Comportement des traders institutionnels et microstructure des marchés : une approche big data

Kevin Primicerio

université PARIS-SACLAY

CentraleSupélec

NNT : 2018SACLC036

Thèse de doctorat

Large-trader behaviour and market microstructure: a big data approach

Thèse de doctorat de l'Université Paris-Saclay
préparée à CentraleSupélec

École doctorale n°573 Interfaces
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Gif-sur-Yvette, le 20 juin 2018, par

Kevin Primicerio

Composition du jury :

Frédéric Abergel CentraleSupélec, Université Paris-Saclay	Président du jury
Nils Bertschinger Frankfurt Institute for Advanced Studies	Examinateur
Fabio Caccioli University College London	Rapporteur
Damien Challet CentraleSupélec, Université Paris-Saclay	Directeur de thèse
Sophie Laruelle Université Paris-Est Créteil	Examinatrice
Fabrizio Lillo Università di Bologna	Rapporteur

Remerciements / Acknowledgements

L'écriture de ces remerciements me fait réaliser à quel point ces quelques années, passées si vite, ont été riches en expériences académiques et humaines. Je souhaite remercier nominativement les personnes ayant contribué à ce travail.

Je tiens tout d'abord à remercier Damien Challet. Merci pour ton aide inestimable, ton temps et ta disponibilité, à la fois par écran interposé (plus de 2500 mails) et au labo. C'est grâce à toi que cette thèse fût aussi enrichissante, merci de m'avoir partagé tes qualités de chercheurs et ta curiosité scientifique, et surtout ta bonne humeur et ton enthousiasme.

Je remercie Frédéric Abergel pour sa disponibilité, nos discussions, et d'avoir fait en sorte que les conditions de travail dans notre équipe soient agréables. Merci également d'avoir accepté d'honorer ma soutenance.

J'adresse un remerciement particulier à Nakahiro Yoshida et Thomas Seligman pour avoir porté intérêt à mon travail, et de m'avoir invité et accueilli à University of Tokyo, et à Universidad Nacional Autónoma de México respectivement. Je remercie Gregory Schehr pour m'avoir donné l'opportunité d'enseigner. Merci à Fabio Caccioli et Fabrizio Lillo pour la relecture et l'évaluation de mon travail en tant que rapporteurs, ainsi qu'à Sophie Laruelle et Nils Bertschinger pour avoir immédiatement accepté de faire partie du jury de cette thèse.

Merci à tous les chercheurs, doctorants et post-docs du laboratoire MICS qui ont fait de cette thèse, avant tout, une aventure humaine et en particulier, à mes anciens camarades de laboratoire Mehdi Lallouache, João Da Gama Batista et Stanislao Gualdi avec qui j'ai eu le plaisir et la chance de débuter cette aventure.

Merci à Sylvie Dervin pour sa disponibilité et sa bonne humeur, et Dany pour son aide avec les aléas informatiques.

Je tiens finalement à remercier mes amis, ma famille, ma soeur, et en particulier mes parents, à qui je dédie ce travail, qui m'ont toujours soutenu mais qui ont, surtout, mis un point d'honneur à me laisser vivre.

Résumé

Cette thèse est composée de quatre chapitres.

Le premier chapitre est une description préliminaire de la base de données Factset Ownership. Nous en donnons une description statistique et exposons quelques faits stylisés caractérisant notamment la structure du portefeuille des institutions financières et fonds d'investissement, ainsi que la capitalisation boursière des entreprises y étant recensées.

Le second chapitre propose une méthode d'évaluation statistique de la similarité entre des paires de portefeuilles d'institutions financières. Une paire statistiquement significative donnant lieu à la création d'un lien de similarité entre ces deux entités, nous sommes en mesure de projeter un réseau à l'origine bi-partite (entre institutions financières et entreprises) en un réseau mono-partite (entre institutions uniquement) afin d'en étudier l'évolution de sa structure au cours du temps. En effet, d'un point de vue économique, il est suspecté que les motifs d'investissements similaires constituent un facteur de risque important de contagion financière pouvant être à l'origine de banqueroutes aux conséquences systémiques significatives.

Le troisième chapitre s'intéresse aux comportements collectifs des gestionnaires de fonds d'investissement et, en particulier, à la manière dont la structure du portefeuille de ces fonds prend en compte, en moyenne, de façon optimale les frais de transaction en présence de faibles contraintes d'investissements. Ce phénomène où, dans de nombreuses situations, la médiane ou la moyenne des estimations d'un groupe de personnes est étonnamment proche de la valeur réelle, est connu sous le nom de *sagesse de la foule*.

Le quatrième chapitre est consacré à l'étude simultanée de données de marché. Nous utilisons plus de 6.7 milliards de trades de la base de données Thomson-Reuters Tick History, et de données de portefeuille de la base FactSet Ownership. Nous étudions la dynamique tick-à-tick du carnet d'ordres ainsi que l'action aggrégée, c'est-à-dire sur une échelle de temps bien plus grande, des fonds d'investissement. Nous montrons notamment que la mémoire longue du signe des ordres au marché est bien plus courte en présence de l'action, absolue ou directionnelle, des fonds d'investissement. Réciproquement nous expliquons dans quelle mesure une action caractérisée par une

mémoire faible est sujette à du trading directionnel provenant de l'action des fonds d'investissement.

Abstract

The thesis is divided into four parts.

Part I introduces and provides a technical description of the FactSet Ownership dataset together with some preliminary statistics and a set of stylized facts emerging from the portfolio structure of large financial institutions, and from the capitalization of recorded securities.

Part II proposes a method to assess the statistical significance of the overlap between pairs of heterogeneously diversified portfolios. This method is then applied to public assets ownership data reported by financial institutions in order to infer statistically robust links between the portfolios of financial institutions based on similar patterns of investment. From an economic point of view, it is suspected that the overlapping holding of financial institution is an important channel for financial contagion with the potential to trigger fire sales and thus severe losses at a systemic level.

Part III investigates the collective behaviour of fund manager and, in particular, how the average portfolio structure of institutional investors optimally accounts for transactions costs when investment constraints are weak. The collective ability of a crowd to accurately estimate an unknown quantity is known as the *Wisdom of the Crowd*. In many situation, the median or average estimate of a group of unrelated individuals is surprisingly close to the true value.

In Part IV, we use more than 6.7 billions of trades from the Thomson-Reuters Tick History database and the ownership data from FactSet. We show how the tick-by-tick dynamics of limit order book data depends on the aggregate actions of large funds acting on much larger time scale. In particular, we find that the well-established long memory of marker order signs is markedly weaker when large investment funds trade in a markedly directional way or when their aggregate participation ratio is large. Conversely, we investigate to what respect an asset with a weak memory experiences direction trading from large funds.

Contents

Introduction

In 1946, W. H. Auden published a poem with a line of stern advice:

> "Thou shalt not sit with statisticians nor commit a social science"

For a long time, even high-ranking decision makers seemed to concur, preferring to base their choices on intuition, personal experiences, and anecdotes. Those days are gone. They have been replaced by an era of "evidence-based decision making" in the major institutions of society: business, government, education, defence, sports. That new era prizes information from big data analysts and behavioural scientists.

The data revolution of the past decade is likely to have a further and profound effect on most organisations. Increasingly, they realise that capturing all the data streams of their applications is not enough, because the collection of data is a cheap procedure, and are looking for new ways to apply statistics in order to maximise the value from their existing businesses while creating new revenue streams. This new era of data is giving rise to new opportunities (e.g. enhance user experience, productivity and sales, or improve fraud detection) and challenges (e.g. analysing both structured and unstructured data). It also leads to the parallel emergence of two major trends in computer systems which are the growth in high-performance computing and big data. The convergence of simulations, big data analytics and high performance computing requires to consider multiple aspects (i.e. hardware, software, algorithm, models, applications) and implies large changes in the IT infrastructures of most companies. Therefore transforming the way industries operate and compete (Fox et al., 2015).

Unlocking the value of the data that already exist inside a company is not an easy task. Big data is not a substitute for common sense or careful research designs: its value does not lie in its quantity but in its processing and analysis. In other words, data volume represents a challenge in term of data analysis and scalability. Therefore appropriate data management and programming capabilities, as well as designing

creative and scalable approaches to summarise, describe and analyse large-scale and relatively unstructured data sets represent the new challenges of big data.

Although data volume is a direct challenge in term of data analysis and scalability, the flip side of volume, which is dimensionality, also needs to be addressed (Zhai et al., 2014). There are statistical tools (e.g. projections, feature selection, fitting) that allow us to reduce the dimensionality of the data and therefore play a crucial role increasing the performance of an algorithm (Xue et al., 2016). However, because of the potential large search space and the feature interaction problems, high dimensional spaces (i.e. with a dimension larger than 2) can be puzzling and the saying *"desperate times call for desperate measures"* never seemed more appropriate. Statistical tools are not a substitute for common sense, and, as Challet (2016) points out, one must learn to place itself in the right space: a space that encompasses the right features.

Financial markets are prominent example of a real-world application that generates a tremendous amount of data with prices that are recorded down to a nano-second resolution. They play a crucial role in the stability and growth of the global economy (Campbell et al., 1997) and have two major functions: providing liquidity and incorporating new information in prices. Financial institutions and organisations, who were precursors in that sector, were probably the first to realise it was an opportunity for better forecasting, and even nowcasting (Giannone et al., 2008). They have widely adopted big data analytics to inform better investment decisions with consistent returns.

Context

Financial engineering refers to the use of tools coming from various fields, such as applied mathematics, economics or computer science, in order to solve a financial problem. It has evolved through time and with technology. Aristotle (350bc) captured one of its earliest example when he related how Thales of Miletus, in ancient Greece, demonstrated that, as a philosopher, he was poor by choice, and not by inability:

> *"He [Thales] was reproached for his poverty, which was supposed to show that philosophy was of no use. According to the story, he knew by his skill in the stars while it was yet winter that there would be a great harvest of olives in the coming year; so, having little money, he gave deposits for the use of all olive-presses in Chios and Miletus, which he hired at a low price because no one bid against him. When the harvest-time came, and many were wanted all at once and of a sudden, he let them out at any rate he pleased, and made a quantity of money. Thus he*

showed the world that philosophers can easily be rich if they like, but that their ambition is of another sort."

Thales used his understanding of meteorology to predict a coming great harvest of olives and to make a significant profit. Today financial engineering uses statistics, stochastics, simulations, data mining and analytics. However it was only in 1900 that a first milestone in quantitative finance was notched with the publication of Louis Bachelier PhD thesis, *Theory of Speculation* (Bachelier, 1900). His thesis is credited to be the first time advanced mathematics was used in the study of finance. He was convinced that the financial markets were a rich source of data for mathematicians and introduced the concept of random walks, that predated Einstein's study on Brownian motion by five years, to approximate asset price's volatile path and calculate option prices. However sixty years were to pass before someone, in the name of Andrey Kolmogorov, took the slightest notice of his work (Weatherall, 2013).

