. The most evident ones concern the frequency of trading, the number of investors, and the specific sequence of buy, sell, and buy-sell trading decisions. For example, clusters 0_05 and 2_05 are characterized by a low frequency of trading. On the contrary, 27_05 and

1_05 are characterized by a high frequency of trading whereas 3_05 and 6_05 show an intermediate level. Similarity of the profile is often related to synchronous buying (red lines) and selling (green lines) decision. However also this synchronicity has a characterizing role for some clusters. Examples are clusters 27_05 and 1_05. Up to now, it was not possible to associate a specific trading strategy to a each cluster of investor. However in the following sections each cluster will be characterized with a number of indicators concerning the profile of trading and some attributes that characterize the investors.

3.5.3 Dynamics of statistically validated networks of investors

The investors' composition and investment profile of clusters are changing year after year. In order to put in relation investors of a cluster of a given year with investors of clusters of the successive year a suitable statistical test has been used. This test looks at the over-representation of the number of investors that are present in both clusters against a null hypothesis that takes into account the heterogeneity of the size of the clusters. The test was performed as described in Ref. [18]. Several pairs of clusters $(m.l)$ detected in consecutive years $(k, k+1)$ present over-expressed intersections of investors. When this is the case cluster m_k is connected with an arc to cluster l_{k+1}. Fig. 3.13 shows the time evolution of several clusters of FDR networks. The time duration of the observed cluster evolutions ranges from a minimum of 2 years to a maximum of 12 years (see the cluster evolution starting from 7_98 and ending at 34_09). The coalescence of several clusters is also observed (see, for example, the coalescence of clusters 2_02, 3_02, 6_02, 22_02 and 34_02 into the cluster 0_03 in the middle of the figure), together with the splitting of a cluster in two clusters (e.g. the splitting of 1_05 into 1_06 and 5_06). For several clusters, their dynamics presents regularities with respect to the type of investors composing them. Fig. 3.13 shows also those clusters that present an over-expression of the number of investors of a given category (or categories). It is evident that several chains of clusters show a persistent over-expression of specific categories of investors. The most prominent example is the cluster evolution starting from 7_98 and ending at 34_09. In this chain all clusters are over-expressed in governmental organizations (nodes with grey color) with the additional over-expression of non-profit institutions (nodes with yellow color) in some years. It is possible also to observe chains of clusters characterized

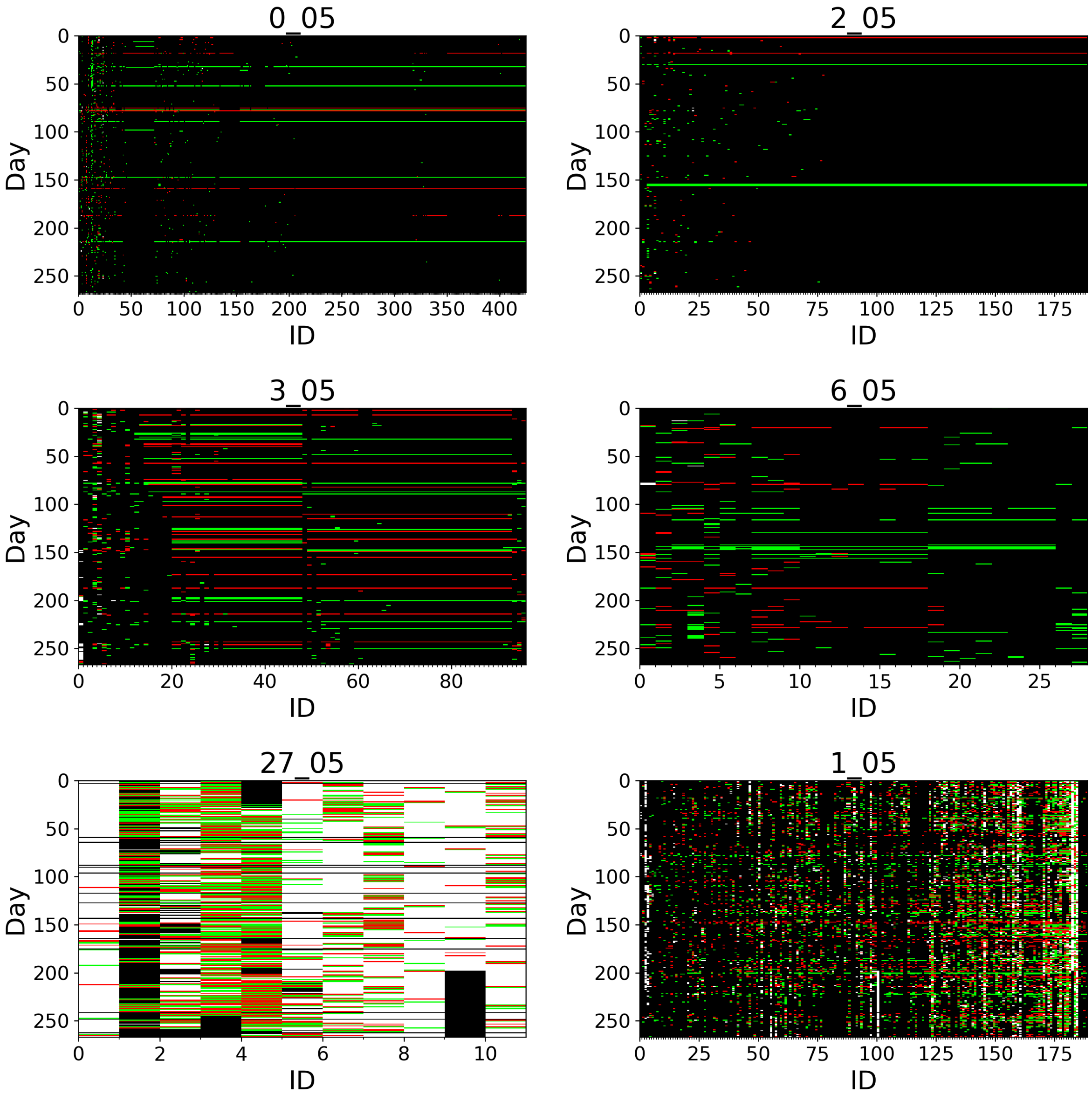


Figure 3.12: Color code representation of the trading action of Nokia investors for clusters 0_05 (top left panel), 2_05 (top right panel), 3_05 (middle left panel), 6_05 (middle right panel),27_05 (bottom left panel), and 1_05 (bottom right panel). The horizontal axis orders different investors whereas the vertical axis is time (in number of trading days from top to bottom). A red spot indicates a buy action, a green spot a sell action and a white spot a buy/sell action. Black spots indicate absence of trading.

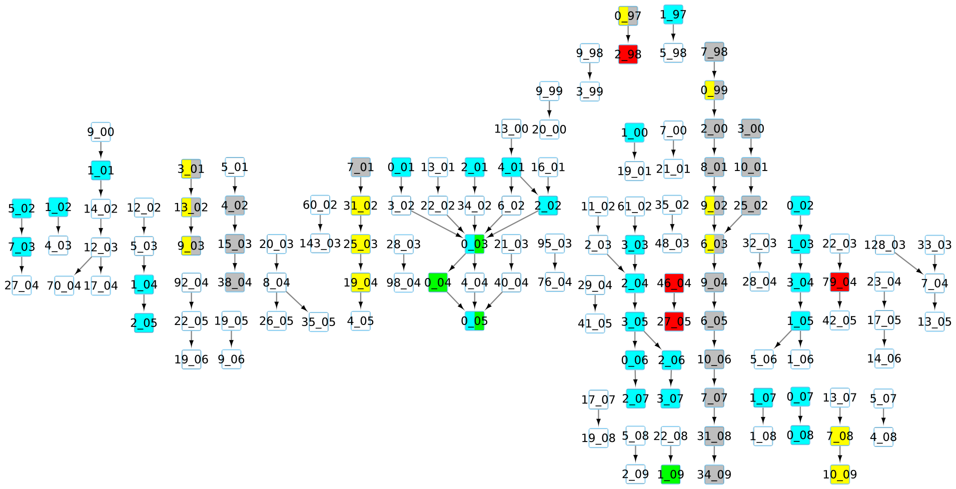


Figure 3.13: Time evolution of the clusters detected in the FDR networks. Clusters are represented by a node labeled with a numerical index and the year of the FDR network. The size of the node is proportional to the logarithm of the number of nodes of the cluster. A link is set between two nodes when the overlap between the number of nodes presents in a community at year i with the number of nodes presents at year $i+1$ is over-expressed with respect to a null hypothesis of random partitioning maintaining the heterogeneity of cluster size observed for the considered years. Paths of cluster evolution lasting several years (up to 12 years) are observed. Splitting and merging of clusters are also observed. Colored nodes are nodes characterized by an over-expression of one or more categories of investors. Colors refer to the different categories as follows: a) non-financial corporations (green), b) financial and insurance corporations (red), c) general governmental organizations (grey), d) non-profit institutions (yellow), e) households (cyan), and f) foreign organizations (brown).

by over-expression of households (see, for example, the chain from 11_02 and 61_02 to 2_07 and 3_07 and the chain from 0_02 to 5_06 and 1_06), households and non financial corporations (see the chain from 13_00 to 0_05), financial corporations (starting from 22_03 and ending at 42_05 and from 46_04 to 27_05), non-profit institutions (see the chain from 13_07 to 10_09), or some other combinations of the different categories.

3.5.4 Long-term ecology of clusters

For each year, four different attributes were investigated for each individual cluster (when the chain is involving just a single cluster per year) or group of clusters (when many clusters are part of a chain in a year, see for example clusters 8_01, 10_01 of the chain from 7_98 to 34_09). The four considered attributes are (i) the average pairwise distance between vectors of individual trading decisions of investors $d(i,j)$ belonging to a cluster or to a group of clusters. The distance between investor i and j is measured

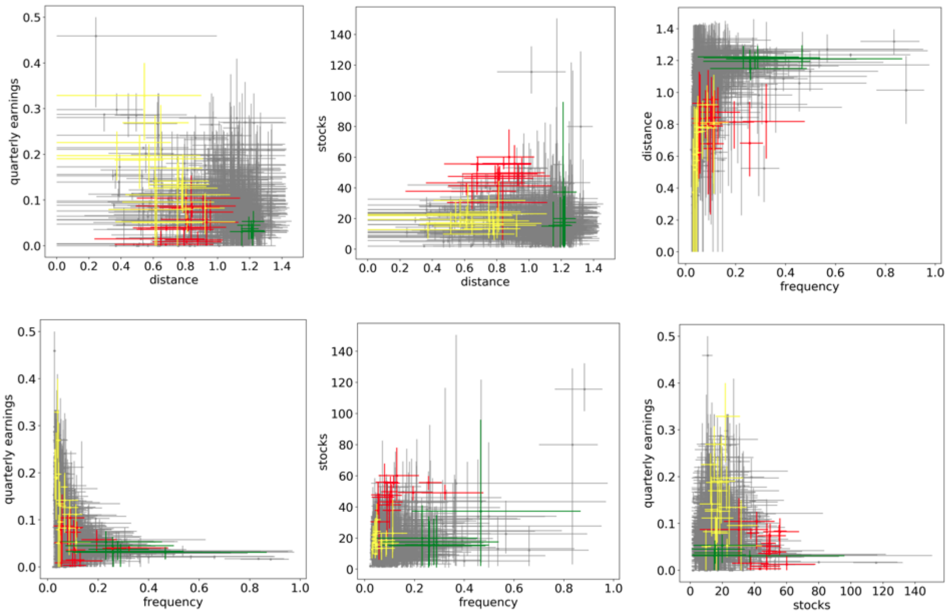


Figure 3.14: Scatter plots of six attributes characterizing a cluster or a group of cluster for each year. Each segment covers the values from the first decile to the last one. Crosses shown in color refer to the cluster chains from 7_98 to 34_09 (red crosses), from 13_00 to 0_05 (yellow crosses), and from 0_02 to 1_06 and 5_06 (green crosses)

first as a Jaccard similarity $\rho_J(i,j)$ and then transformed to a distance according to $d(i,j) = \sqrt{2(1-\rho_J(i,j))}$, (ii) the average value of the ratio of the number of trading during quantitative earning days divided by the total number of trading days, (iii) the average value of the number of stocks each investor of the cluster (or group of cluster) is investing on, and (iv) the average trading frequency of investors of the cluster (or group of cluster). A trading frequency equals to one indicates trading activity for all trading days of the year.

The average distance gives information about the degree of dissimilarity observed between the activity of pairs of investors of a cluster. The average value of the rate of quantitative earning trading days provides information concerning the relevance of these special days in the trading decisions. A high value highlights attention to fundamental news and/or trading decisions associated with market trading days typically characterized by over-reaction of the market. The average number of stocks owned by investors is a proxy of their knowledge about basic financial concepts as the one of investment diversification. The average trading frequency gives information about the average number of trading days of investors during the year. Fig. 3.14 shows six scatter plots of the average values of the 6 pairs of the above indicators. In addition to the average value observed for each cluster (or group of clusters) for each year, two segments indicate the interval from the first decile to the last one. All points and segments are provided in grey with the exception of three groups of clusters referring to three specific chains that are provided with crosses drawn in color. The six panels of the figure show that the chain from 7_98 to 34_09 (red crosses), which is characterized by over-expression of governmental and non-profit institutions, presents attention to diversification (high value of the average number of stocks), low average frequency of trading (with two years of exception when an intermediate frequency of trading was adopted), moderate trading involvement during days of quarterly earnings, and relatively homogeneous trading among investors, as testified by a low value of the average distance between the vectors of trading activities. The points associated with this chain of clusters are quite distinct from the other two selected chains. In fact the chain of clusters from 13_00 to 0_05 (yellow crosses), which is characterized by over-expression of households and non-financial corporations, presents less attention than the previous one to diversification (the average number of stocks is around 15), very low average frequency of trading, high trading involvement during days of quarterly earnings, and relatively low average distance between investment decisions. The third chain of clusters from 0_02 to 1_06 and 5_06 (green crosses), although also characterized by over-expression of households, is characterized by attributes that are quite different from the ones of the previous chain. Specifically, investors of this third chain present moderate attention to diversification, relatively high average frequency of trading, low trading involvement during days of quarterly earnings, and a high average distance between investment decisions (i.e. the trading decisions are rather heterogeneous in this case). Thus, this analysis underlines the presence of investment profiles that are typical of groups of investors, are different one from the other and are present in the market with a time scale of many years or even decades. These types of strategies are typically over-expressed among investors belonging to a specific category or to few categories.

In order to quantify the average similarity of investors belonging to each category, the average Jaccard correlation between the binary vectors of trading activity concerning

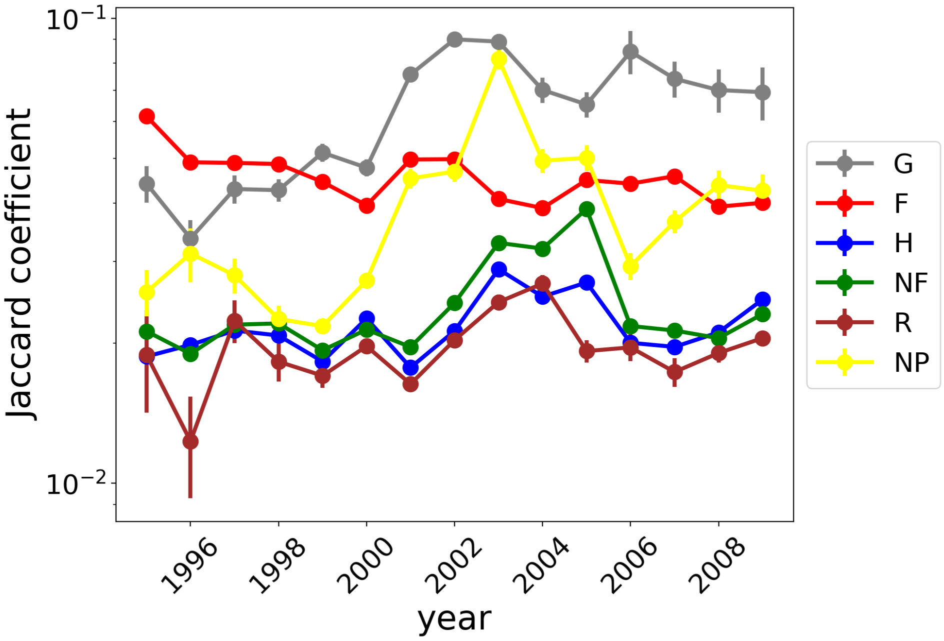


Figure 3.15: Average Jaccard correlation between the binary vectors of trading activity for the six category of the database as a function of the calendar year. Governmental organizations (G), Financial companies (F), Households (H), Non Financial companies (NF), Foreign organizations (R), and Non profit organizations (NP).

the three possible choices for each trading day of a calendar year has been computed. Fig. 3.15 shows the average Jaccard correlation for the six categories of the database as a function of the calendar year. In particular the figure shows that investor categories characterized by lower similarity among their trading activity are the categories of households (H) and foreign organizations (R). On the contrary, categories with higher global similarity are governmental organizations (G) and, to a lesser degree, financial corporations (F) and non-profit institutions (NP). The amount of similarity is rather persistent and stable over the years.