Until the end of the 20th century, technical analysis prospered because of one reason: speed. The trading complexity was considerably lower than it is today. We were unable, at the time, to transmit information quickly, and it limited the number of trades and therefore the pace at which information was incorporated into prices. Therefore mathematical calculus was considered the most powerful tool for describing and understanding the general quantitative relations between the fundamental variables on which the theory was based (Akyıldırım and Soner, 2014).

As organisations grew in size, the tried-and-true methods of managing portfolios that prospered up to that point in time were becoming ill suited for the management of vast sums that accrued to institutions as years went by. Investment objectives, diversification patterns and trading strategies had to be revised. Although portfolio selection was not uncharted territory anymore, thanks to the Modern Portfolio Theory (MPT) introduced by Harry Markowitz (Markowitz, 1952), that formed the foundation for all subsequent theories on how risk can be quantified, investors still struggled in applying new recent concepts. In theory, MPT could help risk-averse investors construct and maximise portfolio expected return while exposing themselves to a sustainable level of risk. However computing the efficient frontier was a time-consuming analytical procedure. In addition, the cost of using computers was prohibitive for all but the very largest investment organisations.

As a consequence, technical and fundamental analyses continued to coexist through much of the 20th century, but the obvious aspect of competition was speed. Whoever was able to run a model the fastest was the first to identify and then to trade upon a market inefficiency capturing the biggest gain. To increase trading speed, make, and execute trading decisions, traders needed faster computers.

Since then, the technology of investing has moved into the space age: computers at every stage of the process. Electronic trading made low-cost brokerages and the NASDAQ possible, trading an entire (and large) portfolio instead of individual stocks. Processing speed meant more accurate data at a faster pace, allowing quants to perform simulations and to stop relying only on mathematical analysis. Everything changed and accelerated: computationally expensive operations such as the simulation of millions of scenarios for the calculation of sensitivities and valuation adjustments, for instance, were and still are daily struggles for a derivatives desk (Cesa, 2017).

Moreover, with the democratisation of electronic exchanges, virtually every economic transaction is recorded down to time scales of nanoseconds. Larger and larger amount of data is collected. The manipulation of large datasets, in principle, provides benchmarks for assessing whether expectations are realistic or fanciful and whether risks make sense or are foolish. This tremendous growth of financial data has meant that firms needed to find ways to acquire, distribute and utilise it. Therefore data consumption is expanding at an accelerating rate as financial institutions increasingly take on more data to conduct analytics, comply with regulation and demonstrate best execution.

Ease of access to tremendous quantities of data on financial market, especially high frequency data, opens the way for new methodologies both in statistics and computer science. Because finance was becoming an empirical science (Bouchaud, 2002), and the fact that financial markets are remarkably well-defined complex systems (Mantegna and Stanley, 1999), opened the door to physics-based approach for understanding financial and economic phenomena. Indeed facts such as these make financial markets extremely attractive for researchers interested in developing a deeper understanding of modelling of complex systems and allowed not only to develop advanced mathematical models that often had little to do with the traditional old-school fundamental and technical thinking but also to uncover universal phenomena that are invariant across systems (stylized facts) and a new way of looking at high-frequency data.

Similar changes have been witnessed in economics. Until the mid-1980s, the majority of papers were theoretical, the remainder relied mainly on "ready-made" data from government statistics or surveys (Hamermesh, 2013), or involved hunting through the library or manually extracting statistics from trade publication in order to gather data on a specific industry. As a result, fundamental analysis, that is, trading on expectation that the price will move according to the level predicted using publicly available information (e.g. cash flows, inflation and trade balances) and economic equilibrium theories it derives the equilibrium price levels developed through the 20th century (known as fundamental analysis). That shift also happened together with the expansion of available data. Apart from simply having more observations and more recorded data in each observation, several features differentiate modern data sets from many used in earlier research.

Nowadays not only tick-by-tick data from the exchanges are available, but also economic data from the companies themselves, which are very different in structure and nature. An often used description of Big Data is the "5 Vs rule" which refers to five distinct "dimensions": Volume, Velocity, Value, Veracity and Variety (Zhai et al., 2014).

High-frequency data, thanks to their electronically recorded small number of features (i.e. typically timestamp, quantity, and price), are particularly clean and readable by nature. However their complexity lies in their Volume and Velocity which are both tied to the speed at which new data is generated from the exchanges. For example the order of magnitude of the number of trades per year recorded of the NASDAQ is billions and increasing at a fast pace. As a result, the data volume is usually too large to be stored and analyzed with traditional database technology and the help of distributed systems becomes necessary. Velocity is associated with the rapidity of decision making, the time it takes for a trading system to react to a news or a change in price which tends to decay as the trading speed increases (e.g. typical lead-lag between stocks).

Conversely financial datasets about companies present a smaller, however still large, number of timestamps (financial institutions and companies usually report at most quarterly), but hundreds of features. Quarterly reports are usually electronically uploaded however the backoffice of financial institutions and companies are known not to be fully automatized and are therefore a source of error. In addition, reports vary from one country to the other, and from one industry type to the other which leads to heterogeneity. On the other hand macroscopic data is associated with Variety and Veracity. Variety refers to the different type of data (tables, relational database, product, region, price, name), which still corresponds to structured data. Veracity refers to the trustworthiness of the data. As opposed to market data that are recorded by machines, macrostructure data such as the 13-F are not fully automated filings and are therefore error prone and need further filtering. Each company. depending on the legislation, might have to report different kind of data, which in turn might be interpreted and reformated differently by the regulators. The challenge is to standardize all that information, or at least to be able to filter out non-homogeneous data, from the dataset under study.

Both of kind of dataset have Value, the proof is they are already extensively used separately and usually by different people (with diverse backgrounds and aims), different objectives, and different timelines. However, as those two kind of data are very different yet complementary, combining both should yield more value than when they are treated independently and separately. Because high frequency behaviour may result from the behaviour of economic agents.

Outline

The thesis is divided into four parts.

Part I introduces and provides a technical description of the FactSet ownership dataset. We show some preliminary statistics emerging from the portfolio structure of large financial institutions and funds, and from the capitalization table of recorded securities. The fact that financial actors do not do charity and tend to maximise their return, and consequently minimise their costs, tends to initiate the emergence of some stylized facts (Farmer, 1999) which we will discuss.

Part II proposes a method to asses the statistical significance of the overlap between pairs of heterogeneously diversified portfolios. In particular, we are using public assets ownership data reported by financial institutions in order to infer statistically robust links between the portfolios of financial institutions based on similar patterns of investment. From an economic point of view, it is suspected that the overlapping holding of financial institution is an important channel for financial contagion with the potential to trigger fire sales and thus severe losses at a systemic level. The method is implemented on a historical database of institutional holdings ranging from 1999 to the end of 2013, but can be in general applied to any bipartite network where the presence of similar sets of neighbours is of interest. We find that the proportion of validated network links (i.e., of statistically significant overlaps) increased steadily before the 2007-2008 global financial crisis and reached a maximum when the crisis occurred. We argue that the nature of this measure implies that systemic risk from fire sales liquidation was maximal at that time. After a sharp drop in 2008, systemic risk resumed its growth in 2009, with a notable acceleration in 2013, reaching levels not seen since 2007. We finally show that market trends tend to be amplified in the portfolios identified by the algorithm, such that it is possible to have an informative signal about financial institutions that are about to suffer (enjoy) the most significant losses (gains).

Part III investigates the collective behaviour of fund manager and, in particular, how the average portfolio structure of institutional investors optimally accounts for transactions costs when investment constraints are weak. The collective ability of a crowd to accurately estimate an unknown quantity is known as the “Wisdom of Crowd”. In many situation, the median or average estimate of a group of unrelated individuals is surprisingly close to the true value. In our work we focused on financial market participants, who satisfy two of the conditions understood to underlie the existence of WoC: diversity of opinions, and better than random behaviour. The results extend the so-called “Wisdom of the Crowd” to much more complex situation in two important ways. First, Wisdom of the Crowd also holds for whole functions instead of a point-wise estimates. Second, this shows that in socio-economic systems, the optimal individual choice may only be found when the diversity of individual decisions is

averaged out. Finally we discuss the importance of accounting for constraints when assessing the presence of Wisdom of the Crowd.

In Part IV, we use more than 6.7 billions of trades from the Thomson-Reuters Tick History database and the ownership data from FactSet. We show how the tick-by-tick dynamics of limit order book data depends on the aggregate actions of large funds acting on much larger time scale. In particular, we find that the well-established long memory of marker order signs is markedly weaker when large investment funds trade in a markedly directional way or when their aggregate participation ratio is large. Conversely, we investigate to what respect an asset with a weak memory experiences direction trading from large funds.

Datasets

Firms are required to publicly report their accounting information (e.g. income statement and balance sheet) and, often, institutional investors and funds are also required (depending on the country's regulation) to report key financial data on a regular basis. These publicly available reports are published on the dedicated website of the country's regulator. For example, in the United States, the majority of these reports comes from SEC filings, which are mandatory and must be submitted on a regular (usually quarterly) schedule. Not only they are openly available through SEC's EDGAR database but, as they are provided in an XML based data structure, are easily parsed by a computer.

However the variation in the data one obtains from the files can be significant, and the data contains custom fields that may add complexities when comparing between companies. Therefore the added value of financial data vendors does not lie in the automation of data collection from different sources, which is an easy task for a decent programmer, but is two folds:

1. Raw data comes in very heterogeneous forms, reports are often specific to the entity's type, sector of activity and country. Data vendors are able to provide, thanks to their expertise, data in a standardized (and proprietary) format. In addition, since the data is not always quantitative, qualitative data can be added and needs to be interpreted. Although data vendors try to automate most of these tasks, some of them still need basic analysis by human in order to make them valuable.

2. Data vendors actively participate in the collection of the missing data. For example, since 2004, FactSet reaches out to institutional investors in Europe individually in order to incentivize them to provide their data, even though it is not required for them by the regulators. As a result they are able to record quarterly as much as 60% of the European funds.