A last investigation concerns the relationship between the logarithmic ratio of validated links and the average daily volatility computed each calendar year. Fig. 3.16 shows the logarithmic ratio of validated links as a function of the average daily volatility for all the investigated years. The figure shows that the two quantities are anti-correlated: in fact, the Pearson's correlation coefficient between them is -0.59. In other words the rate of similar investment profiles observed between pairs of investors is exponentially sensitive to the volatility of the market. Periods of low volatility are associated with

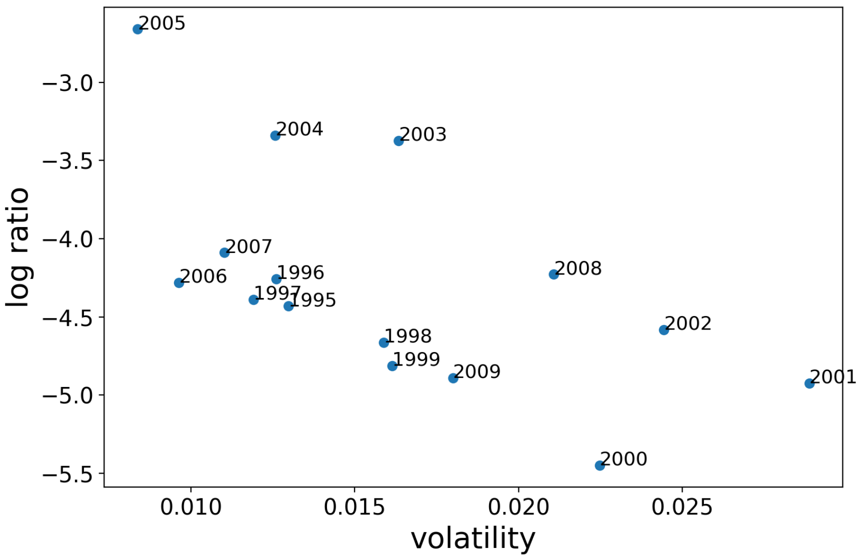


Figure 3.16: Logarithmic ratio of validated links as a function of the average daily volatility estimated for all investigated years.

high values of the logarithmic rate and viceversa. Since the ratio of validated links can be seen as an estimator of the extent to which the activity of investors is synchronized during a year, this value of correlation suggests that the more volatile is the market, the less likely is that investors are organized in stable, synchronized groups. Fig. 3.16 also shows that on top of the role of volatility some other variable (or variables) is crucially affecting the logarithmic rate. This second effect seems to be of bimodal type clustering together years(i) 2001, 2002, 2003, 2004, 2005, and 2008 and (ii) 1995, 1996, 1997, 1998, 1999, 2000, 2006, 2007 and 2009. Although it was not possible to find a clear explanation for this bimodal organization of the scatter plot, it is worth noting that all the years of the onset of the financial bubble are in the second group.

4 Demographic trends and social dynamics of a speculative bubble

In the study of financial systems, a widely adopted assumption is the stationarity of the underlying processes [6]. For a time series, stationarity implies that the distribution of elements of the series does not vary if it is shifted in time. Thus, a set of logarithmic returns $r_{t_1}, r_{t_2}, \dots r_{t_n}$ will have the same distribution of the set $r_{t_1+T}, r_{t_2+T}, \dots r_{t_n+T}$, for any T. The idea behind this assumption is that financial processes do not change significantly over time, and their dynamics is affected equally by the same trends and phenomena. Although this assumption was proven to be valid across several markets in different periods, with the identification of many stable patterns and trends in financial processes [43], there is also strong empirical evidence of different behaviors. Indeed, the search for stable, stationary phenomena is only one aspect of the dynamics of financial systems. Often, the collective behavior of agents interested in making money out of temporary trends brings to the development of unstable, bursty processes that may contradict some of the assumptions of financial theories. Speculative bubbles are a significant example of these processes. In finance, a bubble can be defined as "a situation in which temporarily high prices are sustained largely by investors' enthusiasm rather than by consistent estimation of real value" [69]. Indeed, one frequent observation on speculative bubbles is that, during their inflation, the price of the involved assets is detached by its real value and has no relationship with common financial tools like fundamentalist estimations. In bubbles, a common mechanism is the large diffusion of momentum strategies of investors that focus on consistently buying an asset with a growing expectation of making a profit due to the increase in its price. When the price is pushed too much by the action of investors, the bubble bursts, and a momentum strategy of opposite direction is adopted by the majority of investors, leading to a dramatic drop in the price. In the context of the present dissertation, it is worth translating the definition of speculative bubbles in the language of complex systems. From this perspective, a bubble is the effect of an emergent phenomenon strongly related to the feedback among investors. In fact, the bubble continues to grow until a large number of investors consistently invest money on the inflating asset. Specifically, the feedback among investors is indirectly delivered through the price of the asset, that increases as an effect of the collective behavior of buyers.

Historically, the first recorded occurrence of a speculative bubble was observed in Holland in the XVII century, as it was documented by Charles Mackay in 1841, [70]. At that time, Dutch merchants began to import tulips from the Ottoman Empire. Due to their brightly colors and peculiar shapes, the tulip became a popular item all over the country, leading to an increase in demand and thus in their price. Moreover, the exchange of contracts that worked as modern futures, giving the right to own a given amount of tulips at the fulfilment of extablished conditions, fueled the enthusiasm of customers, since they were allowed to trade with no actual exchange. According to Mackay, the commerce of tulips became so wildly spread that "the rage among the Dutch to possess them was so great that the ordinary industry of the country was neglected". Eventually, the bubble popped when, while taking arrangements for a big purchase, a buyer was not found. This event opened the eyes of traders on the unsustainability of the increase in the price of tulips, which fell to a very small fraction of its previous value. Even though the accuracy of Mackay's account have been recently discussed, [71], this event contributed to introduce the awareness on the occurrence of periods in which a (set of) asset(s) can be strongly mispriced on financial markets due to the collective action of investors. Indeed, recent history is full of examples of speculative bubbles: from the Japanese asset price bubble of late 80's to the housing bubble that led to the outbreak of the big financial crisis of 2008, this phenomenon seems to reiterate regularly over time. Particular interest in the context of this dissertation is going to be put on the *dot com* bubble, which developed in most industrialized countries in the last years of the XX century and affected the assets linked to companies that belonged to the tech sector. Despite the diffusion of studies on the causes and the features of a bubble, an exhaustive comprehension of its dynamics is still missing, together with a clear understanding of the methods to be implemented to reduce or control such a phenomenon.

The existing literature has focused on several aspects of the dynamics of a bubble. One big branch is about the description and characterization of the features of the time series of assets involved in an inflating bubble. In this direction, some studies have applied recursive regression methods to detect signals that may lead to the outbreak of a bubble, in order to provide an early warning on bubbles' development [72], [73]. In other works, an attempt to model the time series of the price of assets involved in a bubble has been carried out by associating the mainly monotone price series with strict local martingales [74]. In [75], instead, a classification of speculative bubbles on the basis of the associated volatility was proposed. Specifically, a distinction is introduced between *fearful* bubbles, associated with a significant increase in volatility during its inflation and *fearless* ones, in which no significant changes in volatility are detected.

Other works try to estimate the impact that different factors have on the dynamics of a bubble. Caginalp et al., in [76], have investigated the relationship between liquidity and the maximum levels of price reached by an asset during a bubble. They found out that the two quantities are positively correlated, and as a consequence deferring dividends (thus reducing liquidity) during the inflation may reduce the size of the bubble. Another work by Lux, [77], instead stresses how the variable that mostly affects the social diffusion of trading decisions among investors during a bubble is the series of return of the involved asset. The fluctuations in this variable are enough to shape a collective behavior that shows herding characteristics.

Other directions involve the characterization of the social demographic attributes of investors entering the market during the inflation of a speculative bubble. For example, Grinblatt et al., in a paper that investigates the same dataset on Finnish investors which is used extensively in this dissertation, characterize the timing of the trading decisions of investors on the basis of their IQ score during the *dot com* bubble in Finland [78]. The analysis performed in this chapter goes in this direction. Indeed, by combining the dataset presented in the previous chapter with a second dataset that records demographic information on the whole population of Finland, a socio-demographic characterization of the sets of investors entering the market in order to buy the Nokia asset for the first time in different moments of the dynamics of the *dot com* bubble has been performed. Moreover, an attempt of modelling the process of buying Nokia on the basis of an opinion dynamics of reservation prices among the Finnish population is proposed in the following sections. This chapter is organized as follows: section 4.1 describes the investigated datasets, section 4.2 presents the methodology and the results for the socio-demographic investigation of investors, and section 4.3 introduces the model for reservation prices as a variation of the Deffuant model.

4.1 Datasets

For the current investigation, the same dataset described in the previous chapter has been used. Indeed, its structure allows to extract the time series of investors at a daily time scale. For this analysis, the focus has been put on the investors that entered the market for the first time in order to buy the Nokia stock. We chose Nokia because at the time was the leading tech company traded in Finland (and it was also the most traded asset in general), and is one of the assets which were most affected by the bubble. Fig. 4.1 plots the closing prices of Nokia during the years 1995-2009. The figure shows the rapid increase in price experienced in the years 1999-early 2000, and the abrupt drop at the end of 2000. In 2007, years after the burst of the bubble, there is another moderate increase in the price which falls back to lower values during 2008.

The second dataset used in this chapter was collected by the Statistics Office of Finland, and contains demographic information on the population of the Finnish country. Specifically, it records information on age levels, income, education and job demographic. For each field, the information is reported at the level of single postal code. For the postal codes whose population is below 50 individuals only the total number of inhabitants is reported.

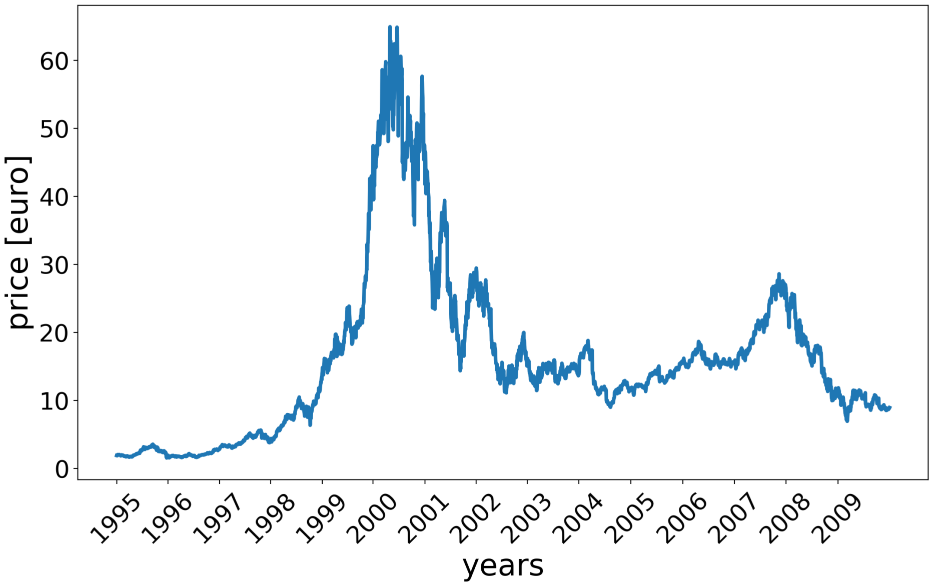


Figure 4.1: Closing price of the Nokia asset in the years 1995-2009. The effects of the inflation of a speculative bubble are evident from 1999 to early 2000. The end of 2000 is characterized by a dramatic drop in the price, associated with the burst of the bubble.

4.2 Characterization of new investors

4.2.1 Time series of new entries

The first step of the present investigation is the detection of the time series of investors accessing the stock market for the first time in order to buy Nokia. We chose to track only investors that had never had ownership of stocks different from Nokia in order to stress the idea that these investors were completely new to a financial investment and were attracted into the market by the performance of Nokia. Moreover, the investigation has been limited only to households in order to focus on the decision of individual investors that acted on behalf of themselves or their families. The top panel of Fig. 4.2 plots the number of newcomers buying Nokia for the first time in the investigated period at a daily scale. The first half of the time series is strongly bursty, with high peaks of new entries occurring in very narrow time windows, often of just one day. An inset, drawn in red color, focuses on the period in which the the inflation process was stronger, 1999-2000, in order to highlight the shape and the dynamics of peaks. The highest peak was observed on the 28^{th} of July in 2000. It is worth noting that the previous day, the

27^{th}, a quarterly earning announcement was released. These announcements, released by the company at regular intervals (four per year) play a significant role in the trading history of an asset because they are usually characterized by the very intense activity of investors reacting to the news. Moreover, as a consequence of this trading activity, the price of Nokia had experienced a very significant drop in price, falling to about the 79% of its precedent value. Thus, the huge peak of the 28^{th} of July is the result of the decisions of a large number of investors that saw in the low price of Nokia a good opportunity to start their investment. This process is at the basis of the model that will be introduced in the following sections. The period that goes from 2005 to 2007 is characterized by a very small rate of new investors. The dynamics becomes again bursty in 2008 and 2009.

The bursty behavior of the time series of new entries reveals also another aspect: the activity of investors in the investigated period was highly heterogeneous, with strong differences in the numbers of investors entering the market before, during and after the bubble. This heterogeneity is summaryzed in the bottom panel of Fig. 4.2, that shows the numbers of household investors entering the market year by year. Thus, in order to characterize the fluxes of new investors, a tool able to deal with such heterogeneity is required. The characterization method adopted in this context, which is taken from the formalism of statistical validation, is presented in next section.