We need high quality historical datasets, with data as close to the raw data as possible. In this regard we used two of them:

- FactSet, and in particular its Ownership database, that collects global equity ownership data for a large number of financial institutions and funds with an history back to 1999. The main advantage of FactSet is that is can easily be used from outside as opposed to Bloomberg which incentivize its client to use it directly with their provided terminal.
- Thomson-Reuters Tick History which provides tick-by-tick data for a large spectrum of instruments (e.g. equities, futures, derivatives). They have their systems incorporated latest thechnologies that provide low-latency data that is well suited for microstructure analysis.

These two datasets provide complementary views from two perspectives which are the macro-finance and the high-frequency data from financial markets. The aim of this chapter is to introduce the reader to both databases, and also to present statistics and stylized facts that are relevant to the following developments discussed in this thesis.

2.1 Factset Ownership - General description

Factset is a financial data provider used by investment professionals (e.g. porftolio managers, analysts, investment bankers, fund managers) for their fund screening and selection process. FactSet provides multiple interconnectable datasets (e.g. FactSet Fundamentals, FactSet Estimates, Factset Ownership). We focused on FactSet Ownership in this thesis and will describe its scope and content in this chapter.

The FactSet Ownership database collects global equity ownership data for about 13 000 institutions, 33 000 fund portfolios, and 280 000 non-institutional stake-holders with history going back until 1999. FactSet aggregates publicly available sources (EDGAR forms 13F, N-O, N-CSR and occasionally 485BPOS) for its ownership data. Therefore every data collected is marked with a source name and a report date. Depending upon the nature of the source, the report date may be updated monthly, quarterly, semiannually or annually.

In addition, FactSet actively enriches its data by approaching mutual funds (mainly in Europe) individually to invite them to provide their data.

factset_entity_id	fs_perm_sec_id	security_name
0024G6-E	B00121-S-US	FIRST AMERICA CAPITAL TRUST TR PFD
002202-E	B00242-S-US	NEXT GENERATION TECH HOLDINGS COM
0017JX-E	B00484-S-US	MEDIA PAL HOLDINGS CORP COM
003DZZ-E	B007G7-S-US	VITAL LIVING INC COM
003DZZ-E	B9BH3P-S	VITAL LIVING INC PFCV SER D
04GHFT-E	B00BNB-S-US	TAX EXEMP BD FD AMER INC COM
04GHFT-E	C5BNTC-S-US	TAX EXEMP BD FD AMER INC CL F
04GHFT-E	H8HTCZ-S-US	TAX EXEMP BD FD AMER INC CL C
04GHFT-E	S5LHW7-S-US	TAX EXEMP BD FD AMER INC CL F-2 SHS
000KJ3-E	B00CQC-S-US	FISERV INC COM

Table 2.1: Security level data, extracted from the *own_basic* table, as represented in the Ownership database: it collects securities permanent identifiers together with the general entity identifier of active and terminated securities.

2.2 FactSet identifiers

I provide a short description of three identifiers (entity based, security based and fund based) generated and maintained by FactSet. They are all defined as unique and never changing with respect to their universe and make it possible to merge datas across the FactSet Ownership tables and also across other FactSet maintained databases.

2.2.1 FactSet Entity Identifier

It is the most general identifier defined by FactSet: it is defined as any organization or individual that is included in any of FactSet content sets. Since FactSet provides its dataset in a relational database format, it allows to seamlessly integrate all the tables together, not only in the Ownership database but also across all the datasets. Financial institutions are defined by their entity identifier.

2.2.2 FactSet Permanent Security Identifier

The security identifier is available under the name fs_perm_sec_id. The first 6 characters of the fs_perm_sec_id are a random alphanumeric combination, followed by "-S-" and lastly followed with a 2 character code representing the country in which the security trades. Therefore multiple security identifiers can correspond to a single entity identifier if it is traded in multiple exchanges, as shown in Table 2.1.

2.2.3 FactSet Fund Identifier

As opposed to the financial institutions which are assigned a unique entity identifier, funds have their own identifier: the FactSet Fund Identifier. The distinction between funds and institutions happen because multiple funds can be related to the same institution (e.g. BlackRock is an institution therefore is assigned an institution identifier, however its multiple funds are assigned a unique fund identifier each).

2.3 Notations

Let us define some necessary quantities to be more precise, the notations defined in this section will be sensibly valid through the whole thesis, unless said otherwise. At quarter q, fund (resp. institution) i has capital $W_i^{(f)}(q)$ (resp. $W_i^{(i)}(q)$) which is invested into $n_i^{(f)}(q)$ (resp. $n_i^{(i)}(q)$) securities among $M(q)$ existing ones. As a result, each security α, whose capitalization is denoted by $C_\alpha(q)$, is found in $m_\alpha^{(f)}(q)$ funds' (resp. $m_\alpha^{(i)}(q)$ institutions') portfolios. $W_{i\alpha}^{(f)}(q)$ (resp. $W_{i\alpha}^{(i)}(q)$) is the position in dollars of fund (resp. institution) i on security α at the end of quarter q.

In all the forthcoming chapters of this thesis we will consider only funds or institutions separately (chapter 3 will focus on institutions whereas chapter 4 and 5 on funds), therefore, for the sake of clarity, we will drop the exponents f and i when there is no ambiguity. Time referring to quarterly updated data will be denoted by q and by t otherwise. However, the explicit time dependence will usually be dropped when not necessary.

2.4 FactSet - Securities

A security is a good or an asset that holds a monetary value. Its legal definition varies by jurisdiction: In the United States, a security is a tradable asset of any kind. Securities are usually categorized into:

- equity securities (common stocks)
- debt securities (banknotes, bonds, and debentures)
- derivatives (forwards, futures, options and swaps)

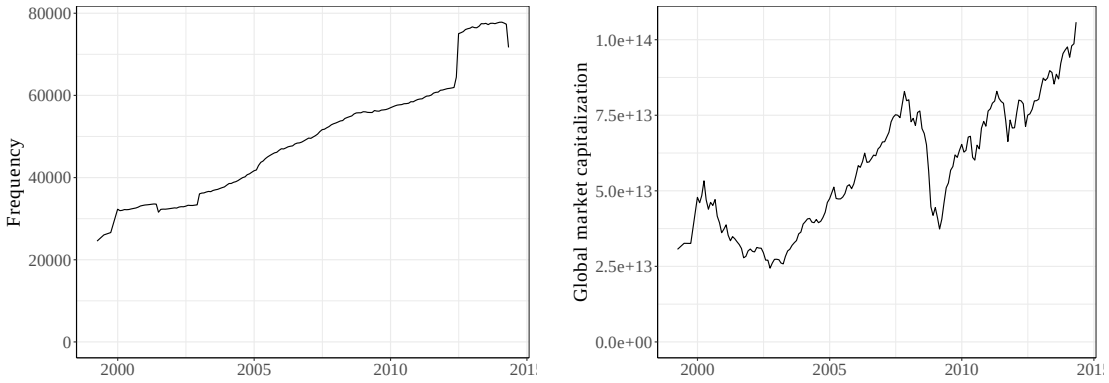

Figure 2.1: The left plot represents the total number of securities collected in FactSet Ownership as a function of time, and the right plot is their corresponding market capitalization as a function of time.

In the scope of this thesis, a security will usually refer to an equity security, that is, an ownership interest held by a shareholder in an entity (company, partnership or trust). Securities included in the database are actively traded and are one of the issue type defined in Table 2.2. As we explained in the introduction, FactSet is constantly increasing its range of coverage. As a result Figure 2.1 shows a sensible increase in the total number of securities and total market capitalization over the studied period of time. One should keep that property of the dataset in mind when studying... as it can have great consequences on, for example, the computation of average degree of a bipartite network.

Since securities are selected by institutional investors, they only play a passive role in the Ownership database. Therefore they only have two dedicated tables which are the "own_basic" defined earlier and the "own_prices" (see Table 2.2) which is a record of the price history and share outstanding of the securities. This table is useful to convert the number of shares into dollar, and therefore to compute the portfolio value of an investor, as well as the market capitalization of a security.

The vast majority, more than ninety percent, of the securities are located in North America, Europe and Asia.

2.4.1 Distributions

We plot the distribution of company size in term of market capitalization (fig. 2.6), number of institutional investors (fig. 2.9), number of fund investors (fig. 2.15) and

Security Issue Types		
Type	Description	Count
EQ	Equity	70452
OE	Open-End Mutual Fund	30277
PF	Preferred	5930
AD	ADR/GDR	4104
ET	Exchange Traded Fund	4017
CP	Convertible Preferred	3622
CE	Closed-End Mutual Fund	2740
WT	Warrant/Right	2325
DR	Derivative	1115

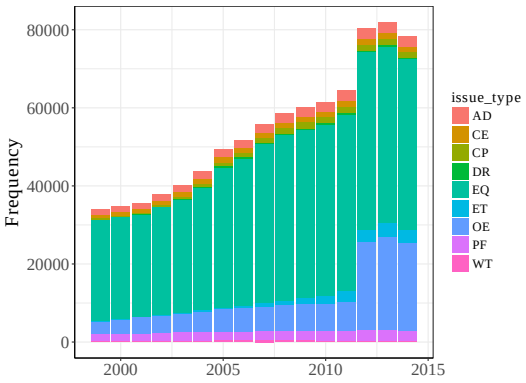

Figure 2.2: The table presents the different security issue type identifiers with their corresponding description and frequency over the whole history. On the right we plot the detailed year-by-year issue type partition as a function of time.

fs_perm_sec_id	price	shares_outstanding	price_date
FV74CR-S-US	29.60001	26674500	2007-03-31
F96L6N-S-GB	1.37000	7913990	2001-02-28
QW2KHK-S-IE	19.06106	0	2013-06-30
K5DYBW-S-US	0.00000	0	2009-02-28
SFCB57-S-US	6.00000	193838	2011-05-31
MBK7R0-S-MY	0.24798	469668997	2007-07-31
RFVH3G-S-US	13.87500	9782000	2000-05-31
XKVSDH-S-AR	0.00000	811185367	2001-03-31
K38ZJZ-S-US	46.91906	1555031	2012-12-31
SK4BC2-S-FR	0.00000	0	2008-08-31

Table 2.2: Excerpt from the "own_prices" table. This table records monthly security prices and shares outstanding of the active and terminated securities followed by FactSet. The data is adjusted as of the adjustment which is found in the own_basic table.

Security region		
Code	Name	Frequency
N	North America	52407
E	Europe	40884
A	Asia	24574
Y	Pacific	2681
M	Middle East	1827
L	Latin America	1820
F	Africa	1299

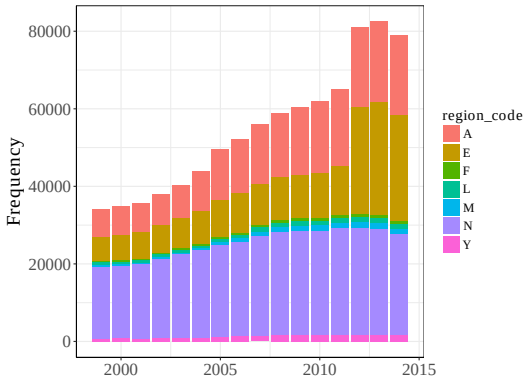

Figure 2.3: Repartition of the securities over the seven region defined by FactSet for the whole history (left table) and on a year-by-year basis (right bar plot).

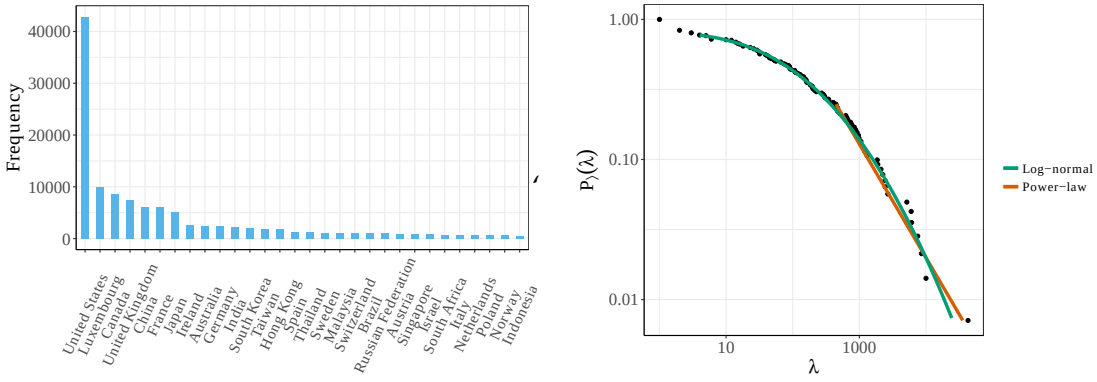

Figure 2.4: Number of securities per country for the 20 most represented countries (left plot). Complementary cumulative distribution function of country frequencies (right plot).

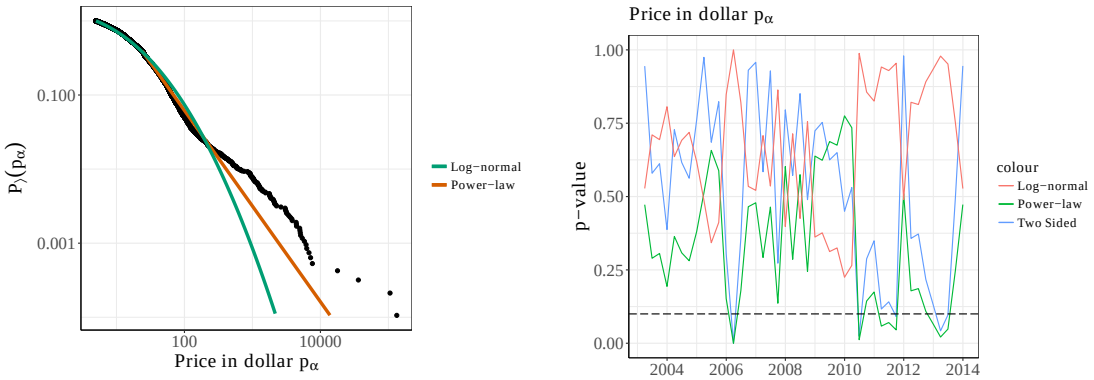

Figure 2.5: Complementary cumulative distribution function of the number of the price p_α of the securities in dollar as of 2012-09-30 (left plot). Two-sided p-values (blue line) and one-sided p-values for log-normal (red line) and power-law (green line) distributions as a function of time (right plot).

stock prices (fig. 2.5). These distributions will be studied in greater details in section 2.9).

2.5 FactSet - Institutions

The primary source for institutional ownership of U.S.-traded equity is the 13F filing. This filing is mandated by the SEC for any investment management institution managing $100 million or more in U.S.-traded securities. Approximately 3,200 institutions file 13Fs on a quarterly basis (filed within 45 days of the calendar quarter-end, and reporting holdings as of that quarter-end). These filings are available electronically on the SEC's EDGAR system. Form 13F is limited to the EDGAR system. As the 13F requires to report only positions greater than 10000 shares or $200000 on listed securities, some positions are already filtered out. Also it represents only the aggregated positions of the institutions: the portfolios of sub-funds, that we will describe in the next section, are merged into a single report.

Using the 13F filing is convenient because all the financial institutions are required to report quarterly and at the same dates. Therefore the discrete time series are observed in a synchronous manner and, as we have seen, small positions are not reported so already filtered out.

Table 2.3 shows the structure of the table that contains all the history of the positions held by the institutions. Here, the factset_entity_id represents an institution. For example, the institution 002H53-E held 8100 shares of the security H0PJ8L-S-US as of

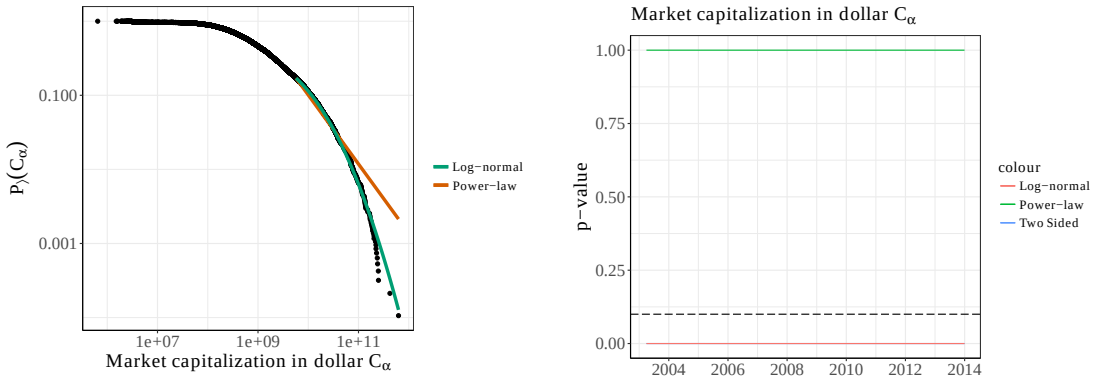

Figure 2.6: Complementary cumulative distribution function of market capitalization C_α as of 2012-09-30 (left plot). Two-sided p-values (blue line) and one-sided p-values for log-normal (red line) and power-law (green line) distributions as a function of time. (right plot)

fs_entity_id	fs_perm_sec_id	holding	report_date
002H53-E	H0PJ8L-S-US	8100	2011-12-31
002J41-E	KB76XV-S-US	234000	2010-09-30
005VN3-E	RPSG79-S-US	50000	2011-03-31
07QM1L-E	T2Z4DM-S-US	9803	2013-09-30
002BZT-E	T8HTVR-S-US	50	2011-09-30
07JGRK-E	P5BL31-S-US	63628	2000-06-30
0BJ0LD-E	B4L9DK-S-US	53604	2009-09-30
05F9JR-E	Q8FXPB-S-US	274955	2006-12-31
009YS4-E	K14DFC-S-US	6400	2008-06-30
000RGG-E	RLCVV6-S-US	8349	2013-09-30

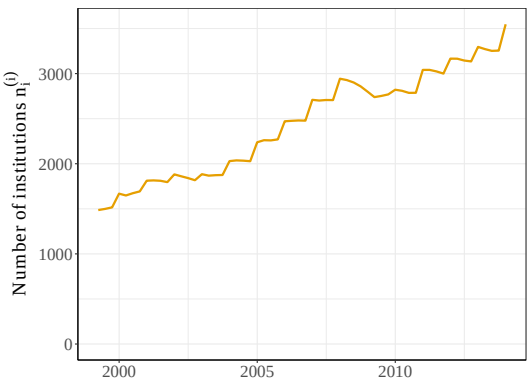

Table 2.3: Excerpt from the "own_13f_holdings_hist" table. This table contains unadjusted historical 13F reported ownership data for active and terminated 13F filers. Active and terminated securities are included.

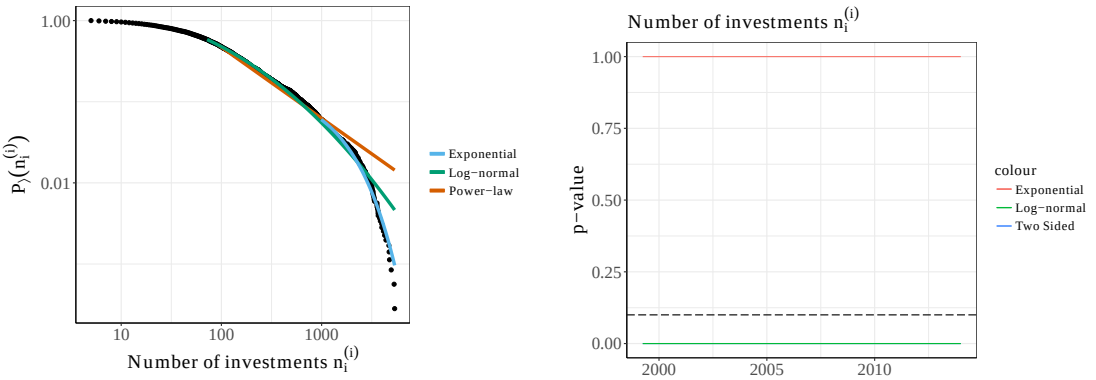

Figure 2.7: Complementary cumulative distribution function, of the number of investments $n_i^{(i)}$ of institutions as of 2012-09-30 (left plot). Two-sided p-values (blue line) and one-sided p-values for exponential (red line) and log-normal (green line) distributions as a function of time. (right plot)

2011-12-31. From that table it is possible to reconstruct the portfolio of any institution in the database at any time.

2.5.1 Distributions

We plot the distribution of institution size in term of total portfolio value in dollar (fig. 2.8) and number of investments (fig. 2.7) these distributions will be studied in greater details in section 2.9).

2.6 FactSet - Funds

Detailed information about funds holdings is also provided, however FactSet uses a wide variety of public filings and other sources to compile fund ownership data. The raw data is formatted similarly to that of the financial institutions in the previous section. Individual funds are uniquely identified by their factset_fund_id, see Table 2.4.

Therefore, as opposed to 13F Filings, the report frequency ranges from as often as once a month to as little as once a year (see fig. 2.10). Moreover the report date is not fixed and can happen at almost any day of the month (albeit often last week of the month). As a consequence the main challenge we had to tackle while pre-processing the data was to deal with the disparity and the non-synchronous time-series. In the following we will cover the cleaning and synchronizing process.