4.2.2 Statistically expressed sets of new investors

The demographic characterization of the dynamics of the *dot com* bubble is based on the detection of those attributes that are over-expressed among the new investors with respect to the whole population. Since both the numbers of investors and the distribution of attributes in the Finnish population are highly heterogeneous, the tool selected for this purpose is again the statistical validation of attributes. In this context the null hypothesis represents a process in which new investors are randomly drawn from the whole population. If in a year the number of inhabitants of Finland is N, with N_i being the new investors and N_A the amount of people in Finland characterized by the attribute A, one can test whether $N_{i,A}$, i.e. the number of investors buying Nokia for the first time that have the attribute A, is compatible with the null hypothesis. This is obtained

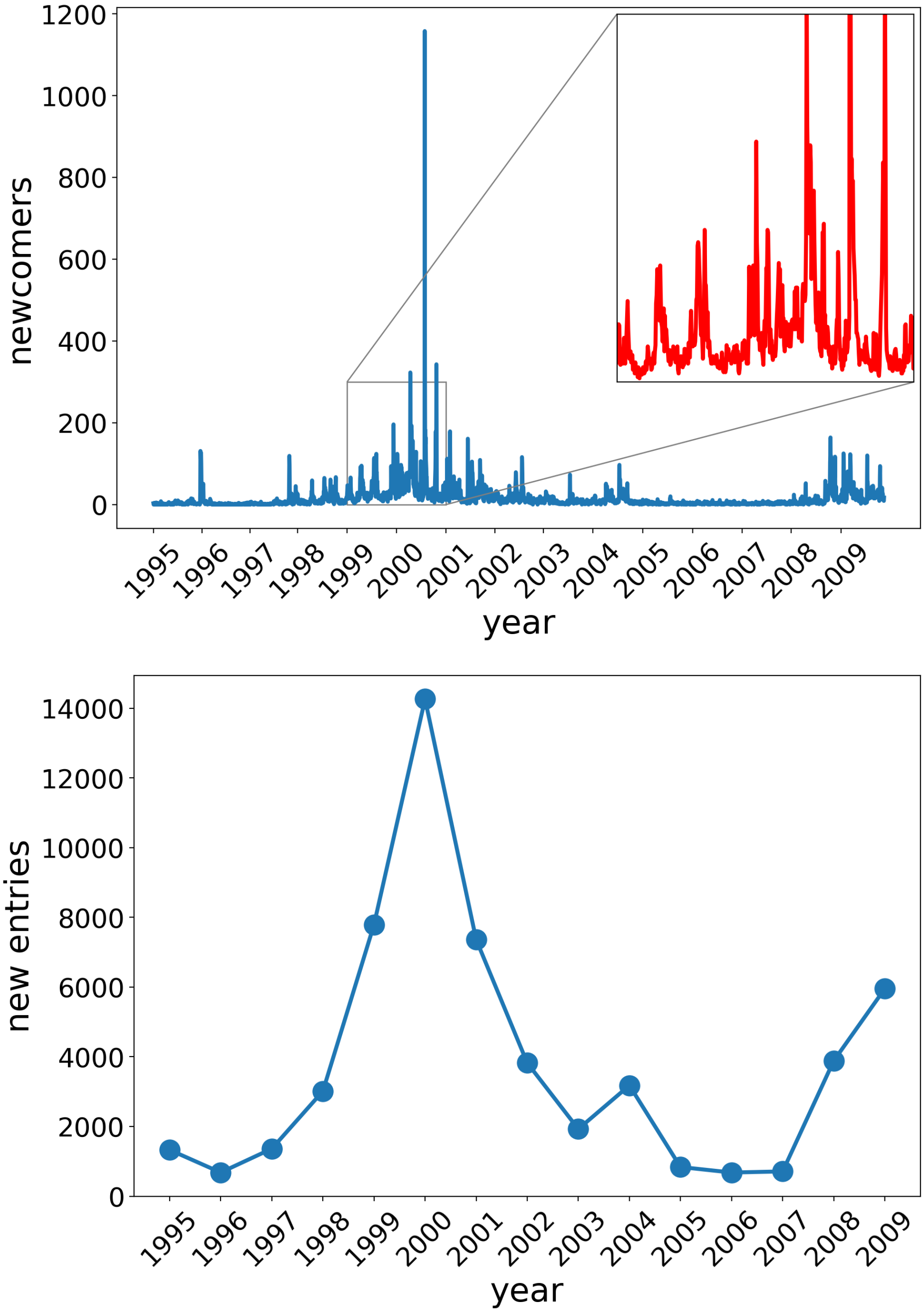


Figure 4.2: Time series of the numbers of investors accessing the stock market for the first time in order to buy the Nokia stock. The top panel plots the time series at a daily scale, with an inset that zooms the peaks in the years 1999 and 2000. The bottom panel plots the same time series aggregated on an yearly scale.

by computing the corresponding p-value,

$$p_1(N_{i,A}) = 1 - \sum_{X=0}^{N_{i,A}-1} H(X|N, N_i, N_A). \tag{4.1}$$

As already discussed in the previous chapter, Eq. 4.1 represents the probability of finding a quantity equal or larger than $N_{i,A}$ investors with the attribute A with a random draw from the whole population. Thus, testing this p-value allows to detect whether the attribute A is *over* expressed among the investors, i.e. whether it occurs more than expected according to the null hypothesis. Thus, an over expressed attribute identifies a category of people in the whole population whose rate of entrance was so high to deviate from the heterogeneity of the system. In this context, it is interesting to detect also the attributes that are *under* expressed among the investors, i.e. those attributes that occur less than expected according to the null hypothesis. Indeed, under expressed attributes represent the categories of investors that were entering the market at rates so low to deviate from heterogeneity. Thus, considering both kinds of expressions allows to obtain a clearer demographic characterization of the categories of investors more/less involved in the bubble. In order to detect the under expressed category, one should look at the other tail of the hypergeometric distribution, as shown in [79]. Specifically, the p-value for testing under expressions is computed according to

$$p_2(N_{i,A}) = \sum_{X=0}^{N_{i,A}} H(X|N, N_i, N_A). \tag{4.2}$$

The p-value p_2 represents the probability of finding a quantity equal or smaller than $N_{i,A}$ investors with the attribute A with a random draw from the whole population. Thus, a small value of p_2 indicates that the attribute A is observed among investors less than expected according to the null hypothesis. Once the two-tail p-values are computed, they are tested taking into account the corrections for multiple comparisons, with a significance threshold of 0.01. Specifically the correction used is the control for the False Discovery Rate (FDR) [51].

The motivation for such a test can be found by looking at Fig. 4.3. The top panel of the figure plots the heatmap of the ratios of the number of new investors of different age levels on the total number of Finnish people of the same age levels, for different years. The bottom panel instead plots the ratios of the number of new investors of different age levels on the total number of investors, for different years. Thus, the top panel shows that the fractions of Finnish people of different ages entering the market grow significantly in the years of the bubble, reaching a peak in 2000, but it does not contain information on the distribution of age levels between all the investors. On the other hand the bottom panel shows that the youngest and oldest population groups are less represented among investors, but it does not take into account how this relationship evolves with respect to the global population. The test on over/under expressions, instead, focuses on the age levels more/less involved in the bubble dynamics by naturally taking into account both sources of heterogeneity. Fig. 4.5 plots the outcomes of the statistical validation of age levels in the investigated period. In the figure, each dot is red if the corresponding age level was over expressed in the corresponding year, it is blue if it was under expressed and it is white otherwise. A careful investigation of the figure reveals that people younger than 18 years and those older than 70 are almost always under expressed, while, people between the age of 25 and 45 are almost always over expressed. However, in 2005, 2006 and 2007, which were year characterized by small rate of new investors, the pattern of over/under expression is less pronounced. Moreover, it is interesting to observe the behavior of the remaining age groups: people of 18-19 years are under expressed only in a part of the investigated period, which broadly overlap the complete dynamics of the bubble. People of 20-24 years instead are also under expressed only in 2000 and 2001, which were the years in which the bubble popped, while they were over expressed in 1995. Thus, people of this age level proved to be moderately interested in trading, but on average more aware of the risks of the bubble once it was bursting. On the other hand, people of age between 45 and 65 had an opposite behavior. Indeed, they are over expressed only in a set of year that broadly covers the dynamics of the bubble, and continued to enter at significant rates also when it popped. The case of people of age between 60 and 64 is meaningful: these investors were over expressed only in the worst years, 2000 and 2001, in which the price of Nokia dropped dramatically. Before going on, it is worthwhile comparing the pattern of the top panel of Fig. 4.3 with those of Fig. 4.4. Indeed, the latter plots the ratios of all the investors with an open position on the Nokia stock at a yearly scale. Thus, Fig. 4.4 shows the distribution in age of all investors that are active on Nokia in a given year, not only those that start investing in that year. From the comparison, it is evident that the two patterns are quite different, and the peak observed in 2000 for the new investors did not affect significantly the whole distribution of investors. This is an additional evidence of the extraordinary nature of the dynamics of the bubble, whose patterns deviated significantly from those observed during the whole period.

Fig. 4.6 instead plots the pattern of over/under expressions of gender among investors in the investigated period. In this case the situation does not show an evolution, since the number of male investors is always greatly larger than the number of female ones. This evidence shows that the dynamics of the bubble did not affect significantly the role of gender in characterizing the interest towards a financial investment.

The other attributes of the census data (education, income and job) required additional work in order to be investigated. Indeed, gender and birth date are included in the metadata on investors of the Euroclear dataset. This implies they can be extracted directly from the data. Instead, the Euroclear data does not contain information on education, income and job; thus, it is not possible to obtain directly the distribution of investors with respect to these attributes. In order to overcome this issue, the following procedure has been followed: we first extracted from the census data the empirical probability distribution of attributes conditioned on postal codes. These empirical probabilities are then conditioned also on age levels, according to the principle of maximum entropy, which assumes the prior distribution to be uniform between age levels. Whenever it is possible, linear constraints on age are assigned to the prior distributions. For example, when dealing with job information, people younger than 65 years are never classified as retired. Since the adoption of constraints can alterate the normalization of the conditioned probabilities, each discrete probability is then corrected by dividing it for the overall sum. Thus, once the conditioned probabilities for the whole population are computed, they can be used to obtain the number of investors with the investigated attributes for each postal code and age level (both information are present in the Euroclear data). Then, aggregating on postal codes and age levels one obtains the distributions of investors with respect to the investigated attributes. Thus, by applying this methodology, one is able to reconstruct the profiles of investors belonging to a category of attributes by taking into account their distribution in space and in age. If many

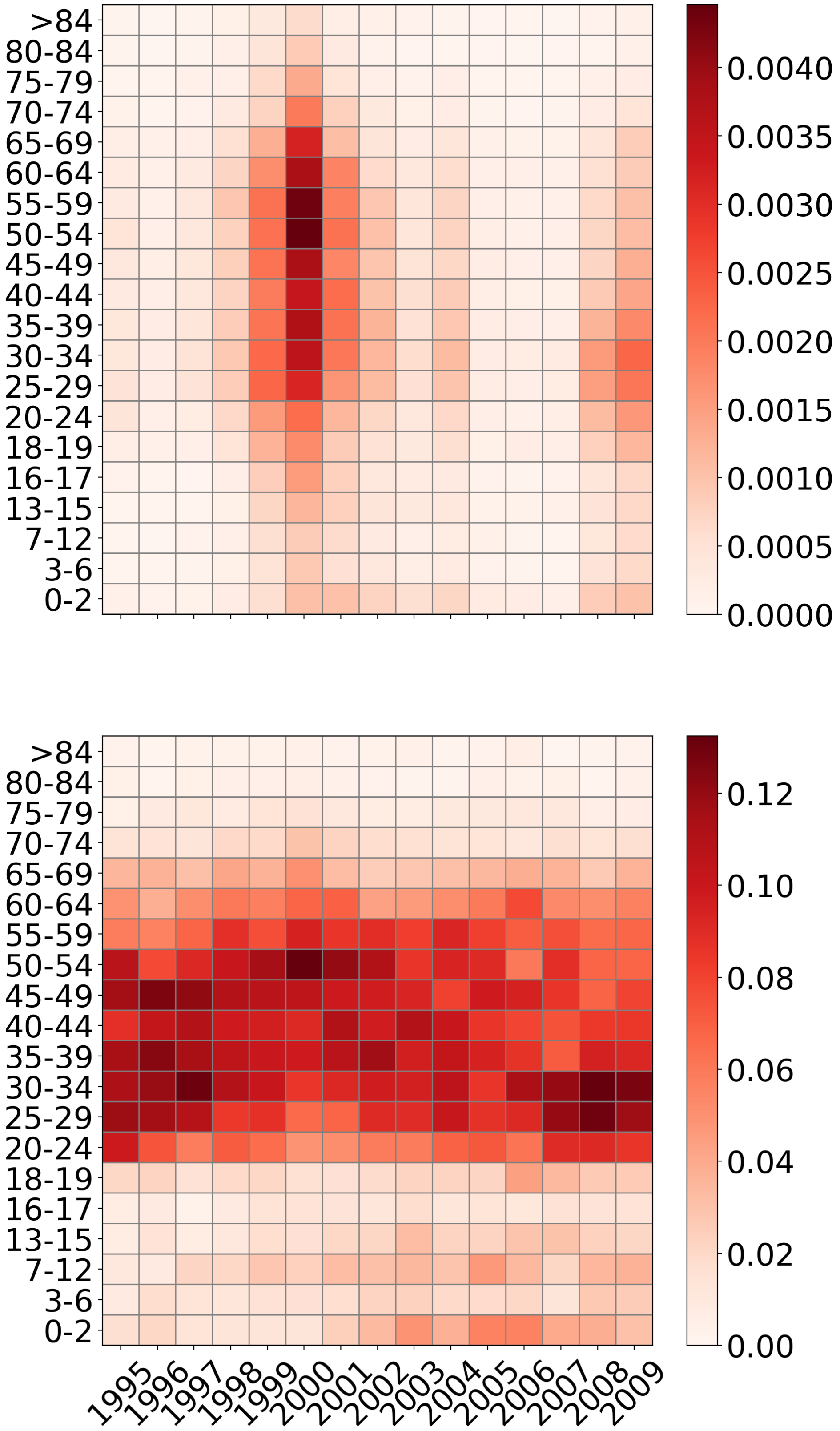


Figure 4.3: Heatmaps of the distribution of age levels among new investors as a function of time. The top panel plots the ratios of new investors of different age levels to the total number of Finnish people of the same age. The bottom panel plots the ratios of new investors of different age levels to the total size of the set of investors.

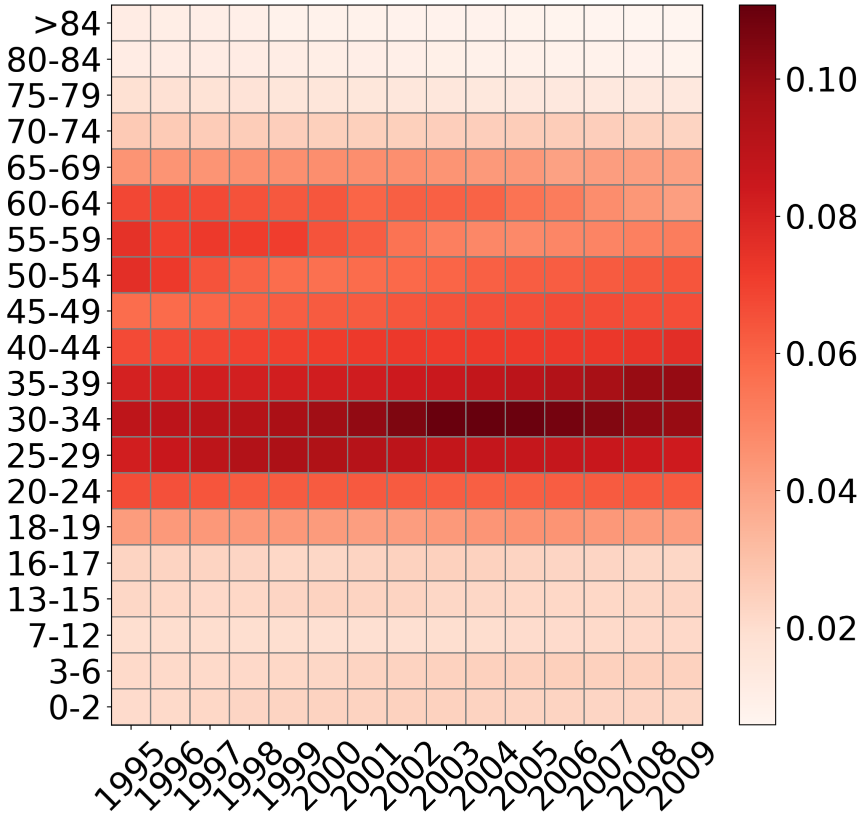


Figure 4.4: Heatmaps of the distribution of age levels among all investors with an open position on the Nokia stock. For each investor, the age is computed when he opens the position and remains constant in the following years. The panel plots the ratios of investors of different age levels to the total number of Finnish people of the same age.

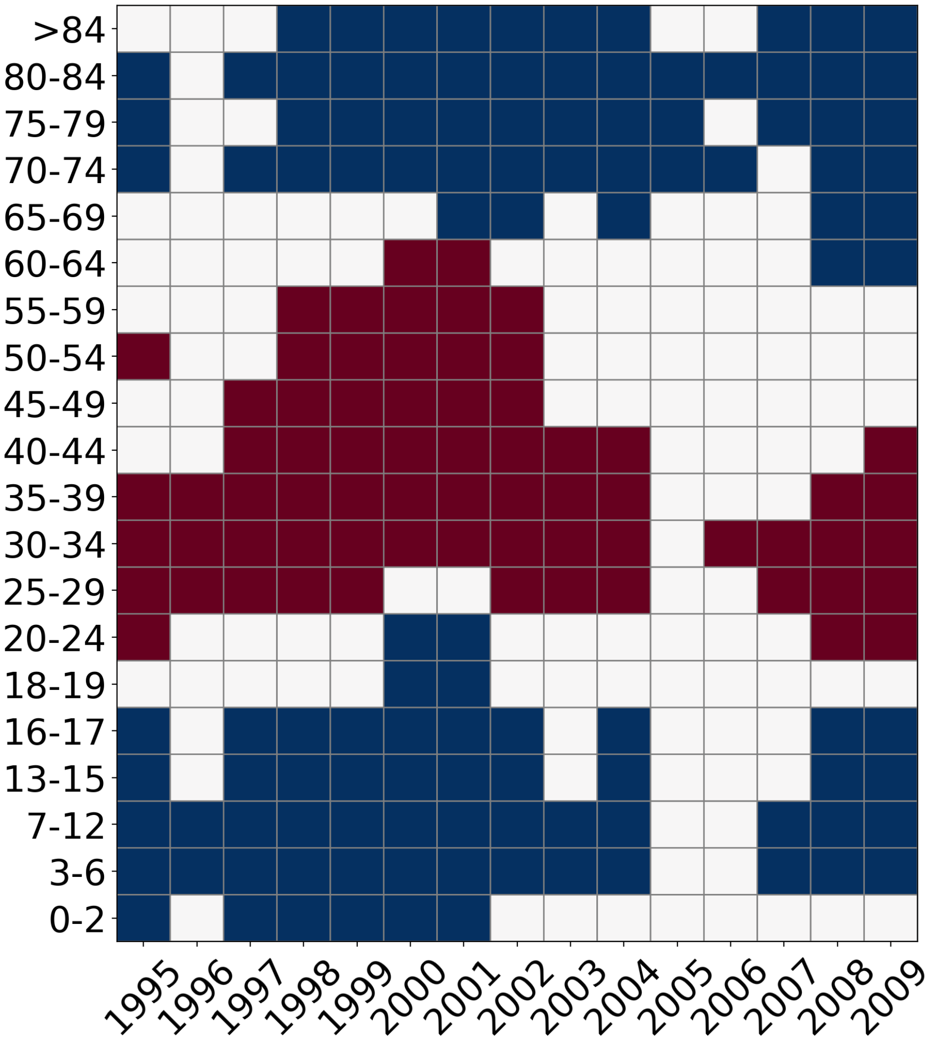


Figure 4.5: Heatmap of over and under over expressions patterns for age levels as a function of time. Each dot is red in the presence of over expression, blue in the presence of under expression and white in the presence of neither.

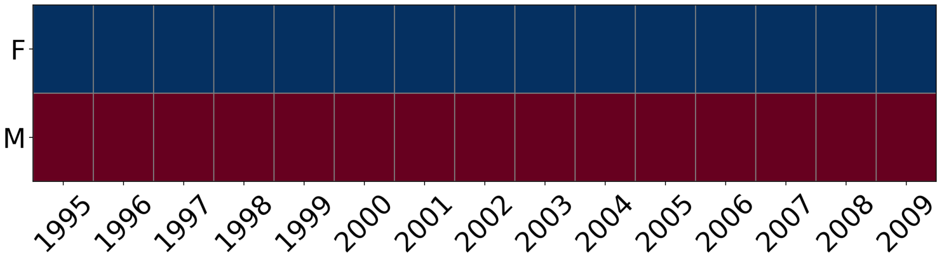


Figure 4.6: Heatmap of over and under over expressions patterns for gender as a function of time. Each dot is red in the presence of over expression, blue in the presence of under expression and white in the presence of neither.

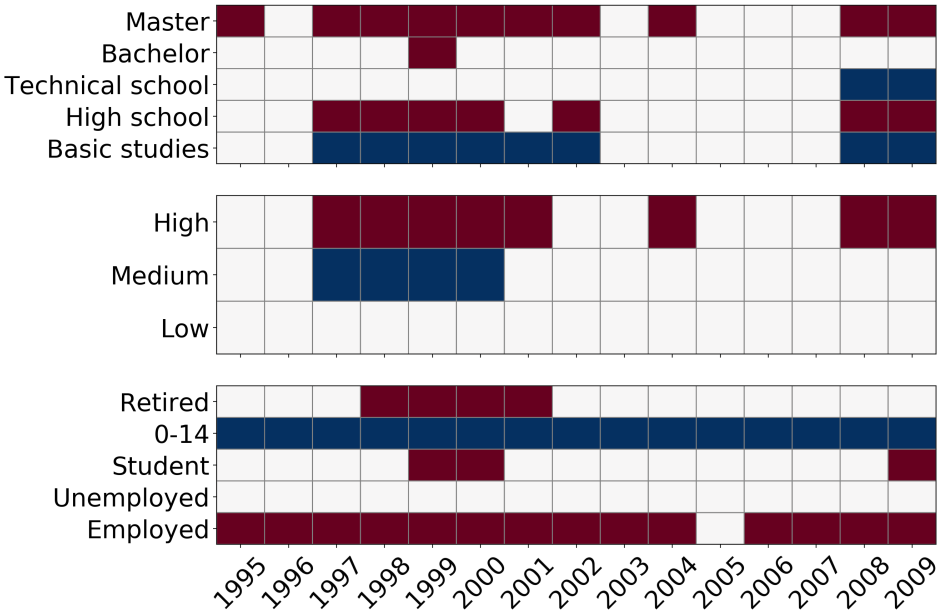


Figure 4.7: Heatmap ofver and under over expressions patterns for education (top panel), income (center panel) and job (bottom panel) as a function of time. Each dot is red in the presence of over expression, blue in the presence of under expression and white in the presence of neither.

investors come from a postal code and have an age which are strongly characterized by a given attribute, this will be properly reflected in the distribution of that attribute among all the investors.

Fig. 4.7 plots the over/under expressions patterns for these attributes. The top panel shows the statistical expressions for education levels. People with a master are over expressed in 1995, in the years of the bubble, from 1997 to 2002 and later in 2004 and 2007-2008. People with high school education are over expressed in 1997, 1998, 1999, 2000, 2002, 2008 and 2009. People with a bachelor result over expressed only in 1999, signaling an increase in investment decisions for this category strongly localized in the period in which the bubble was steadly growing. People with the lowest level of education, indicated as basic studies, are always under expressed in the period from 1997 to 2002, showing a low interest in investing in the bubble, and then again in 2008 and 2009. People with a technical education instead never deviate from the heterogeneity of the system apart from the last two years, in which they are under expressed. The panel in the center plots the pattern of statistical expression for income levels. The levels are computed looking at the global distribution of income in Finland. The high (low) level refers to people with an income that lies above (below) the 80^{th} (20^{th}) percentile of the nationwide distribution of income, while all the other investors belong to the medium level. The panel shows that newcomers with a high income are always over expressed from 1997 to 2001, in 2004 and in the years 2008-2009. In the years from 1997 to 2000, investors with medium level of income are always under expressed. Low income investors instead never deviate from heterogeneity. The bottom panel instead plots the patterns of over/under expressions for the different job groups reported in the census data. In this case, employed investors are always over expressed with the exception of 2005, while people younger than 14 are always under expressed, signaling that these two category are permanently more and less represented among investors with respect to the whole population. Retired people instead were over expressed only in the period from 1998 to 2001, proving to be more sensitive to the dynamics of the bubble. Students were over expressed only in 1999 and in 2000, which were the years in which the bubble dynamics was more pronounced, and again in 2009. Unemployed people instead never deviated from heterogeneity.

4.2.3 Characterization of postal codes

The last step in the characterization of new investors involves their geographical distribution, indicated by the postal code of their residence address. Finland has 3200 single postal codes of five digits that span the whole country. In this case, the statistical investigation of investors has been performed by looking at the over/under expression of investors living in different postal codes with respect to the whole population living in the same area. Since a large number of postal codes, especially those in the North of the country, are inhabited by small numbers of people, the postal code grid has been aggregated in order not to reduce the power of the statistical test. Specifically, the aggregation was performed by putting together all the postal codes that share the first three digits. This aggregation produces 839 groups of postal codes located in areas which are geographically close. For example, the aggregated postal code 001** contains all the single postal codes of the central area of Helsinki. Thus, the over/under expression of a given postal code in an year represents the deviation from heterogeneity of the numbers of investors living in it when taking into account the size of the population in the same area, the size of the whole Finnish population and the global number of new investors. Table 4.1 provides a summary of the numbers of over and under expressed postal codes detected in the investigated period. The table shows that the number of statistical expressions is rather limited, ranging from 0.2% to 3% for over expressions, while under expressions never occurs, with the only exception of 2000, in which only one under expression is detected. In order to understand the results of such a statistical characterization, the existence of a relationship between over expression and average income has been investigated. However, since the fraction of over expressed postal codes is always very small, a test on the statistical significance of such an investigation has been conducted. Specifically, for each year, a bootstrap sampling of the income values of postal codes without an over expression has been made. The adopted bootstrap procedure is based on obtaining random samples (allowing replacement) of a given set, in order to obtain several realizations of the investigated variable. In this case, the sizes of the bootstrap samples were fixed equal to the number of over expressed postal codes in the corresponding year, in order to compare sets of the same size. Then, for each realization, the average income of the bootstrap sample is compared with the average income of the set of over expressed postal codes. The meaning of this test is to detect whether the incomes of the two sets come from the same distribution or not. Table 4.2 reports some statistics about the bootstrap procedure. The table reports year by year the percentiles at 2.5% and 97.5% for the distributions of the average income on 100,000 bootstrap realizations, the average income of the over expressed postal codes and the fraction of times in which the average income of the bootstrap realizations was larger than the corresponding quantity for the over expressed postal codes. The last column thus indicates the p-values of a test that detects whether distributions of income in the two sets are different or not. The p-values are so low that the null hypothesis of equal distribution is always rejected with a significance threshold of 0.05, properly corrected on the total number of tests. The results of this bootstrap procedure is summarized also in the top panel of Fig. 4.8. The blue line in the panel plots the average value of income for the set of over expressed postal codes. The orange band instead plots the intervals between the percentiles at 2.5% and 97.5% for the average income of the bootstrapped samples. It is evident how the line and the band never overlap, revealing

Table 4.1: Summary statistics of the amount of over and under expressed postal codes in the different years. The total number of postal codes is 839 and remains stable over time.

Year	OE	UE
1995	5	0
1996	2	0
1997	6	0
1998	10	0
1999	17	0
2000	24	1
2001	10	0
2002	6	0
2003	2	0
2004	3	0
2005	2	0
2006	2	0
2007	2	0
2008	12	0
2009	14	0

how the distributions in the two sets remain always distinct, with the over expressed postal codes always characterized by a higher level of income.

In order to quantify the extent of the difference in income between the over expressed postal codes and the others, a modified version of the Goodman and Kruskal gamma indicator [80] has been used. The measure was computed in the following way: the set of postal codes is splitted into two sets, the over expressed ones and the others. Then all the couples between elements of the two sets are considered, distinguishing between N_a, the couples in which the over expressed postal code has a higher income and N_b, the couples in which the opposite hypothesis holds. The gamma is then computed as

$$\gamma = \frac{N_a - N_b}{N_a + N_b}. \tag{4.3}$$

The Goodman and Krukal gamma can take values between -1 (over expressed postal codes have always a lower income) and 1 (over expressed postal codes have always a larger value), with 0 signaling no relationship between income and over expression. The bottom panel of Fig. 4.8 plots the values of the gamma as a function of different calendar years. The figure shows that the over expressed postal codes are always associated with higher income, as already shown by the bootstrap procedure. However, the Goodman and Kruskal gamma adds some information. Indeed, although over expression and average income of a postal code are always positively correlated, during the bubble the magnitude of this correlation decreased. The gamma starts with values very close to one at the beginning of the investigated period, it decreases in the following years until it reaches its minimum in 2000, and then it gradually recovers to its previous values. This pattern is an evidence of the fact that investing during the bubble was a social phenomenon that strongly affected also regions associated with lower levels of income, that usually do not show high rate of new investors. The same phenomenon occurs also in 2008 and in 2009, that were characterized again by pronounced, bursty series of new investors.

4.3 Modeling the inflation of a bubble

In the previous section it was shown that the largest wave of investors buying Nokia for the first time during the bubble dynamics occurred the day after a quarterly earning announcement. Specifically, immediately after the announcement the price fell significantly, closing at 45.00 euro (the opening price in the same day was 57.15 euro). Thus, the large number of investors entering the market after the announcement saw in this drop in price an opportunity to start their investment on Nokia at a convenient price. This event suggests the presence of a class of potential investors that, before entering the market during the inflation of a bubble, monitorate the price of the involved assets waiting for a temporary drop. Indeed, since their expectation is that afterwards the price will keep growing at its usual rates, they see in the drop a possibility to beat the market, i.e. to maximize their profit by entering the market at a price which is lower than usual. In financial literature, this process is associated with the fixing of a *reservation price*. A reservation price is a threshold fixed by an investor that represents the maximum

Table 4.2: Summary statistics of the bootstrap procedure on the income variable. The table reports the year, the percentiles at 2.5% and 97.5% for the distribution of average income in the samples bootstrapped from the postal codes without an over expression, the actual average income for the over expressed postal codes and the fraction of times in which the average income of the bootstrap sample was higher than the actual value. The performed bootstrap replicas are 100,000.

Year	2.5 Perc. BS	97.5 Perc. BS	Average OE	p-value
1995	8155.1	9517.4	11500.7	0.00000
1996	7897.2	10245.3	11963.3	0.00011
1997	10156.1	11805.3	14537.0	0.00000
1998	10659.9	12118.1	14490.9	0.00000
1999	11292.3	12610.7	15270.2	0.00000
2000	11719.0	12872.3	16035.1	0.00000
2001	12291.9	14448.5	17662.0	0.00003
2002	11451.9	13861.6	18466.6	0.00007
2003	12245.4	16881.3	21465.1	0.00002
2004	13865.2	18635.5	26889.2	0.00000
2005	13944.3	19893.7	28735.1	0.00003
2006	14243.5	20964.2	27419.9	0.00209
2007	15389.4	22456.7	29932.7	0.00010
2008	17621.7	20499.5	26073.5	0.00000
2009	17897.3	20425.1	25327.6	0.00000

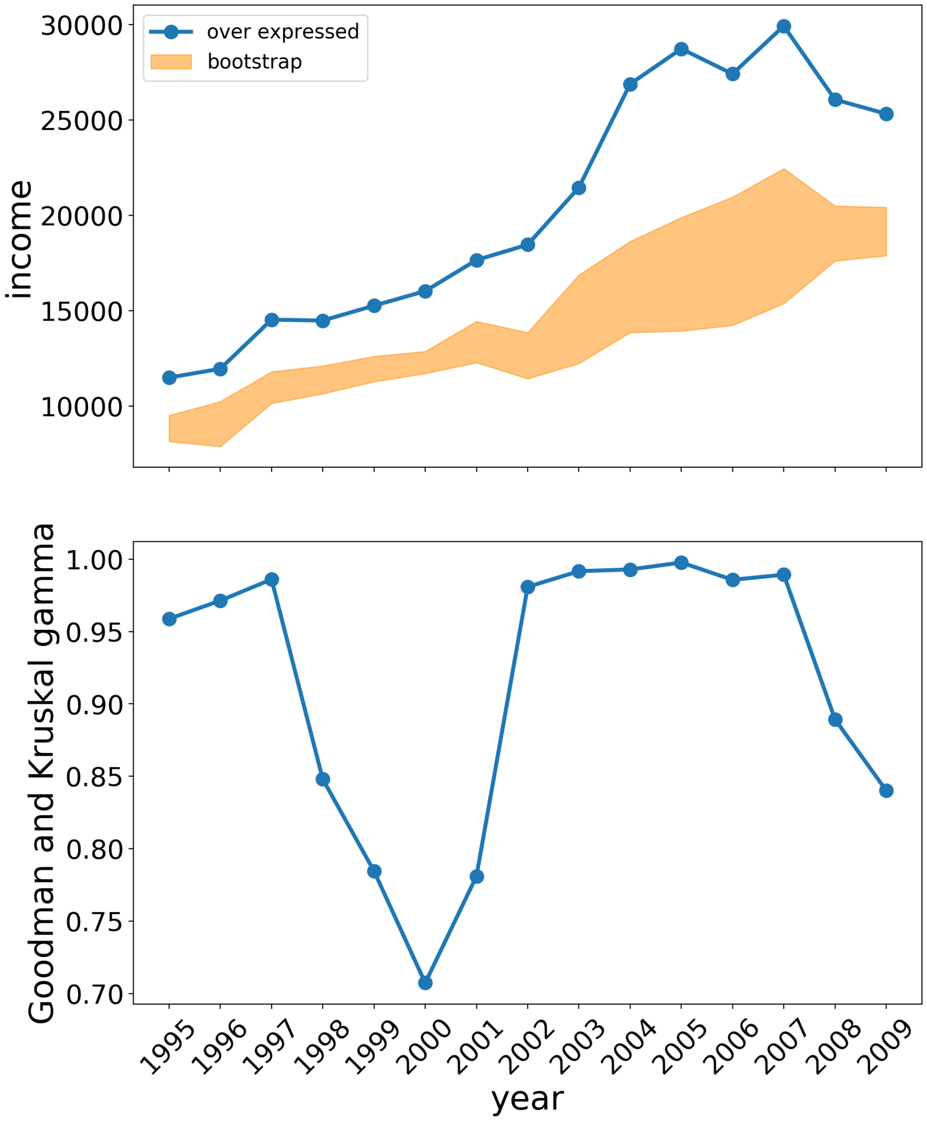


Figure 4.8: The top panel plots the average income of the over expressed postal codes (blue line) versus the intervals between the 2.5% and the 97% percentiles of the average incomes on the bootstrap samples (orange band) as a function of time. The line and the band are never intersect, showing that the distribution of income in postal codes with and without an over expression is different. The bottom panel plots the Goodman and Kruskal ranking correlation gamma between over expression and income as a function of time. Although the gamma is always positive, it shows a decreasing dynamics in the years in which the bubble is inflating, recovering to its precedent values after the bubble bursts. It decreases again in 2008 and 2009.