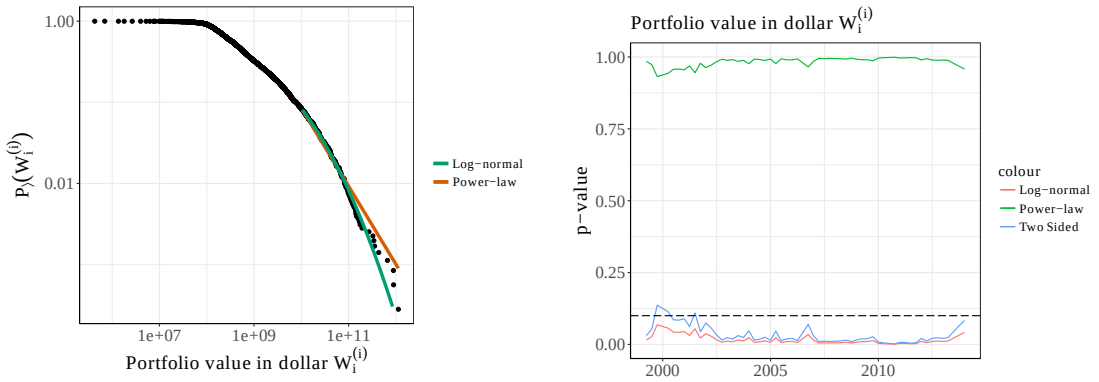

Figure 2.8: Complementary cumulative distribution function of the portfolio value $W_i^{(i)}$ of institutions as of 2012-09-30 (left plot). Two-sided p-values (blue line) and one-sided p-values for log-normal (red line) and power-law (green line) distributions as a function of time. (right plot)

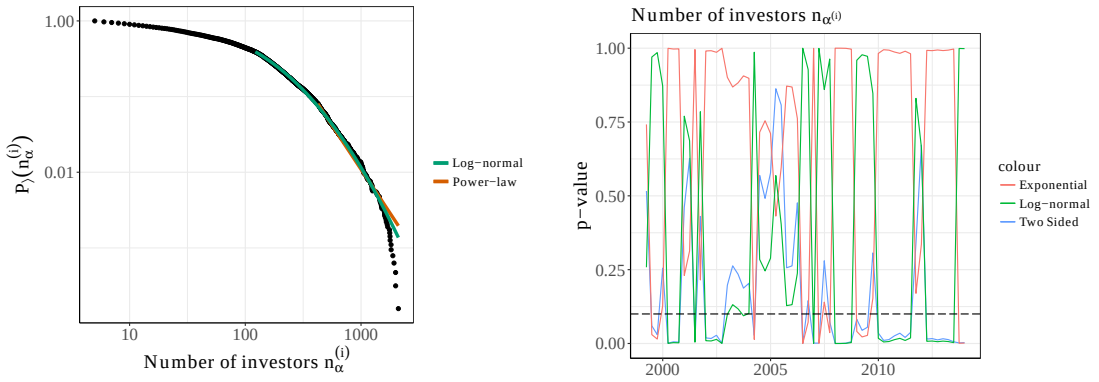

Figure 2.9: Complementary cumulative distributions function of the number of investors $n_\alpha^{(i)}$ as of 2012-09-30 (left plot). Two-sided p-values (blue line) and one-sided p-values for exponential (red line) and log-normal (green line) distributions as a function of time. (right plot)

fs_fund_id	*fs_perm_sec_id*	*holding*	*report_date*
04BK15-E	*XLSRNP-S-US*	*1100*	*2001-12-31*
04BPX1-E	*MPXT90-S-JP*	*352800*	*2010-09-30*
04BNZ0-E	*NYJLQ1-S-CH*	*2635324*	*2010-03-31*
04CS9Q-E	*H8KLSN-S-IN*	*117800*	*2012-03-31*
04CQC8-E	*JCQKGG-S-GB*	*11344*	*2010-05-31*
04G7C6-E	*D0LWQH-S-DE*	*760*	*2012-08-31*
04GRB7-E	*H7K4T4-S-US*	*75*	*2012-11-30*
04C0FY-E	*B4Q6W7-S-DE*	*50000*	*2008-11-30*
04CSLR-E	*GB58PD-S-US*	*1475*	*2008-03-31*
04BCZX-E	*F6XS56-S-US*	*19500*	*2005-09-30*

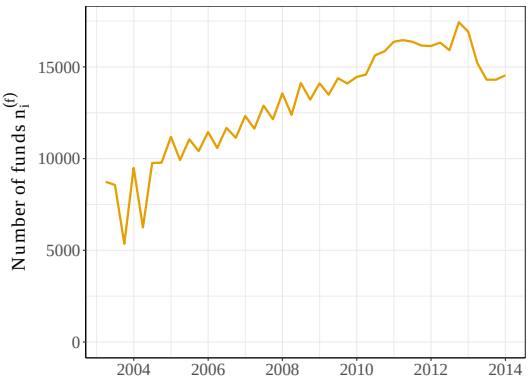

Table 2.4: Excerpt from the table that collects historic holdings of recorded funds.

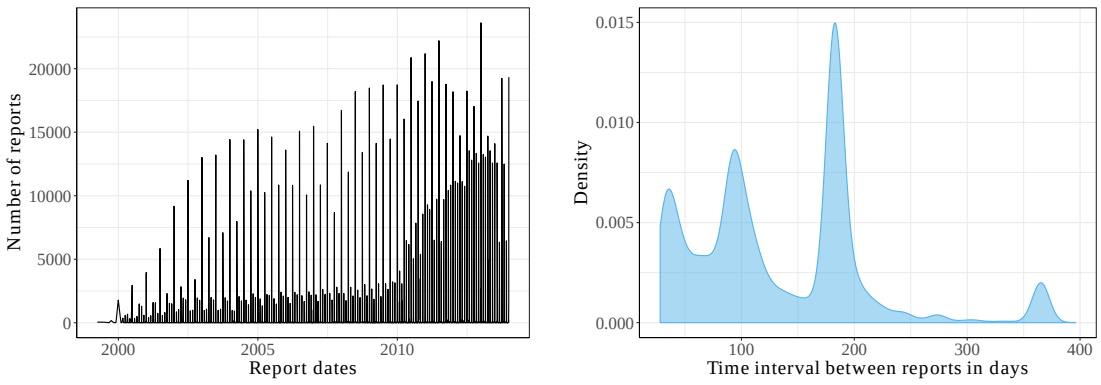

Figure 2.10: Number of reports filled by fund managers for every dates in the studied history (left plot). Distribution of the time interval, in days, between two reports over the whole history (right plot)

2.6.1 Non-synchronous time-series

The two main factors that cause the time-series to be non-synchronous are the reporting frequency and the date of the reports. The combination of these two factors causes large and small spikes in the frequency of dates of report (see left plot in fig. 2.10). The reporting frequencies can be appreciated by looking at the time-interval distribution between two reports and reveals that funds usually report monthly, quarterly, bi-annually or annually (see density plot in fig. 2.10), with the majority of the reports happening bi-annually. The disparity in the date of the reports will be observed more easily afterwards. Hereafter we describe the three-steps procedure we followed in order to synchronize the time-series:

1. We have seen that the time resolution ranges from one or two data points a year (annual or bi-annual reports) to twelve (a minority of funds, mostly US-based, reports monthly). Quarterly updated funds appear to be a good compromise between statistics and time-resolution. Therefore we filter out funds that are not updated at least quarterly on average (see left plot of Figure 2.11). The report dates frequencies appear to be cleaner, however there is still an alternate appearance of large and small spikes. The small spikes are now coming from two sources: monthly updated funds, and report dates disparity.
2. We generate quarterly updated time-series of dates occuring on the last day of the month (similarly to 13F filings), that we will refer to as *merge dates*. We replace the old *report dates* with the nearest *merge date* happening in the future (it can be achieved using a rolling join). Then, if the combination of *merge dates* and *report dates* is not unique (e.g. a monthly updated fund will have three *report dates* assigned to a single *merge date*) only the last one is kept (i.e. the closest to the assigned *merge date*). The resulting time series are at most quarterly updated and fully synchronized.
3. The final step of the procedure is to require the funds to appear a minimum of four times in a row (a full year).

The final result is synchronous quarterly updated time-series, as shown in right plot of Figure 2.11.

2.6.2 Cleaning and filtering

As we have already seen so far the FactSet Ownership dataset presents only few blatant errors that are easily cleaned up. Additionally filtering is required for some quantities. We describe the full process hereafter.

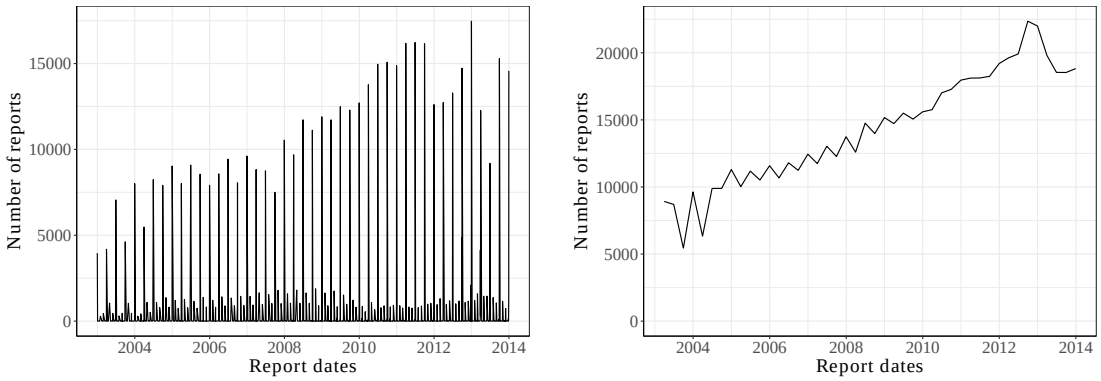

Figure 2.11: Number of reports by date. Left plot: For funds that report at least quarterly. Right plot: After applying the full filtering procedure.