price he/she is willing to pay in order to buy a given asset. If the price goes below the reservation price, the investor will buy a certain amount of stocks of that asset. This mechanism belongs to the class of strategies related to the concept of realization utility, as shown in [81]. Specifically, the adoption of a reservation price can be connected to the evidence that the majority of individual investors have a greater propensity to sell stocks whose price has grown since purchase. This mechanism is known as *disposition effect.* In order to understand whether the presence of a distribution of reservation prices in a population could contribute to the bursty behavior of investors entering the market during the *dot com* bubble observed in Fig. 4.2, the relationship between the rate of new entries and the variations of the price has been investigated. Specifically, for different time windows τ, the moving average of logarithmic returns has been computed, and the Pearson coefficient of correlation with the time series of new investors has been evaluated. This measure investigates whether large variations in the price at different time scales affect the entrance rate of investors. Fig. 4.9 plots the Pearson correlation coefficient as a function of the time window τ at which the moving average is computed. The τ can be seen as the memory of the previous price history that the investors take into account when making trading decisions. The figure shows that the Pearson coefficient is always negative, signaling that the investors are more likely to enter when the price decrease. A possible explanation for this mechanism is the adoption of reservation prices by new investors. Moreover, the Pearson correlation changes with the adopted memory. In fact, it shows a negative peak for $\tau = 8$, and it ranges from -0.13 to -0.22. Thus, the reaction time to the variations of price vary between the investors.

In the present context, observing this aspect of the dynamics of the bubble gave support to the idea that the diffusion of increasing reservation price among investors is one of the factors that fuel the inflation of a bubble. Indeed, as an effect of the significant increase in price shown by an asset during a bubble, an increasing number of potential investors start to became interested in the possibility of an investment. Due to the steady increase in price, they adjust their reservation prices over time, waiting for the right occasion. As soon as there is a drop in the price that goes below their reservation prices, they massively enter the market. In order to reproduce this dynamics, an agent based model that describes the diffusion of increasing reservation prices among investors

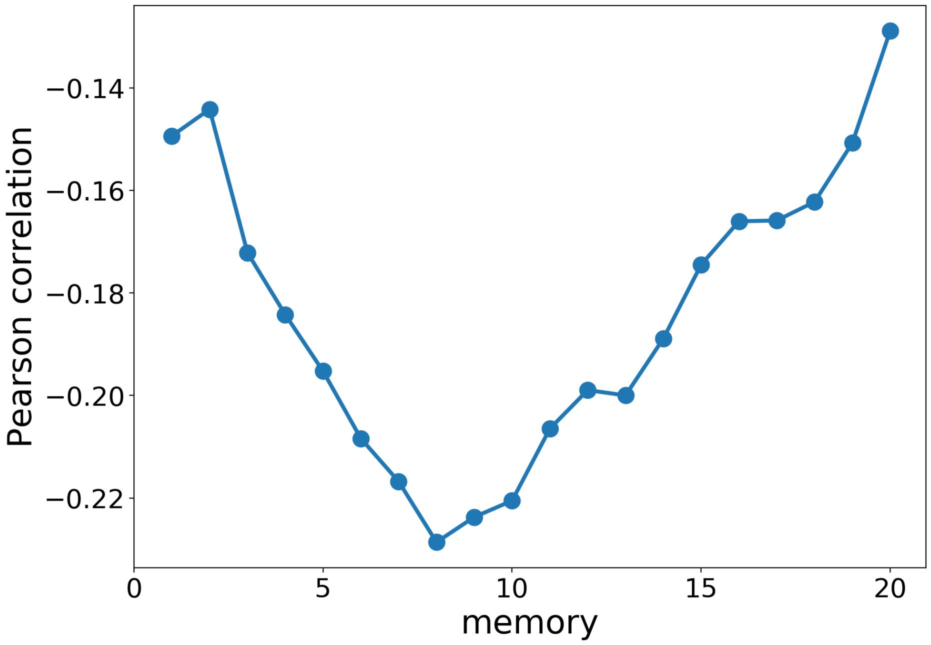


Figure 4.9: Pearson correlation between the time series of new investors buying Nokia for the first time and the moving average of logarithmic returns of Nokia in the investigated period as a function of the time window at which the moving average is computed.

has been designed. This model takes inspiration from the Deffuant model, an ABM developed at the beginning of the XXI century [82].

4.3.1 The Deffuant model

In a paper published in 2000, Deffuant et al. introduced a model to describe the opinion dynamics within a society. Their ABM followed a vein of similar models, most of whom were developed as a modified version of the Ising model, such as, for example, in [83]. The Deffuant model was the first attempt to describe an opinion not just as a binary variable, that can oscillates only between two states, but as a continuous variable that can take any value in the range [0,1]. Moreover, the model exploits networks in order to model actual social structures. Indeed, in the original version of the Deffuant model, social interactions are mimicked using a regular lattice. The moderately high and constant values of the local clustering coefficient assure that in a regular lattice the nodes are organized in locally correlated groups that are able to mimick the existence of cohesive group of friends. Thus, in a Deffuant model, after fixing an initial distribution of opinion that is assigned to all the nodes, the discussion between nodes is carried out by selecting random couples of nodes that are connected in the underlying network. Each single discussion works as follows: if the absolute difference between the opinions of the two nodes is below a given threshold ϵ, which is a parameter valid for all couples, then the two nodes end out with an opinion which is the average of their initial ones. Otherwise, they maintain their starting opinions. For high values of ϵ, consensus is reached in the lattice and almost every individual converge to the same opinion. For smaller values of ϵ, polarization of opinions in few groups or complete fragmentation are observed. In the present dissertation, the Deffuant model has provided the inspiration for an ABM capable to describe the dynamics of an inflating bubble. Indeed, in this context the opinions represent the reservation prices of the pool of potential investors. However, some effects which are peculiar to speculative bubble, such as an increasing interest in the involved assets and a constant increase of the reservation price itself have been added to the model, making it significantly different from the original Deffuant model. In fact, the former effect is related to the increasing focus that is put on an asset by media and society once its price performance becomes evident. The latter instead relates to the fact that investors are pushed to continuously update their reservation prices as a consequence of the increasing price of the asset. This model can be put in relation to that of Barberis et Xiong, [81]. Indeed, both focus on the implications of the disposition effect. However, the model of Barberis and Xiong is based on the predictions that investors make on the dynamics of the price when starting an investment, while the current model focuses on the effect of social dynamics on the rate of new investors.

4.3.2 A model for reservation prices

The model has been designed in order to reproduce the dynamics of people buying Nokia for the first time in the period from 01/01/1997 to 31/12/2000, which is the interval in which the price continued to grow before starting to significantly decrease. In the present model, small world networks [23] have been chosen in order to mimick social interaction among investors. Indeed, while maintaining high values of local clustering coefficient, small world networks are more realistic than regular lattice because they have a small diameter. In fact, in a small world network, given any couple of nodes, the shortest path that connect them will be relatively small. Specifically, a small world network with 4 neighbors and a rewiring probability of $p = 0.05$ has been chosen. The next step was to fix the initial reservation prices of all investors. This step was performed by extracting values from an exponential distribution with exponent $\gamma = P_0/4$, where P_0 is the real price of Nokia on the first day of the investigated period. A value $p^* = 0.95P_0$ was fixed as the threshold at which the distribution of starting reservation prices is cut. This step has been made in order to prevent investors from starting with a reservation price which is higher than the real price. The exponential distribution was chosen in order to describe an initial situation in which Nokia still obtained small interest by investors. Indeed most of the investors at the beginning have a reservation price close to 0, which can be interpreted as the total absence of interest in investing in the asset. After the initial setup, the dynamics of reservation price starts by iterating the same procedure several times. The time step t represents a day of trading activity. The sequence of actions for each time step is:

- **Update of the tolerance parameter.** Before the actual discussion between investors starts, the tolerance parameter ϵ is updated. This is a significant difference from the original Deffuant model. Indeed, in this context the tolerance parameter is not constant in time but evolves with respect to price dynamics. This feature has been included in order to reproduce the significant increases experienced in the time series of new entries. If the price grows significantly, investors are more willing to change their reservation price. Specifically, the tolerance parameter is updated according to $\epsilon_t = c_0 * P_t * (1 + 4\frac{\Delta P(t,\tau)}{P_{t-\tau}})$, where c_0 is a parameter that can range from 0 to 1, P_t is the price of Nokia at time t and $\Delta P(t, \tau)$ is the variation of price experienced in $[t - \tau, t]$. Thus, after a significant price increase ϵ_t can be equal to a significant fraction of the price of Nokia.

- **Discussion.** Once the tolerance is updated, all the investors that are not in the market yet select randomly one node from their neighbors and "discuss" their reservation price in couples. Here the discussion is different from what observed in the Deffuant model. If the difference between the two reservation prices is below the tolerance ϵ, the investor with the lowest reservation price takes the highest value, while the other investor leaves its reservation price untouched. Otherwise, nothing happens. This asymmetry in the discussion was introduced in order to describe the intense expectations of investors during the dynamics a bubble. In fact, the system is biased toward higher values because many investors expect a further increase in the price.

- **Entrance in the market.** After the discussion, the reservation prices are compared with the real price of Nokia at time t. At this point, the process of entering the market is then modelled in the following way: each investors has a probability $p = \alpha * p_{in}(t) + (1 - \alpha * p_{in}(t)) * p_r$ to enter the market, with $\alpha = 1$ if the reservation price of the investor is lower than the real price and 0 otherwise. In p, $p_{in}(t)$ is the probability to enter the market when the real price is lower than the reservation price. The probability $p_{in}(t)$ is a function of time and follows the same dynamics of the tolerance parameter ϵ_t: indeed, $p_{in}(t) = p_{in}(0) * (1 + 4\frac{\Delta P(t,\tau)}{P_{t-\tau}})$, with $p_{in}(0)$ a parameter fixed in the setup in the model. Thus, the probability of following the indication of reservation price is positively correlated with the recent history of the price of Nokia. This choice also contributes to shaping a more pronounced dynamics of people entering the market when the increase in price is stronger. Instead, p_r represents a random probability of entering at any time, without taking into account the reservation price. The probability p_r does not depend in time and it was introduced in order to reproduce the small waves of new investors that enter the marke in absence of big drops in the price.

- **Update of reservation prices.** The last action of each time step is to update the sequence of reservation prices of all investors according to the recent dynamics of price. Indeed, once the starting distribution of reservation prices is assigned when the model is set up, the dynamics of discussions among investors leaves the price with the superior bound represented by the price of Nokia at $t = 0$, P_0. Since the time series experiences a sharp overall increase during the investigated period, there is the need of regularly redefining the reservation prices towards the recent values. In order to do this, each $\tau/2$ time steps the reservation prices RP of all investors is updated according to $RP_i(t) = RP_i(t-1) * (1 + \frac{\Delta P(t,\tau/2)}{P_{t-\tau/2}})$, only when $\Delta P(t, \tau/2) > 0$.

The overall dynamics of the model is thus based on a key principle: the variation in the price of the asset strongly influence the intensity of the actions of investors. Indeed, sharp increases in the price of Nokia are reflected in a larger tolerance, making investors change more easily their reservation prices for higher ones, and in a larger probability of entering the market when the reservation price is higher than the real price. Table 4.3 reports a list of all the values assigned to the parameters used in the model. Fig. 4.10 plots the cumulative density functions (cdf) of new entries in the investigated period, both the real one and the one obtained by the model. Looking at the figure it is evident that, although the model cdf is less smooth than the real one, the model is able to produce a bursty time series with a profile similar to the one observed in the real process. Indeed, both curves appear as strongly convex, as a result of the combination of a first period in which the investors entered at low rates and a second period in which the dramatic rise in the price of the asset drove increasingly large waves of investors to enter the market, producing a steep curve. Thus, by modeling the entrance of investors in the market as the result of a process determined by the diffusion of reservation price, whose intensity grows with the price, it is possible to properly fit the real time series, characterized by peaks of increasing size.

Table 4.3: The table reports the parameters used to set up the model. N is the total number of investors of the model. Its choice does not affect the proportions of the final time series of investors, but only its absolute values. The parameter τ, which represent the time window at which the investors monitor the variation of price, was set to 21 days, which is the number of trading days in a month. The parameter c_0 represents the fraction of price which is set at the beginning as the tolerance threshold. The tolerance ϵ_t is proportional to c_0 but changes in time according to $\epsilon_t = c_0 * P_t * (1 + 4\frac{\Delta P(t,\tau)}{P_{t-\tau}})$. The parameters p_{in} and p_r are respectively the probability for an investor of entering the market when his/her reservation price is above the real price and the probability of randomly entering the market at any time. The latter was fixed to a value which is several orders of magnitude smaller than the first one, since in the model the dynamics of reservation price is the predominant one.

Parameter	Value
N	10000
τ	21
c_0	0.01
p_{in}	0.01
p_r	$2.5\ 10^{-5}$

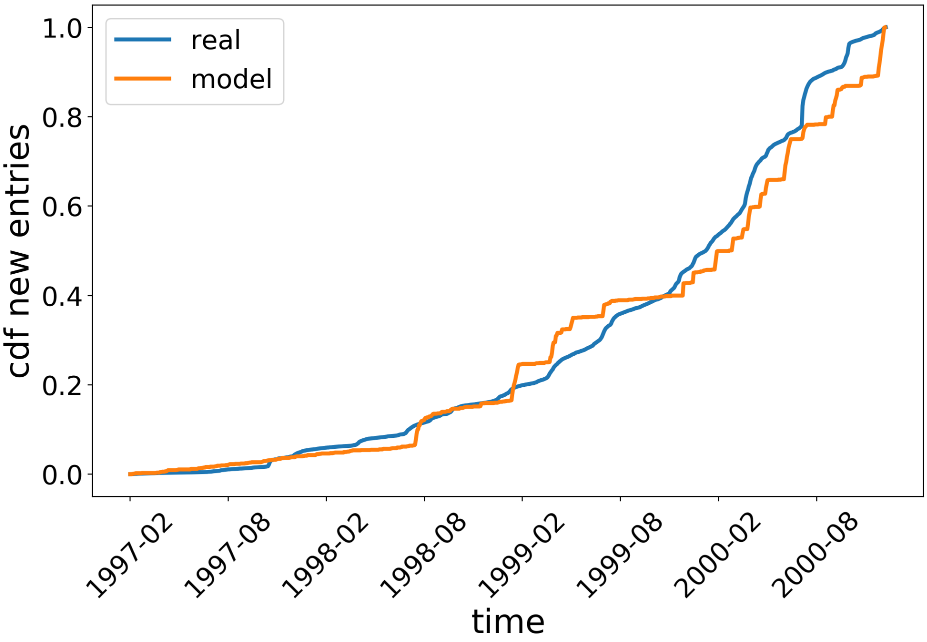


Figure 4.10: Cumulative density function of the time series of investors entering the market obtained in the real process (blue line) and by using the model (orange line).

5 Networked structure of trading networks and high frequency trading

A key principle in the operation of modern stock markets is guaranteeing a high degree of anonymity to investors. Indeed, each investor approaching the market for a transaction should not know who is going to be his/her counterparty. This is meant to minimize the amount of adverse selection, i.e. to minimize the release of information associated with the decision of an investor to trade. Thus, anonymity is a common fundamental characteristic of modern stock exchanges. It should ensure that financial markets are not networked markets. A market is networked when its participants have a constrained number of possibilities for potential counterparties. In fact, networked markets are markets with constrained intermediaries and/or markets where the underlying network of interactions is crucial for the outcome of the trading transaction (a classic example of networked market is the job market [84]). This type of interconnections should be avoided in stock markets in order to minimize the inhomogeneity in the distribution of information among investors, that usually has a large impact on the trading process [61]. However, the modern evolution of financial markets has created a new type of heterogeneity among investors, that can be be divided in two categories: the first is the category of investors having access to high frequency trading, whereas the second is the one without this access. In the present context, high frequency trading is intended as the act of operating in the market at high speed, on the basis of algorithmic computation and on time intervals below human reaction time.

In this chapter we investigate if and how the introduction of high frequency trading implicitly induces the establishment of preferential or avoiding networked relationships among market members. In what follows, the expression *market member* will be used to refer to the traders that are registered in the market. The registration allows them to access directly the market in order to perform a transaction.

The landscape of high frequency trading platforms and technologies is changing fast (nowadays, the completion time of a market activity of computer trading algorithms has reached the time scale of microseconds [85]), and the percentage of transactions performed in this way covers more than 50% of market transactions. The research question is whether the heterogeneity of strategies and performances of investors reinforces the presence of statistically persistent preferential/avoiding relationships between market members trading in a financial market. The presence of these statistically detectable relationships would therefore characterize a fully electronic anonymous financial market as a "de facto" networked market.

In order to properly study the dynamics of a stock market at such short time scales, the order book of the market has to be considered. An order book is an electronic register that contains information on each event occurring at the stock market. Specifically, it stores and continuously updates a list of all the orders that are placed by investors. The information contained in the order book can be very useful when trading, since knowing the price level and the type of the orders that are currently being placed can help in forecasting the direction in which the price of a given asset is going. The orders placed in the order book can belong to two different categories: i) limit orders if they express the will to buy (sell) a specified amount of shares at a given price. These orders stay in the order book until they are matched by ii) market orders, which express the will to buy (sell) a specified amount of shares at the current best ask (bid). The ask (bid) level is set by limit orders, and is the lowest (highest) amount of money that is being asked by other investors to buy (sell) a share. Each time a market order is issued, a transaction occurs between the placer of the market order (called *aggressor*) and the placer of the limit order (called *counterpart*).

The analysis of order books and high frequency trading is characterized by some specific aspects:

- **Large data size.** Since the time resolution of order books nowadays goes down to the level of microseconds, [86], the amount of data available for even a single day of trading can require a large amount of computer memory. This poses challenges on computational time and power.

- **Irregularity in the temporal spacing of events.** At this time scale, the events are not occurring on a time grid, and they are tracked one by one. This makes the sequence of events irregular with respect to temporal spacing, and provides the motivation to adopt an approach based on discrete temporal scales.

- **Intraday patterns.** When looking at the dynamics within a single day of trading, stationarity cannot be assumed to be valid anymore. Indeed there is evidence of intraday patterns, like the U-shaped form shown by volatility during the day [87].

- **Temporal correlation.** Another peculiar feature of the intraday dynamics of price is the presence of short lasting non negligible negative values of autocorrelation for the returns, the so-called *bid-ask bounce* [88], that again contradict the empirical evidence found at larger time scales.

5.1 Database

The data used for this investigation is taken from a database maintained by Nordic Nasdaq, a local subsidiary of Nasdaq that provides financial services and infrastructures in the Baltic area and in Scandinavia. The dataset consists of a coded data stream containing the unfolding of the order book of Nordic Stock Exchange, from February 2010 to December 2011. From this data stream it is possible to reconstruct the information about each order inserted in the book, together with the resulting transactions. In the present investigation only the activity of the venue of Stockholm has been considered. For each order, it is possible to extract the time at which it had been placed, with a resolution of one millisecond. Other available information is the asset the order is related to, its size and its price. The information on the identity of the issuer is not reported. Whenever a market order is placed, a transaction is reported, and the identity of the two market members is recorded. Moreover, from the information reported for the transaction it is possible to compute the *time to fill*, i.e. the time that passes between the placement of the limit order and its execution through a market order. It is worth noting that in this database the information about aggressor and counterpart is an exact information and, differently from other studies, it is not obtained by using a proxy associated with the presence of the transaction price at the best bid or best ask.

5.2 Summary statistics of orders and transactions

In the current investigation, orders and transactions have been investigated on a monthly basis before being analyzed. This implies that the analysis on the effects of high frequency trading has been conducted on 23 sets of data which contained all the orders and transactions performed in a calendar month (usually 21 trading days). Table 5.1 reports the number of transactions occurring each month. Additional information concerns the number of market members active on the market on each time span, the number of ISINs[1] for which at least one transaction was performed and the number of transactions performed for the most and the least traded ISINs. The table shows that, month by month, there are fluctuations in the numbers of market members, ISINs and transactions.

Table 5.2 reports the summary statistics related to the order being placed in the same time intervals. Specifically, it reports the overall number of orders, the number of orders which are deleted before their execution, the number of ISINs with at least one order, and the number of orders placed for the most recurring ISINs. For the orders, the information on market members is not available. It is worth noting that, for each month, at least 97% of orders are deleted without being fulfilled. Moreover, it can be checked that the number of ISINs present in this table does not match the corresponding column of Table 5.1. This happens because for some ISINs (mainly warrants), all the placed orders are deleted and no related transactions occur.

Whenever a transaction occurs, the corresponding *time to fill* can be computed. The time to fill can be seen as an indicator of the possibility of a market member to operate at high frequency scales. In fact, if one looks at the dynamics of transactions from the point of view of market members, the distribution of time to fill of the transactions in which a given market member operates as aggressor contains information on the time scale at

[1] ISIN is an acronym for International Securities Identification Number, and is a code that uniquely identifies a specific security. It may refer to stocks, warrants, bonds and all kinds of financial assets.

Table 5.1: Summary statistics of the transactions performed month by month at the Nordic Stock Exchange, venue of Stockholm. For each month, the number of overall transactions, the number of active market members, the number of traded ISINs and the number of transactions for the most traded ISIN are reported. The number of transactions for the least traded ISIN is always 1.

Month	Year	All transactions	Market Members	ISIN	Max ISIN
2	2010	2112543	75	1196	118745
3	2010	2932648	72	1258	153313
4	2010	3180196	72	1450	170222
5	2010	3966700	73	1468	161784
6	2010	3111806	72	1380	152399
7	2010	2886504	74	1467	149161
8	2010	2947415	74	1517	128933
9	2010	3134295	74	1491	143201
10	2010	3203359	76	1578	181046
11	2010	3065697	80	1513	134826
12	2010	2908710	79	1612	122183
1	2011	3279643	78	1742	173092
2	2011	3592409	80	1656	154490
3	2011	3741057	79	1503	167537
4	2011	2770656	83	1522	144484
5	2011	3512017	81	1545	154938
6	2011	3322234	79	1309	162460
7	2011	3159236	83	1435	166391
8	2011	6238368	83	1754	450914
9	2011	4951353	85	1375	309365
10	2011	4659666	86	1468	270822
11	2011	4613463	83	1313	235156
12	2011	3598750	87	1125	181707

Table 5.2: Summary statistics of the orders submitted month by month at the Nordic Stock Exchange, venue of Stockholm. For each month, the number of overall orders, the number of deleted ones, the number of ISINs for which at least one order was submitted and the number of orders for the most traded ISIN are reported. The number of orders for the least traded ISIN is always 1.

Month	Year	All orders	Deleted orders	ISIN	Max ISIN
02	2010	80183507	78471547	2539	3319910
03	2010	85750874	83341336	2699	2484291
04	2010	99788039	97262359	2594	2388965
05	2010	182443507	179296066	2951	6353257
06	2010	186214540	183639747	3138	4759539
07	2010	179279654	176804280	3040	4388838
08	2010	191142502	188630743	3307	4401413
09	2010	179201653	176548176	3184	4450367
10	2010	184497855	181859901	3358	4990035
11	2010	184198881	181616060	3598	5058352
12	2010	157258377	154860344	3889	4035948
01	2011	182295350	179620325	3763	6308408
02	2011	174137163	171190872	3831	6205200
03	2011	209725983	206641919	3554	5610651
04	2011	136269268	133962124	3669	4296711
05	2011	167620037	164725937	3870	4525160
06	2011	171414745	168708426	3775	5338413
07	2011	212321422	209715733	3777	6648121
08	2011	435489001	430384111	3708	12605615
09	2011	312579682	308491781	3721	7857166
10	2011	235694315	231917347	3437	6134586
11	2011	256873721	253095077	3168	7563479
12	2011	176806487	173673030	3314	5908463

which it is able to react to the updates of ask and bid inside the order book. Indeed, if an aggressor consistently places market orders few milliseconds after the corresponding limit order is inserted in the order book, this is a strong indicator that he is operating through algorithmic trading below human reaction time. Fig. 5.1 shows the pdf for the time to fill of the transactions occurring on January 2011. Similar pdfs have already been investigated for different stock markets in [89] and [90]. The x axis, whose scale is milliseconds, ranges from less than one to about $3{\cdot}10^7$, i.e. approximately the number of milliseconds contained in 8 hours, which is the duration of the opening of stock market at the Stockholm venue. The pdf is close to a power law, with the lowest values of time being by far the most frequent. Moreover, between ∼10 and ∼100 there is a local peak in the pdf. This is likely to be associated with the fixed time scales at which the trading algorithms are set to operate. Although Fig. 5.1 does not provide information on the distribution of time to fill at an individual level, a closer look to the individual behavior of the various market members reveals a pronounced heterogeneity. Fig. 5.2 shows the cumulative pdf of time to fill for two different market members, Citadel Securities Ltd (CDG) on the right and Svenska Handelsbanken AB (SHY) on the left. Data refers to January 2011. From the figure it is evident that these two market members operate at rather different time scales. Indeed, while about 10% of the transactions for which CDG operates as aggressor have a time to fill within 1 ms (and about 25% have a time to fill smaller than one second), SHY never operates as aggressor with a time to fill smaller than 100 ms, with far less than 1% of transactions with a time to fill smaller than ten second.

Fig. 5.3 shows the cumulative density function of the 5_{th} percentiles of time to fill for each market member active in January 2011. This figure spreads light on the extent of the heterogeneity of the system. In fact, the 5_{th} percentiles spans four orders of magnitude, from 10 to ∼10,000 ms. It is evident that about 40% of the market members are almost never active below a time to fill of 200 ms, which is a typical human reaction time.

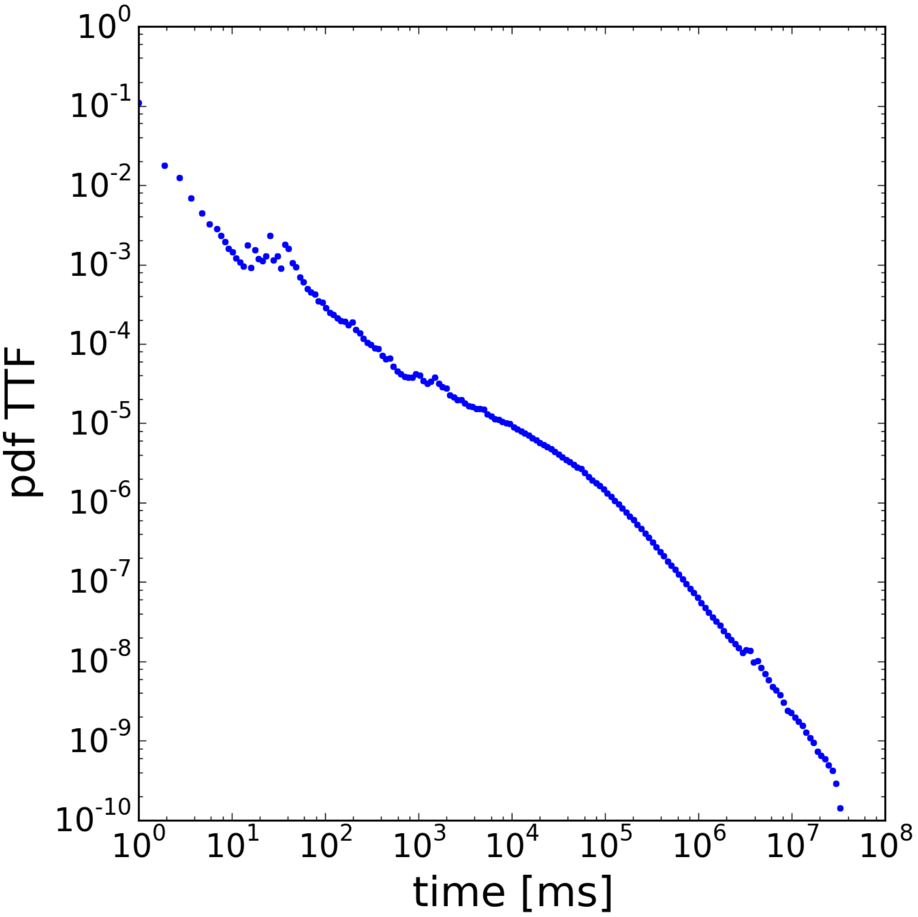


Figure 5.1: Probability density function of the time to fill of all transactions performed in January, 2011. A peak in the range [10,100] signals the potential presence of algorithmic trading at fixed time scales.

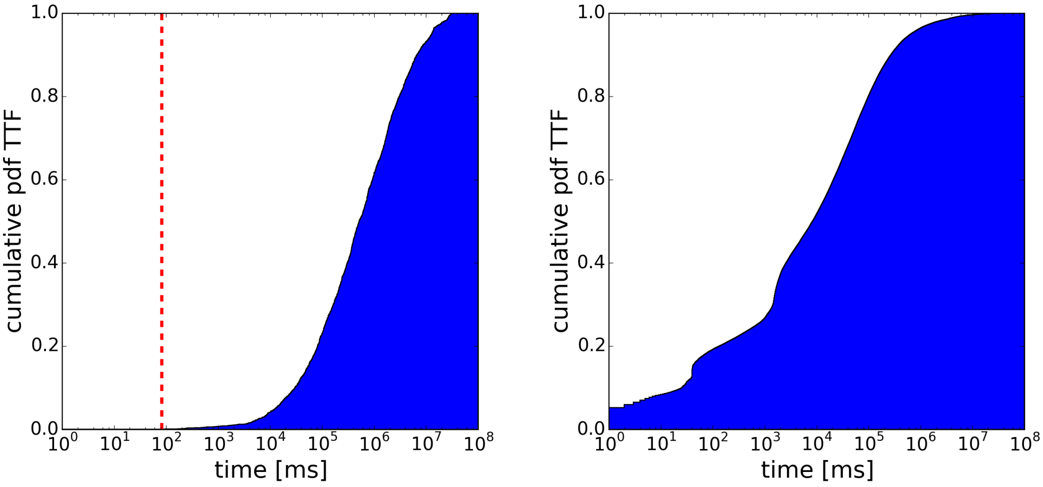


Figure 5.2: The left panel shows the cumulative distributions of the time to fill of the transactions in which the market member Svenska Handelsbanken AB (SHY) was involved as aggressor on January, 2011. The dashed red line points to the minimum value of the time to fill. The right panel shows the same information for the market member Citadel Securities Ltd (CDG). It is worth noting that in the investigated period the two market member were operating as aggressors at radically different time scales.