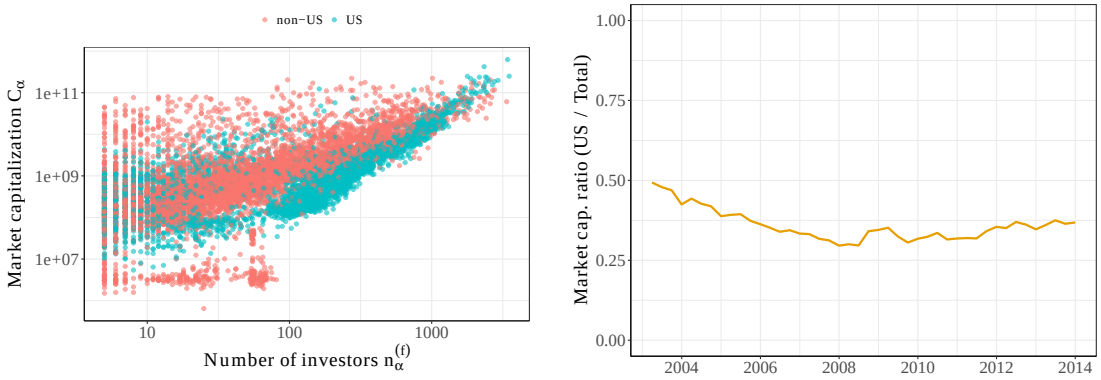

Figure 2.12: Left plot: Market capitalization as a function of the number of investors for US securities (orange points) and non-US securities (green points). Right plot: Temporal evolution of the aggregated market capitalization of US over the total market capitalization.

Period. Figure 2.11 shows that year 2014, as the number of securities and funds decreased, is suspicious. There is a lag between the date of a report and when FactSet adds it to its database and this lag probably varies depending on the source, the country, and the sector. Since our subscription to FactSet ended in March 2015, the year 2014 is not fully furnished. Therefore, in this thesis, we restrict our study to at most 36 quarterly snapshots starting from the first quarter of 2005 and ending with the last quarter of 2013.

Country of origin. Plotting the market capitalization of the securities as a function of their number of investors (see left plot in fig. 2.12) results in two distinct cloud of dots, each of them corresponds to a different region of origin: green (resp. orange) cloud corresponds to non-US (resp. US)-based securities. The origin of this large difference

between these two regions is not clear: it could for example come from differences in regulations in non-US countries, and that FactSet is more effective to collect data in the US than anywhere else. The fact that the market capitalization is a single number, so easily collected by FactSet, and that the number of investors requires to be collected one by one seems to point out to the latter. It turns out that the ratio of the investment values in US and non-US assets varies little as a function of time (see Fig. 2.12). As a consequence we focused on US securities. For that reason we decided to focus only on US securities and removed the non-US securities from the dataset.

Penny stocks. Refers to securities which trade below \$5 per share in the USA. Since they are considered highly speculative investments and are subject to different regulations are not listed on a national exchange. We filtered them out (price $p_\alpha \geq 5$ USD).

Size. We set a threshold on the size of securities (number of investors $n_\alpha \geq 10$ and market capitalization $C_\alpha \geq 10^5$ USD) and on the size of the institutions and funds (number of investments $n_i \geq 5$ and portfolio value $W_i > 10^5$USD).

Buys and sells. A convenient quantity to compute (that we will use in...) is the difference in ownership between two quarters $W_{i\alpha}^{q} - W_{i\alpha}^{q-1}$(buy or sell). However, FactSet does not always explicitly specify a sold out position with a zero shares held and usually represents it by an absence of the position. In order to compute ... we add the missing zeros to correctly identify sold out positions.

At this stage the filtering on the funds is finished and we can compute the final portfolio value W_i and the number of investments n_i that include all the reported positions (on U.S.-based securities) of the selected funds.

2.6.3 Distributions

We plot the distribution of institution size in term of total portfolio value in dollar (fig. 2.14) and number of investments (fig. 2.13) these distributions will be studied in greater details in section 2.9).

2.7 Thomson-Reuters Tick History

Thomson-Reuters Tick History (TRTH) provides access to historical high-frequency data, with tick-by-tick market resolution, across global asset classes (over 5 million instruments from various exchanges) dating to 1996. It reports anonymous transactions and updates of the aggregated quotes in *real-time*. TRTH provides for every instrument at least two, and up to three, distinct files:

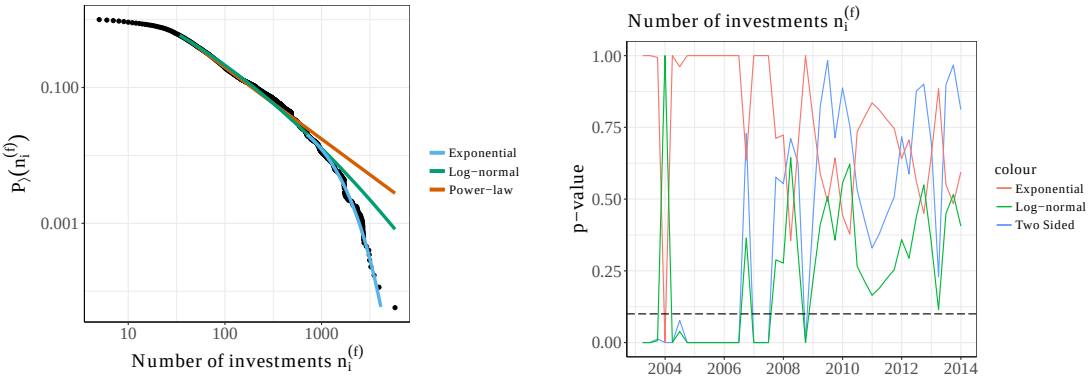

Figure 2.13: Complementary cumulative distributions functions of the number of investments n_i^f of funds as of 2012-09-30 (left plot). Two-sided p-values (blue line) and one-sided p-values for exponential (red line) and log-normal (green line) distributions as a function of time. (right plot)

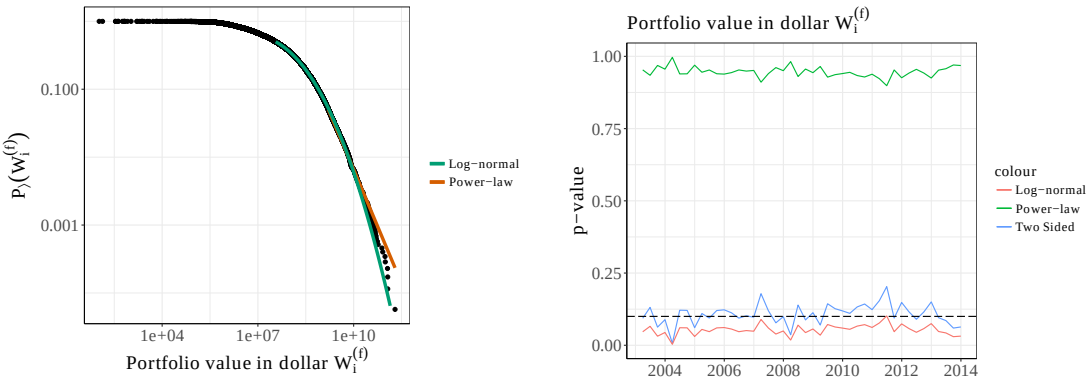

Figure 2.14: Complementary cumulative distributions functions of the portfolio value W_i^f of funds as of 2012-09-30 (left plot). Two-sided p-values (blue line) and one-sided p-values for log-normal (red line) and power-law (green line) distributions as a function of time. (right plot)

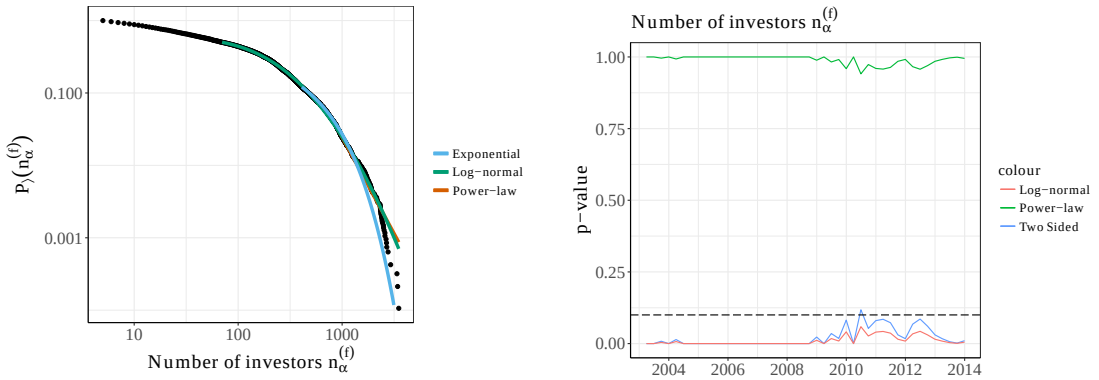

Figure 2.15: Complementary cumulative distributions functions of the number of investors n_α^f of funds as of 2012-09-30 (left plot). Two-sided p-values (blue line) and one-sided p-values for log-normal (red line) and power-law (green line) distributions as a function of time. (right plot)

- *Trades messages*, it contains all the trades timestamped.
- *Quotes messages*, it contains the best limit.
- *Market depth*, which is not always available, especially in the US. Contains the order book up to 10 limits.

The scope of this thesis includes the study of the long-memory length of order signs however TRTH does not provided signed trades, in addition both files are not exactly synchronized and there is a non-constant lag of about a few milliseconds in between. A standard Lee-Ready (Lee and Ready, 1991) procedure could be applied or a more refined one (Toke, 2016) (or Easley et al. (2016)). However the accuracy of the Lee-Ready trade classification algorithm is controverted (Theissen, 2001), and the use of a more refined algorithm (Toke, 2016) was not feasible given the substantial amount of trades we had to process (more than 6 billions). Therefore we used the simpler tick-test method.

2.8 Thomson-Reuters Tick History and FactSet matching

This section is relevant for Chapter 5 that focuses on the relationship that exists between large-trader activity and the long memory of limit-order markets. We limited our study to 32 quarterly snapshots that correspond to the 2007-2013 period, which starts before the 2008 crisis and ends with the validity of our FactSet dataset. We focused our work on US-based securities (as explained in section 2.6.2) and most of

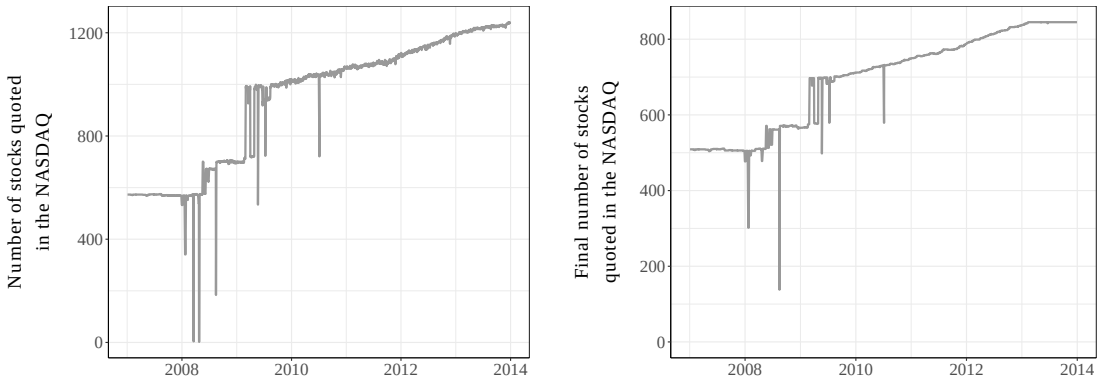

Figure 2.16: Number of stocks matched between FactSet and TRTH before (left plot) and after (right plot) filtering as a function of time.

them were traded on the NYSE or the NASDAQ. For simplicity we focused the analysis of intraday data from only one US exchange and, because it is fully automatized, we arbitrarily chose the NASDAQ which left us with 2480 securities identified as being traded on the NASDAQ by FactSet.