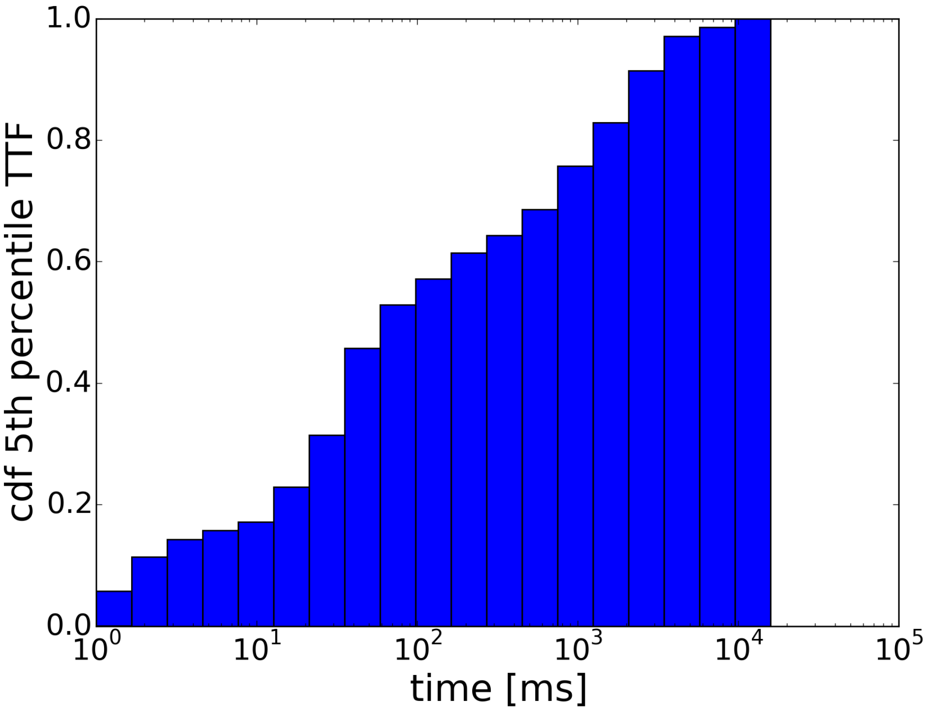


Figure 5.3: Plot of the cumulative density function of the 5_{th} percentiles of time to fill for each market member active in January 2011.

5.3 Statistical characterization of a networked market

The first step in this investigation is to detect whether there is empirical evidence of networking effects in the market. In order to do this, each couple of market members operating in the stock market has to be tested against a null hypothesis of random pairing that takes into account the heterogeneity of their activity. The test is designed as follows: given two market members A and B active on an ISIN i, the number of transactions N_A in which A is active as an aggressor with any counterpart and B is active as a counterpart N_B with any aggressor are computed. Then the number of transactions made by the ordered couple of market members (A, B), with A aggressor and B counterpart, N_{AB} is computed. At this point, given N the total number of transactions made on the ISIN i one can compute the probability that the N_{AB} transactions made by the couple (A, B) are compatible with a null hypothesis of random co-occurrence. The random pairing is intended as a draw. This implies that, according to the null hypothesis, for each transaction the aggressor select randomly its counterpart from the pool of all market members. A significant divergence from this random behavior would mean that the pairing of market members is biased, and would represent empirical evidence for a networked stock market. Thus, given the numbers N, N_A, N_B and N_{AB} one can compute two p-values, according to what seen in [79],

$$p_1(N_{A,B}) = 1 - \sum_{X=0}^{N_{A,B}-1} H(X|N, N_A, N_B), \tag{5.1}$$

$$p_2(N_{A,B}) = \sum_{X=0}^{N_{A,B}} H(X|N, N_A, N_B). \tag{5.2}$$

The p-value p_1 represents the probability that A and B operate together according to the null hypothesis N_{AB} times or more. If this probability is low, then the occurrence of the couple (A, B) is over-expressed with respect to the activity of A and B. The other p-value measures the opposite effect. In fact, p_2 is the probability that A and B operate together according to the null hypothesis N_{AB} times or less. If p_2 is low, it means that the couple (A, B) is under expressed and market members A and B are avoiding each other on the stock market. In both cases the null hypothesis is violated, and the trading pairing is not compatible with a random matching. In order to rigorously detect the over/under expressed couples of market members, all the p-values should be tested against a significance threshold, suitably corrected for multiple comparisons, as shown in the previous chapters. In order to detect whether high frequency trading plays a role in the networking process, one should first identify the couples of market members that operate significantly at high frequency time scales. A preliminary step in this direction is the identification of high frequency transactions. In order to discriminate between transactions associated to high frequency trading and the rest, in this analysis time to fill has been used. The classification works as follows: a temporal threshold t is chosen and all transactions whose time to fill is below t are labelled as high frequency ones. The threshold was set to 200 ms, which is a typical human reaction time, in order to be sure that trading activity could not be associated with a human decision. All the tests performed have been repeated trying different thresholds, ranging from 50 to 200 ms, and the results did not show significant changes. Fig. 5.4 plots the ratios of high frequency transactions for each of the 20 most traded ISINs. Each point represents the ratio averaged on the 20 ISINs, with the error bars spanning from the first to the last decile. A clear increasing trend is evident at the end of the investigated period, showing how the relevance of high frequency trading was steadily growing in late 2011.

In order to detect the couples of market members that operate significantly through high frequency trading a statistical test has been designed. The idea behind the test is to check whether each pair of market members is acting on a given stock homogeneously through both high frequency and "low" frequency trading. In fact, if this is not happening then the couple of market members is showing over or under expression with respect to the participation in trading labeled as of high frequency trading to a statistically significant extent. Thus, if N is the total number of transactions performed on an ISIN i, N_{AB} is the number of transactions made by the aggressor A and the counterpart B and N_{HF} is the number of transactions performed in high frequency mode on i by all market members, by computing $N_{AB,HF}$, i.e. the number of transactions performed by the couple (A, B) in high frequency mode, one can apply the formalism of over/under expressions and compute the p-values

$$p_1(N_{12}) = 1 - \sum_{X=0}^{N_{AB,HF}-1} H(X|N, N_{AB}, N_{HF}), \tag{5.3}$$

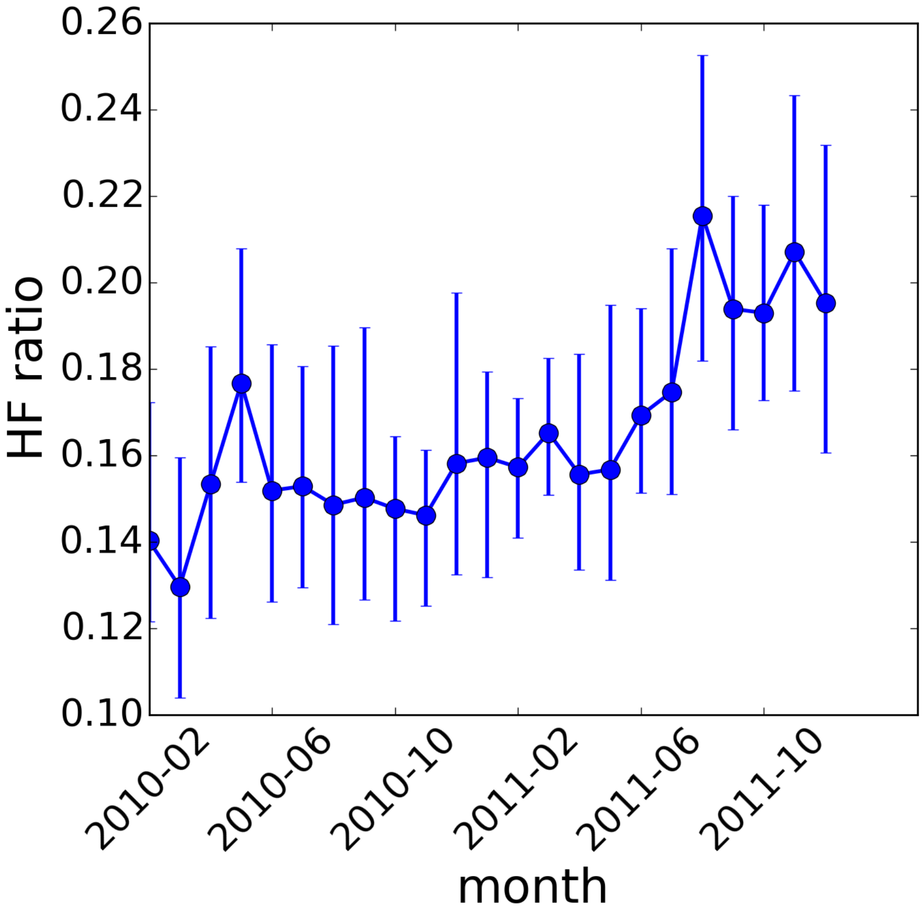


Figure 5.4: Plot of the fractions of high frequency transactions performed on the 20 most traded ISINs as a function of the trading month. Each point represents the average between the ISINs, with the error bars going from the first to the last decile.

$$p_2(N_{12}) = \sum_{X=0}^{N_{AB,HF}} H(X|N, N_{AB}, N_{HF}). \tag{5.4}$$

Testing p_1 and p_2 with a statistical threshold allows to detect over/under expressed couples. Here the over (under) expression of a couple represents the fact that their transaction are labeled more (less) in high frequency mode than what expected if the high frequency transactions were homogeneously distributed among all the couples taking into account their heterogeneous levels of activity.

Table 5.3 reports the outcome of the two tests on the 20 ISINs which were most traded at the venue of Stockholm in the period 2010-2011. The tests were performed on a monthly basis. The table reports the total number of couples which were active at least once on one of the 20 ISINs, and the number of couples with at least one over/under expression in the two tests. From the table it is evident that the number of couples over/under expressed in the test on the networking of the market is always larger than the corresponding amount for the test on high frequency trading. However, the information in Table 5.3 is not enough to assess whether the two phenomena are correlated. Fig. 5.5 plots the Jaccard coefficient between the couples of the two test, both for over (left panel) and under expressions (right panel). Each point represents the averaged Jaccard coefficient on the ISINs, with the error bars spanning from the first to the last decile. The figure shows that the Jaccard coefficient increases significantly during the last months of 2011. Moreover, this effect is more pronounced for over expressions. Even if this result does not imply causality between the two phenomena, it shows that the probability that a couple over/under expressed in its high frequency activity is also networking the market increases over time. Thus, although high frequency trading is probably not the unique factor that contributes to building a networked market, its impact grows over time, especially when establishing preferential relationships, represented by the over expressions.

Table 5.4 reports the outcome of the two tests from the perspective of the 25 most active market members, for January 2011. The table reports the number of times in which each market member appears in a over/under expressed couple in the two tests performed on the 20 most traded ISINs. The case in which the market members act as aggressors and the case in which they act as counterparts are considered as distinct cases. A detailed investigation of Table 5.4 confirms the connections between the networking of the market and the heterogeneous adoption of high frequency trading. Indeed, the Pearson coefficients between the sequence of over/under expressions as aggressor/counterpart in the two tests is always statistically significant, ranging from 0.51 (for the over expressions as counterpart) to 0.77 (for the over expressions as aggressor). This result indicates that the market members that have the higher number of reinforcing (avoiding) relationships in the stock market have a high probability of being strongly involved also in adopting high frequency trading more (less) than expected according to the null hypothesis, and viceversa. However, Table 5.4 spreads light also on the differences between the two phenomena. Indeed, when looking at the test on the networking of the market, the Pearson coefficient between over and under expression is high (0.62 for if one considers over and under expressions as aggressors and 0.74 for the case of counterparts), signaling that the process of networking is symmetrical: a market member that establishes an high number of enforcing relationships has an high probability of establishing also an high number of avoiding relationships, and viceversa. This is not true for the test on high frequency trading: in this case, the two sequences are anticorrelated for aggressors, with a Pearson coefficient of -0.31, and are not correlated for counterparts (Pearson coefficient of 0.05). This result can be seen as an other confirmation of the fact that, in the investigated period, high frequency trading was not homogeneously distributed among market members.

Table 5.3: Summary statistics of the couples operating on the 20 most traded ISINs of the Nordic Stock Exchange, venue of Stockholm, in the period February 2010 - January 2011. For each month, the table reports the total number of couples, the number of over and under expressed couples in the test on the networked structure of the market and the number of over and under expressed couples in the test on the adoption of high frequency trading.

Month	All couples	OE networked	UE networked	OE HF	UE HF
2010-02	5025	635	342	344	306
2010-03	4823	788	484	425	369
2010-04	4688	736	450	413	431
2010-05	4684	827	534	503	489
2010-06	4687	742	458	407	381
2010-07	4964	811	511	407	366
2010-08	4890	731	459	423	352
2010-09	4963	810	488	437	412
2010-10	4892	788	446	465	363
2010-11	5162	840	500	486	385
2010-12	5450	809	474	452	410
2011-01	5307	967	547	546	448
2011-02	5683	995	580	625	488
2011-03	5767	1036	646	597	541
2011-04	6137	940	547	537	411
2011-05	5973	959	648	604	513
2011-06	5849	897	617	513	503
2011-07	6136	958	646	515	484
2011-08	6288	1134	913	608	793
2011-09	6363	1186	930	699	696
2011-10	6301	1162	879	665	698
2011-11	6209	1163	968	656	699
2011-12	6337	928	710	539	513

Table 5.4: Summary statistics of the number of times in which the 25 most active market members are over/under expressed (OE/UE) in the two tests (networked/HF) in different roles (aggressor A/counterpart B).

MM	OE A networked	UE A networked	OE B networked	UE B networked	OE A HF	UE A HF	OE B HF	UE B HF
FORA	125	24	218	109	300	25	116	203
CSB	114	81	52	43	370	1	80	56
CDG	97	77	14	3	285	23	0	5
SWB	78	109	77	104	42	172	142	60
ENS	78	115	58	105	8	205	114	117
SRE	71	46	88	49	138	1	90	38
BRC	69	55	32	19	35	42	0	71
SHB	66	90	50	43	8	155	127	24
SAB	65	63	49	52	172	4	38	91
OHM	55	30	36	23	26	21	96	3
ABC	55	51	61	36	2	95	142	7
MSI	53	42	40	57	21	15	16	171
NIP	50	43	16	12	33	112	18	3
UBS	49	29	34	18	190	0	2	61
GEL	47	25	95	76	58	3	128	91
NDS	45	41	33	36	2	100	70	29
SGP	44	36	41	32	47	47	147	9
MAN	43	22	13	2	0	18	7	0
CAR	41	22	57	40	10	18	119	14
EPB	37	19	29	18	2	66	58	13
SGL	34	15	19	5	0	21	2	2
BPP	30	36	57	47	1	51	6	128
CAD	30	9	35	12	0	33	43	3
GSI	28	10	28	11	7	9	3	18
CDV	28	9	48	30	23	8	15	15

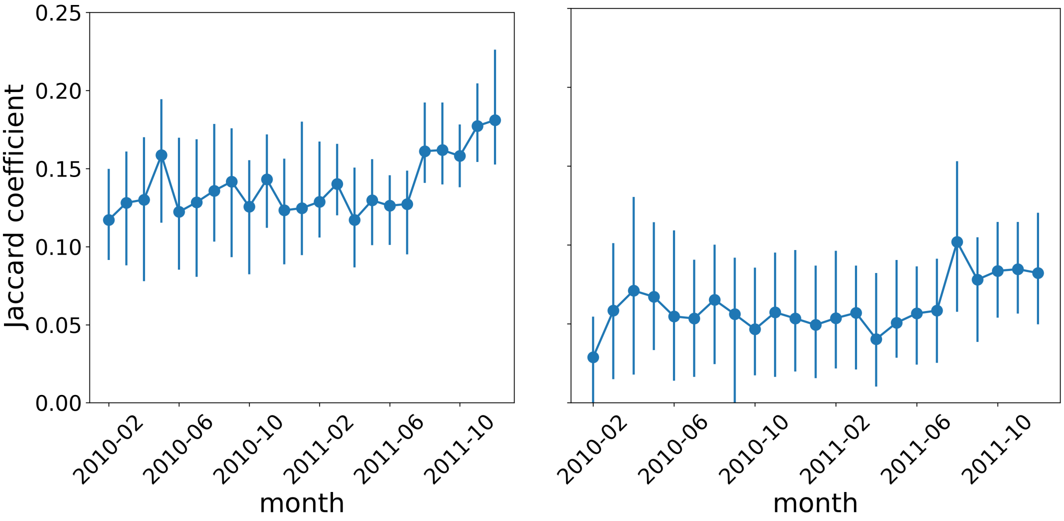


Figure 5.5: The left panel shows the Jaccard coefficient between the sets of couples over expressed in the test on the networked structure of the market and those over expressed in the test on the adoption of high frequency trading as a function of the trading month. The right panel shows the same plot for the under expressed couples.