The next step was to match securities from FactSet with their corresponding Reuters Instrument Code (RIC) in TRTH (a ticker-like code to identify financial instrument and indices). A simple method would have been to use the ISIN, however our TRTH dataset does not contain a RIC/ISIN correspondence table. Therefore a less trivial matching method was to use the company names of the securities in FactSet and searched for their corresponding RIC on the Reuters website [1]. Out of the 2480 securities traded on the NASDAQ from FactSet, using automated methods that involve string comparison algorithm based on the Levenshtein distance, we link assets found in both FactSet and the TRTH databases. For each asset traded on the NASDAQ and each day, we extracted all the trade prices together with the best bid and ask prices just before the trades. We extract the NASDAQ's tick-by-tick data provided by the Thomson-Reuters Tick History (TRTH) database. For each stock, it reports on two different files the anonymous and unsigned trades and updates on the order book. We remove 19/03/2008, 24/04/2008 and 25/04/2008 that have significantly less traded stocks than other dates (as shown in fig. 2.16). Finally, we keep assets traded for at least 200 days and with more than 200 trades per day on average. That leaves us with 846 stocks and more than 6.7 billion trades (see global daily volume and number of trades in fig. 2.17).

[1] https://www.reuters.com/finance/stocks/lookup

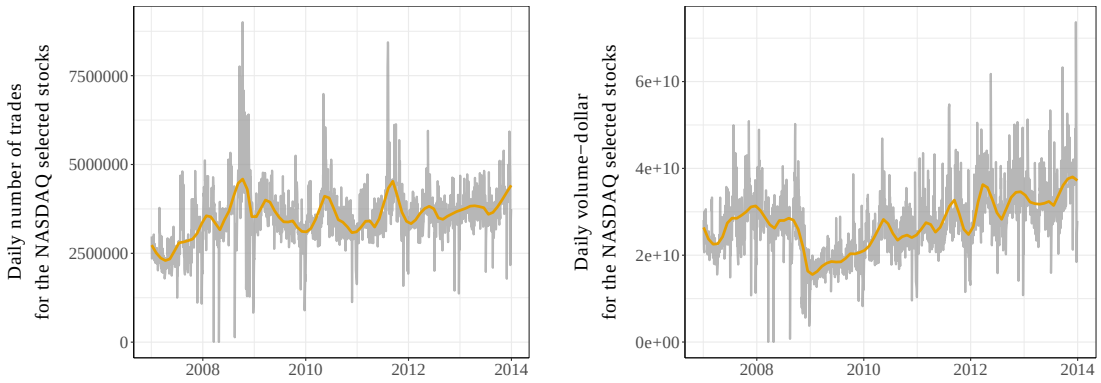

Figure 2.17: Daily number of trades (left plot) and total daily volume-dollar (right plot) for the NASDAQ stocks under study as a function of time.

2.9 Heavy tail distributions

The investigation of distributions is a large area of research in economy (Pareto, 1964; Stanley et al., 1996; Amaral et al., 1997; Axtell, 2001; Gabaix et al., 2003a; Mizuno et al., 2004) and finance (Mandelbrot, 1967, 1997; Farmer, 1999; Bouchaud, 2001; Solomon and Richmond, 2001)), they have in particular been observed in the distribution of wealth (Solomon and Richmond, 2001), volatility, mutual fund size, or firm size. A system that exhibits a heavy tail can often be linked to key properties of an underlying stochastic process that are responsible for creating the observed distribution. The aim of this section is to provide a short review of diverse distribution that can be observe with the FactSet dataset, and to use a systematic and rigorous approach in order to compare candidate distributions. However we shall not relate our observations with generic mechanisms.

2.9.1 Method

We use the `poweRlaw` R package (Gillespie, 2015), which implements the method of Clauset et al. (2009), which enables power-laws and other heavy tailed distributions to be fitted in a straightforward manner. It is argued that only the tail of the distribution follow a power-law in practice, therefore a first step is to estimate the scaling region parameter $x_{min}$of the power-law model. This threshold is estimated using a Kolmogorov-Smirnov approach as recommended by (Clauset et al., 2009).

The second part of the analysis is to asses which distribution is closer to the true distribution. We will compare the candidate distribution (e.g. power-law) with another plausible distribution (e.g. log-normal) using we will first compare the power-law

distribution with the log normal distribution using Vuong's test statistic (Vuong, 1989), which is a likelihood test for model selecting using Kullback-Leibler criteria.

This is a two-step process where the first hypothesis being tested is:

- H_0: Both distribution are equally far from the true distribution.
- H_1 : One of the test distributions is closer to the true distribution.

If H_0 is rejected during that first step, i.e. two-sided p-value smaller than 0.1 (named "Two Sided" in figures' legend), it means that it is worth choosing between one or the other distribution. The one-sided p-value then gives an upper limit on obtaining that small log-likelihood ratio if the selected distribution is true.

2.9.2 Results

The distribution of fund sizes was reported to have a power-law tail (Gabaix et al., 2003b,a, 2006), and more recently a log-normal tail (Schwarzkopf and Farmer, 2008, 2010). Our observations for fund sizes (see fig. 2.14) is in agreement with the later, however not statistically significant for every dates, as shown by the two-sided p-value. Interestingly, institution sizes, which often represent an aggregated size of multiple funds, displays a more statistically significant log-normal tail when compared with a power-law (see fig. 2.8).

Measuring fund (see fig. 2.13) and institution (see fig. 2.7) sizes distribution in term of number of investments yields a different observation: a thinner tail. We did not investigate this result any further, however a constraint such as an upper-bound in the number of available investments could be at play: the largest institutions are an order of magnitude bigger than the largest funds, however their largest number of investments is almost the same.

The functional form of the upper tail of the firm size distribution is debated. The size distribution of firms aggregated across industries is accepted to display a power-law tail (Axtell (2001); Bottazzi and Secchi (2003); Dosi (2005)). Whereas, for firms belonging to a single industry, the size distribution of firms is sometimes found to be a log-normal (Stanley et al. (1995, 1996); Bottazzi and Secchi (2003); Dosi (2005)) or a power-law (Axtell (2001)). However we find that security sizes distribution (see fig. 2.6) displays a log-normal tail.

We also measured the size distribution of firms in term of number of investors that we could divide in two categories: institutions and funds. When measured in term of

funds, the size distribution seems to have a log-normal tail (see fig. 2.15) whereas it is unclear when measured in term of institutions (see fig. 2.9).

The distribution of prices, which can be manipulated with stock split and reverse stock split, displays a heavy tail which tends to be closer to a power-law in the latest years. (see fig. 2.5).

Chapter 3

Statistically validated network of portfolio overlaps and systemic risk

The 2007-2008 global financial crisis has drawn the attention of both academics and regulators to the complex interconnections between financial institutions (Glasserman and Young, 2015) and called for a better understanding of financial markets especially from the viewpoint of systemic risk, i.e., the possibility that a *local* event triggers a *global* instability through a cascading effect (Brunnermeier, 2009; Chan-Lau et al., 2009; Staum, 2012; Acemoglu et al., 2015; Battiston et al., 2016; Gai and Kapadia, 2010). In this respect, while much effort has been devoted to the study of counterparty and roll-over risks caused by loans between institutions (Allen and Gale, 2000; Eisenberg and Noe, 2001; Iori et al., 2006; May and Arinaminpathy, 2010; Haldane and May, 2011; Bluhm and Krahnen, 2011; Krause and Giansante, 2012; Cimini et al., 2015; Cimini and Serri, 2016; Barucca et al., 2016), the ownership structure of financial assets has received relatively less attention, primarily because of lack of data and of adequate analysis techniques. Yet, while in traditional asset pricing theory assets ownership does not play any role, there is increasing evidence that it is a potential source of non-fundamental risk and, as such, can be used for instance to forecast stock price fluctuations unrelated to fundamentals (Greenwood and Thesmar, 2011; Anton and Polk, 2014). More worryingly, if investment portfolios of financial institutions are too similar (as measured by the fraction of common asset holdings, or portfolio overlap), the unexpected occurrence of financial distress at the local level may trigger fire sales, namely assets sales at heavily discounted prices. Fire sales spillovers are believed to be an important channel of financial contagion contributing to systemic risk (Shleifer and Vishny, 1992; Cifuentes et al., 2005; Shleifer and Vishny, 2010; Caccioli et al., 2014; Cont and Wagalath, 2014; Greenwood and Thesmar, 2011): when assets prices are falling, losses by financial institutions with overlapping holdings become self-reinforcing and trigger further simultaneous sell orders, ultimately leading to downward spirals for asset prices. From this point of view, even if optimal portfo-

lio selection helps individual firms to diversify risk, it can also make the system as a whole more vulnerable (Glasserman and Young, 2015; Corsi et al., 2016). The point is that fire sale risk builds up gradually but reveals itself rapidly, generating a potentially disruptive market behavior.