6 Conclusions

This dissertation presents different statistical investigations of social and financial systems from Finland and Sweden. Specifically, the focus has been put on detecting the presence of significant deviations from the heterogeneity naturally present in such systems. Indeed, since heterogeneity is a peculiar and ubiquitous feature when dealing with complex systems, it should be properly taken into account when looking for statistically significant patterns in the activity of agents. In the context of the present dissertation, this result has been achieved by adopting and expanding the formalism of statistically validated networks (SVN). The approach of SVN is applied to bipartite networks in order to obtain a projection where only the links that are detected as the outcome of a non-random process are left. The random process is modeled as a random draw from the pool of all possible neighbors, whose size is equal to the degree of each agent. In this way the significance of each link is considered by taking into account the heterogeneous activity of the corresponding couple of agents.

In this thesis, the formalism of statistical validation has been applied to three different systems: i) a series of bipartite networks that describe the trading decisions in time of investors active at the venue of Helsinki of the Nordic Stock Exchange during the period 1995-2009, ii) the sets of investors entering the market to buy Nokia stock in the same venue during the evolution of the *dot com* speculative bubble and iii) the network of market members active at the order book of the venue of Stockholm of NSE in the years 2010 and 2011.

In the first case, statistical validation has been used to obtain clusters of investors characterized by similar trading patterns. Indeed, after applying Infomap to the SVN, a partition of clusters is obtained. This method detects partitions of clusters which are overlapping with the hierarchical structure of the system obtained from the dissimilarity matrix computed from vectors that describe the trading profile of a pair of investors. It is possible to combine the two methods to extract additional information from the data. In fact the method based on statistically validated network and community detection is giving results of high precision but of unknown accuracy. The combination of the approach of statistically validated networks with the correlation-based approach expands the set of information available about the clusters of investors providing an increase of the statistical accuracy at a minimal cost of decrease in terms of precision (i.e. in terms of increase of false positive). Thus it is proposed the use of a combination of the two methods when there are indications that the accuracy of the statistical validation network approach is too limited to properly describe the system of interest. Moreover, by expanding the formalism of SVN, the characterization of the clusters of investors with respect to the category they belong to has been performed. A different version of the same approach was used to detect the evolution of clusters of investors active on Nokia in different years. The combination of these methods leads to the observation of an ecology of groups of investors presenting a multiplicity of time scales that can last up to more than one decade. These groups are characterized by attributes that clearly distinguish them and their trading. The groups are often showing an over-expression of investors of specific categories, with this type of over-expression being observed over many years. Moreover, the similarity of investment actions between pairs of investors is exponentially sensitive to the level of market volatility. Further lines of research could focus on an extensive investigation of the factors that determine the different strategies followed by investors, such as technical analysis, release of news on the traded company, etc.

When dealing with the dynamics of the *dot com* speculative bubble, the statistical validation approach has been adopted in order to characterize the fluxes of household investors entering the market in order to buy the Nokia stock during the period 1995-2009, at a yearly time scale. The characterization is based on the census data collected by the Statistics Office of Finland, that includes information on age levels, education, income and job demography. In this context, the investigation has focused both on over and under expression of investors with specific attributes, in order to detect the categories which were involved in the bubble more and less than expected if taking into account the heterogeneity of new investors and of the Finnish population. The characterization on age levels reveals that people between 25 and 64 years are likely to be over expressed while younger and older investors tend to be under expressed. Moreover, there are significant cases in which an age level is under expressed only in the years of the most intense bubble dynamics (20-24 year cohort) or viceversa is over expressed in the same period (60-64 year cohort). Further investigation reveals that people with higher income and education are more likely to be over expressed as newcomers in the market, although this pattern is less pronounced in the years in which the bubble was inflating, as shown in the bottom panel of Fig. 4.8. Thus, our approach allows to detect the categories of investors which were more and less reactive during the dynamics of the bubble. A further line of research could focus on estimating the impact that factors such as focused marketing or the presence of cultural biases have in shaping such patterns of over/under expressions.

This dissertation presents also an attempt to model the process of entrance in the market of investors that are attracted by an asset that is experiencing a bubble-driven inflation. This result was obtained by defining a modified version of the Deffuant model that describes the spreading dynamics of reservation price. In order to reproduce the highly bursty time series of investors entering the market in the model, the intensity of the process depends on the recent history of the price of the asset. After incorporating this feature, the model is successful in reproducing the cumulative density function of new investors entering the market. Additional work should be done to develope a rigorous method to calibrate the model, and applying it to multiple assets, belonging or not to the technology sector, in order to see if a clear pattern or clusterization emerge from the sets of parameters of the different assets.

The investigation of the activity of market members active at the venue of Stockholm was focused on detecting the impact of high frequency trading on the process of networking of the market. Observing a networked market means detecting strong reinforcing/avoiding relationships between market members. In this context, the approach of statistical validation was first applied to the couples of market members active on a given ISIN, in order to detect whether the number of transactions performed by a couple is compatible with a random matching process that takes into account the global activity of its components. The results of this statistical test reveals that the market is networked. Indeed, for multiple ISINs there are several couples of market member which are over and under expressed. A similar test was employed to detect the couples of market members that are over/under expressed with respect to a trading activity characterized by the condition that at least one of the two market members is performing high frequency trading. The overlap between the two phenomena proved to be significant and increasing during the investigated period, especially when looking at the patterns of over expressed couples. Thus, algorithmic high frequency trading played a significant and increasing role in networking the market, although it was also proved that it cannot be considered as the only cause of the networking phenomenon. The current investigation has been performed considering each traded ISIN separately. A potentially more complete approach should involve the analysis of the multiplex networks obtained by putting together the over/under expressed couples detected on different ISINs.

Bibliography

[1] L. Euler. Solutio problemat is ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 8:128–140, 1741.

[2] G. Alexanderson. Euler and königsberg's bridges: a historical view. *Bulletin of the American Mathematical Society*, 43:567, 2006.

[3] J.L. Moreno. *Who shall survive? A new approach to the problem of human interrelations.* Nervous and Mental Disease Publishing Company, Washington, 1934.

[4] G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8), 4 1965.

[5] H. Helbing and E. Pournaras. Society: Build digital democracy. *Nature*, 527:33–34, 2015.

[6] R. N. Mantegna and H. E. Stanley. *An Introduction to Econophysics: Correlations and Complexity in Finance.* Cambridge University Press, New York, NY, USA, 2000.

[7] J. P. Bouchaud. Economics need a scientific revolution. *Nature*, 455, 2008.

[8] Tumminello M., Miccichè S., Lillo F., Piilo J., and Mantegna R.N. Statistically validated networks in bipartite complex systems. *PLOS ONE*, 2011.

[9] J. D. Farmer. Market force, ecology and evolution. *Industrial and Corporate Change*, 11(5):895–953, 2002.

[10] E. D. Adrian. The all-or-none principle in nerve. *The Journal of Physiology*, 47(6):460–474, 1914.

[11] P. W. Anderson. More is different. *Science*, 177(4047):393–396, 1972.

[12] M. Newman. *Networks: An Introduction.* Oxford University Press, Inc., New York, NY, USA, 2010.

[13] A.L. Barabási and M. Pósfai. *Network science.* Cambridge University Press, Cambridge, 2016.

[14] H. Ebel, L. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Phys. Rev. E*, 66:035103, 9 2002.

[15] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[16] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews. Neuroscience*, 10(3):186–198, 3 2009.

[17] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *SIGCOMM Comput. Commun. Rev.*, 29(4):251–262, 1999.

[18] L. Marotta, S. Miccichè, Y. Fujiwara, H. Iyetomi, H. Aoyama, M. Gallegati, and R. N. Mantegna. Backbone of credit relationships in the japanese credit market. *EPJ Data Science*, 5(1), 2016.

[19] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger. Network topology generators: Degree-based vs. structural. *SIGCOMM Comput. Commun. Rev.*, 32(4):147–159, 2002.

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

[21] K. Suchecki, V. M. Eguíluz, and M. San Miguel. Voter model dynamics in complex networks: Role of dimensionality, disorder, and degree distribution. *Phys. Rev. E*, 72:036132, Sep 2005.

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

[23] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):409–10, 1998.

[24] Vilfredo Pareto. *Cours d'Economie Politique.* Droz, Genève, 1896.

[25] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, October 2000.

[26] A.L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[27] B. Corominas-Murtra, R. Hanel, and S. Thurner. Sample space reducing cascading processes produce the full spectrum of scaling exponents. *Scientific Reports*, 7:11223, 2017.

[28] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323–351, 2005.

[29] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.

[30] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review*, E 69(026113), 2004.

[31] B.H. Good, Y.A. De Montjoye, and A. Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106, 2010.

[32] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[33] F. Battiston, V. Nicosia, and V. Latora. Structural measures for multiplex networks. *Physical Review E*, 89(3):032804, 2014.

[34] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral. Module identification in bipartite and directed networks, 2007.

[35] G. N. Gilbert. *Agent-based models*. Quantitative applications in the social sciences. Sage, Los Angeles, CA, 2008.

[36] E. Ising. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, February 1925.

[37] K. Sznajd-Weron and J. Sznajd. Opinion evolution in closed community. *International Journal of Modern Physics C*, 11:1157, 2000.

[38] A. Hosseiny, M. Bahrami, A. Palestrini, and M. Gallegati. Metastable features of economic networks and responses to exogenous shocks. *PLOS ONE*, 2016.

[39] Thomas C. Schelling. Models of segregation. *The American Economic Review*, 59(2):488–493, 1969.

[40] W. B. Arthur. Complexity in economic theory: Inductive reasoning and bounded rationality. *The American Economic Review*, 84(2):406–411, May 1994.

[41] Challet D. and Zhang Y. C. Emergence of cooperation and organization in an evolutionary game. *Physica A: Statistical Mechanics and its Applications*, 246(3):407 – 418, 1997.

[42] E. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417, 1970.

[43] R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

[44] G. Gan, C. Ma, and J. Wu. *Data clustering - theory, algorithms, and applications.* SIAM, 2007.

[45] P. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining.* Addison Wesley, 5 2005.

[46] Musciotto F., Marotta L., Miccichè S., Piilo J., and Mantegna R. N. Patterns of trading profiles at the nordic stock exchange. a correlation-based approach. *Chaos, Solitons & Fractals*, 88(Supplement C):267 – 278, 2016.

[47] R. N. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B - Condensed Matter and Complex Systems*, 11(1):193–197, Sep 1999.

[48] M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences USA*, 102:10421–10426, 2005.

[49] M. Tumminello, C. Coronnello, F. Lillo, S. Miccichè, and R. N. Mantegna. Spanning Trees and Bootstrap Reliability Estimation in Correlation-Based Networks. *International Journal of Bifurcation and Chaos*, 17:2319–2329, 2007.

[50] R. G. Miller. *Simultaneous Statistical Inference.* New York: Springer-Verlag, 1981.

[51] Benjamini Y. and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B*, pages 289–300, 1995.

[52] M. Tumminello, S. Micciché, F. Lillo, J. Varho, J. Piilo, and R. N. Mantegna. Community characterization of heterogeneous complex systems. *Journal of Statistical Mechanics: Theory and Experiment*, (01):P01019, 2011.

[53] L. Marotta, S. Miccichè, Y. Fujiwara, H. Iyetomi, H. Aoyama, M. Gallegati, and R. N. Mantegna. Bank-firm credit network in japan. an analysis of a bipartite network. 2014.

[54] C. Bongiorno, A. London, S. Miccichè, and R. N. Mantegna. Core of communities in bipartite networks. *Phys. Rev. E*, 96:022321, 2017.

[55] A. W. Lo. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *The Journal of Portfolio Management*, 2004.

[56] H. A. Simon. *Administrative behavior*. Free Press, New York, 1947.

[57] H. A. Simon. *Utility and Probability*. The New Palgrave, London, 1990.

[58] J. A. Frankel and K. A. Froot. Chartists, Fundamentalists, and Trading in the Foreign Exchange Market. *The American Economic Review*, 80(2):181–185, 1990.

[59] K. C. Chan. On the contrarian investment strategy. *The Journal of Business*, 61(2):147–63, 1988.

[60] K. C. Chan, N. Jegadeesh, and J. Lakonishok. Momentum strategies. *Journal of Finance*, 51(5):1681–1713, 1996.

[61] S. Grossman and J. Stiglitz. Information and competitive price systems. *American Economic Review*, 66(2):246–53, 1976.

[62] B. Barber, Y. Lee, Y. Liu, and T. Odean. Just how much do individual investors lose by trading? *Review of Financial Studies*, 22(2):609–632, 2009.

[63] M. Grinblatt and M. Keloharju. The investment behavior and performance of various investor types: a study of finland's unique data set. *Journal of Financial Economics*, 55(1):43–67, 2000.

[64] M. Grinblatt and M. Keloharju. Sensation seeking, overconfidence, and trading activity. *The Journal of Finance*, 64(1):549–578, 2009.

[65] Tumminello M., Lillo F., Piilo J., and Mantegna R.N. Identification of clusters of investors from their real trading activity in a financial market. *New Journal of Physics*, 1(14):013041, 2012.

[66] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[67] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, December 1985.

[68] J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, August 2002.

[69] R. J. Shiller. *Irrational Exuberance.* Princeton University Press, stu - student edition, 3 edition, 2015.

[70] C. Mackay. *Memoirs of Extraordinary Popular Delusions and the Madness of Crowds.* Office of the National Illustrated Library, London, 1852.

[71] P. Garber. Famous first bubbles. *Journal of Economic Perspectives*, 4(2):35–54, 1990.

[72] P. Phillips, Y. Wu, and J. Yu. Explosive behavior in the 1990s nasdaq: when did exuberance escalate asset values? *International Economic Review*, 52(1):201–226, 2011.

[73] P. Phillips and J. Yu. Dating the timeline of financial bubbles during the subprime crisis. *Quantitative Economics*, 2(3):455–491, 2011.

[74] Protter P. Mathematical models of bubbles. *Quantitative Finance Letters*, 4(1):10–13, 2016.

[75] Andersen J. V. and Sornette D. Fearless versus fearful speculative financial bubbles. *Physica A: Statistical Mechanics and its Applications*, 337(3):565 – 585, 2004.

[76] G. Caginalp, D. Porter, and V. Smith. Financial bubbles: Excess cash, momentum, and incomplete information. *Journal of Psychology and Financial Markets*, 2(2):80–99, 2001.

[77] T. Lux. Herd behaviour, bubbles and crashes. *Economic Journal*, 105(431):881–96, 1995.

[78] M. Grinblatt, M. Keloharju, and J. T. Linnainmaa. Iq, trading behavior, and performance. *Journal of Financial Economics*, 104(2):339–362, 2012.

[79] V. Hatzopoulos, G. Iori, R. Mantegna, S. Miccichè, and M. Tumminello. Quantifying preferential trading in the e-mid interbank market. *Quantitative Finance*, 15(4):693–710, 2015.

[80] L. A. Goodman and W. H. Kruskal. *Measures of Association for Cross Classifications*, pages 2–34. Springer New York, New York, NY, 1979.

[81] N. Barberis and W. Xiong. Realization utility. *Journal of Financial Economics*, 104(2):251–271, 2012.

[82] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 03(01n04):87–98, 2000.

[83] Katarzyna Sznajd-Weron and Jozef Sznajd. Opinion evolution in closed community, 2000.

[84] A. Calvo-Armengol and M. O. Jackson. The effects of social networks on employment and inequality. *The American Economic Review*, 94(3):426–454, 2004.

[85] O. Kaya. High frequency trading - reaching the limits. *Deutsche Bank Research*, 2016.

[86] S. DeCovny. Microseconds under a microscope. *CFA Institute*, 25(4), 2014.

[87] T. Andersen, T. Bollerslev, and J. Cai. Intraday and interday volatility in the japanese stock market. *Journal of International Financial Markets, Institutions and Money*, 10(2):107–130, 2000.

[88] J.Y. Campbell, A.W.C. Lo, and A.C. MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, 1997.

[89] D. Challet and R. Stinchcombe. Analyzing and modeling 1+1d markets. *Physica A: Statistical Mechanics and its Applications*, 300(1):285–299, 2001.

[90] Z. Eisler, J. Kertesz, F. Lillo, and R. N. Mantegna. Diffusive behavior and the modeling of characteristic times in limit order executions. *Quantitative Finance*, 9(5):547–563, 2009.