In this contribution we propose a new statistical method to quantitatively assess the significance of the overlap between a pair of portfolios, with the aim of identifying those overlaps bearing the highest riskiness for fire sales liquidation. Since we apply the method to institutional portfolios we will use interchangeably the terms institution and portfolio throughout the paper. In practical terms, the problem consists in using assets ownership data by financial institutions to establish links between portfolios having strikingly similar pattern of holdings. Market ownership data at a given time t consists of a set $I(t)$ of institutions, holding positions from a universe of $M(t)$ securities (or financial assets in general). The $|I(t)| \times |M(t)|$ ownership matrix $\mathcal{W}(t)$ describes portfolios composition: its generic element $W_{i\alpha}(t)$ denotes the number of shares of security $\alpha \in M(t)$ held by institution $i \in I(t)$. The matrix $\mathcal{W}(t)$ can be mapped into a binary ownership matrix $\mathcal{A}(t)$, whose generic element $A_{i\alpha}(t) = 1$ if $W_{i\alpha}(t) > 0$ and 0 otherwise, which allows to define the degree $d_i(t) = \sum_\alpha A_{i\alpha}(t)$ of an institution i as the number of securities it owns at time t, and the degree $d_s(t) = \sum_i A_{i\alpha}(t)$ of a security s is the number of investors holding it at time t. The number of securities held by both institutions i and j, namely the overlap of their portfolios, is instead given by $o_{ij}(t) = \sum_\alpha A_{i\alpha}(t)A_{j\alpha}(t)$ (with $i \neq j$), which is the generic element of the $|I(t)| \times |I(t)|$ portfolio overlap matrix $\mathcal{O}(t)$. In network theory language, $\mathcal{O}(t)$ represents a projected monopartite network of institutions obtained as a contraction of the binary ownership matrix $\mathcal{A}(t)$, which instead represents a bipartite network of institutions and securities. However, in such a projected network two institutions are connected as soon as they invest in the same security: this generates too many links and fails to filter out less risky overlaps. For example, a security held by a large number of investors would trivially determine a correspondent number of projected links without a clear meaning. Although there is no unique way to tackle this problem, the point of view we take here can be roughly summarized as follows: if we were to reshuffle links in the original bipartite network without changing the degree of each node, how likely is the observed overlap? Thus, the problem is that of building a *validated* projection of the original bipartite network containing only the most significant overlaps that cannot be explained by a proper null network model. In this way we can drastically reduce the original amount of links and obtain a much sparser validated network with a clearer meaning.

All methods to build validated projections proposed in the literature involve the use of a threshold to determine which links are retained in the monopartite network, but vary in how the threshold is chosen (Neal, 2014). The simplest and most common approach is to use an unconditional global threshold (Latapy et al., 2008; Neal, 2013), which

however suffers from arbitrariness, structural bias and uniscalarity—by systematically giving preference to institutions with many holdings (Neal, 2014). Using a threshold which depends on institution degrees can overcome the latest two limitations (Serrano et al., 2009; Borgatti and Halgin, 2011). In particular, the threshold can be determined using a null hypothesis of random institutions-to-securities matching constrained to institutions degree, for which the probability that two institutions share a given number of securities is given by a hypergeometric distribution (Goldberg and Roth, 2003; Sudarsanam et al., 2002). Yet, also this approach is biased by implicitly treating securities as equivalent and interchangeable. A recently proposed improvement to this method consists in building homogeneous networks of securities, that is, in splitting the original bipartite network into subnetworks each consisting of securities with the same degree and of all institutions linked to them (Tumminello et al., 2011). In this way, the null hypothesis can be properly cast, for each layer separately, with the hypergeometric distribution. Problems however arise when securities are characterized by a strongly heterogeneous number of investors: the process of creating homogeneous subnetworks with securities having the same degree often translates into almost empty subsets, causing a serious resolution problem and leading to almost empty validated networks (see section Methods). A possible solution here is to perform link validation without taking into account degree heterogeneity (Tumminello et al., 2011), which however cannot be formalized analytically since the events of choosing different securities have now different occurrence probabilities. An alternative approach consists in using a null model of random institutions-to-securities matching constrained not only to institutions degree but also to securities degree. The fixed degree sequence model (FDSM) (Zweig and Kaufmann, 2011; Horvát and Zweig, 2013) and the stochastic degree sequence model (SDSM) (Neal, 2014) belong to this category. In the FDSM, the null hypothesis cannot be formalized analytically and the method relies on a conditional uniform graph test by generating a microcanonical ensemble of random graphs whose overlaps can be compared with the empirical ones. However, algorithms to sample the graph configuration space are impractically complex or biased (Blanchet and Stauffer, 2013), or suffer from arbitrariness (Gionis et al., 2007). In contrast, in the SDSM the null hypothesis can be formalized at least numerically, but it is computationally impractical in most cases (Neal, 2014). Thus, also the SDSM relies on a conditional uniform graph test, which is however easy to achieve by using the linking probabilities between institutions and securities obtained with the Link Probability Model (LPM) (McCulloh et al., 2010). Yet, in the LPM these probabilities are basically the proportion of link occurrences over multiple observations of the data, which requires much more information than that contained in the ownership matrix, and, more importantly, represents a valid approach only when the underlying network is assumed to be stationary in time—which is clearly not the case for stock markets.

The method we propose here overcomes all the limitations of its predecessors by building on a null hypothesis described by the *Bipartite Configuration Model* (BiCM) (Saracco

et al., 2015), which is the extension of the standard *Configuration Model* (Park and Newman, 2004) to bipartite graphs. In the null BiCM network, institutions randomly connect to securities, but the degrees of both institutions and securities are constrained on average to their observed values in real ownership data. This is achieved through maximization of the Shannon entropy of the network subject to these constraints, which remarkably allows to analytically and numerically formalize the null hypothesis (see section Methods). The additional advantages of the BiCM with respect to Tumminello et al. (2011) is that of not requiring the homogeneity of neither layer of the network, and with respect to Neal (2014); McCulloh et al. (2010) of using only the information contained in a single snapshot of the data. The method works as follows. For each date t, in order to distinguish the true signal of overlapping portfolios from the underlying random noise, every link of the projected network has to be independently validated against the BiCM null hypothesis. Thus, for each pair of institutions (i, j) having overlap $o_{ij}(t)$, we compute the probability distribution $\pi(\cdot|i, j, t)$ of the expected overlap under the BiCM (see section Methods). The statistical significance of $o_{ij}(t)$ is then quantified through a p-value:

$$P[o_{ij}(t)] = 1 - \sum_{x=0}^{o_{ij}(t)-1} \pi(x|i, j, t), \tag{3.1}$$

where the right-hand size of Eq. (3.1) is the cumulative distribution function of $\pi(\cdot|i, j, t)$, namely the probability to have an overlap larger or equal than the observed one under the null hypothesis. If such a p-value is smaller than a threshold $P^*(t)$ corrected for multiple hypothesis testing (see section Methods), we validate the link between i and j and place it on the monopartite validated network of institutions. Otherwise, the link is discarded. In other words, the comparison is deemed statistically significant if the observed overlap would be an unlikely realization of the null hypothesis according to the significance level $P^*(t)$. This procedure is repeated for all pairs of institutions, resulting in the validated projection $\mathcal{V}(t)$ of the original network: a monopartite network whose generic element $V_{ij}(t) = 1$ if $P[o_{ij}(t)] < P^*(t)$, and 0 otherwise.

When applied to a historical database of SEC 13-F filings (see section Methods for details and Fig. 3.1 for the temporal evolution of the main dataset statistics), our method yields statistically validated networks of overlapping portfolios whose properties turn out to be related to the occurrence of the 2007-2008 global financial crisis. In particular, we propose to regard the average number of validated links for each institution as a simple measure of systemic risk due to overlapping portfolios. Such a measure gradually built up in years from 2004 to 2008, and quickly dropped after the crisis. Systemic risk has then been increasing since 2009, and at the end of 2013 reached a value not previously seen since 2007. Note that because there is only one large crisis in our dataset, we refrain from making strong claims about the systematic coincidence of highly connected validated networks and the occurrence of financial crises. We

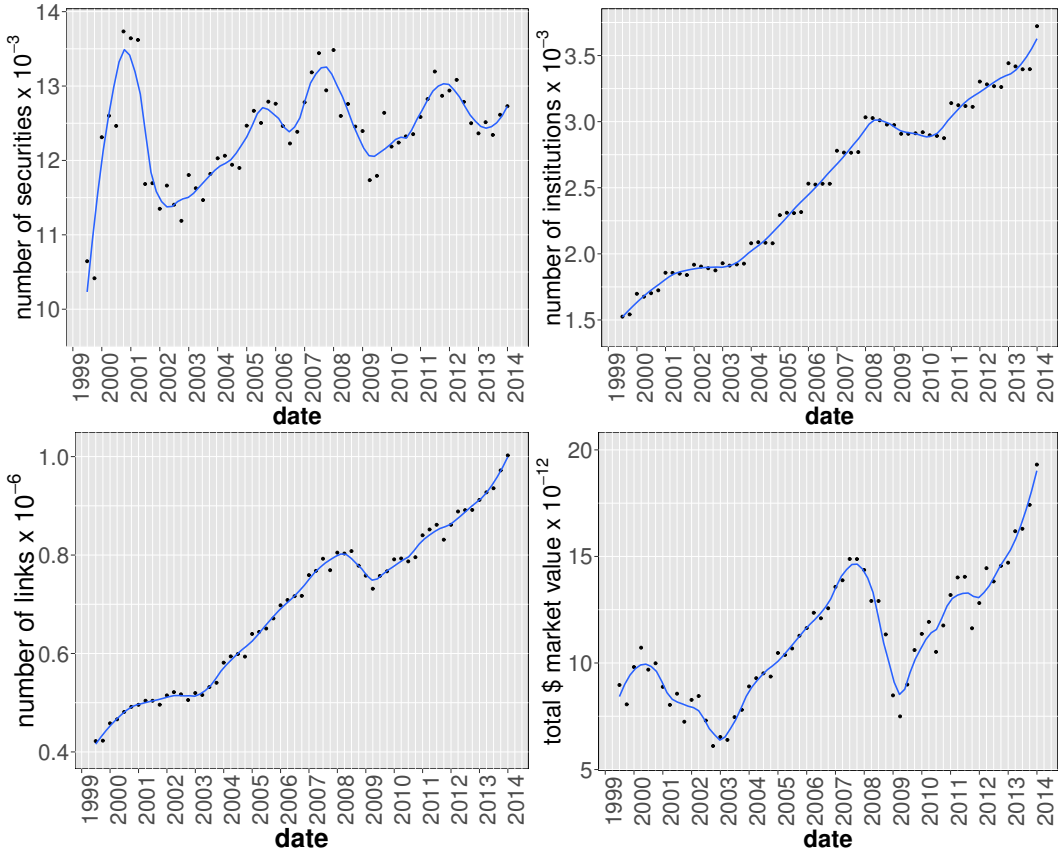

Figure 3.1: Temporal evolution of main aggregate quantities characterizing the bipartite ownership network. From left to right: number of institutions $|I(t)|$, number of securities $|S(t)|$, number of different ownership relations $L(t) = \sum_{is} A_{is}(t)$ and total market value $MV(t) = \sum_{is} W_{is}(t)p_s(t)$, where $p_s(t)$ and $\sigma_s(t) = \sum_i W_{is}(t)$ are the price and number of outstanding shares of security s at time t, respectively. Solid lines correspond to a locally weighted least squares regression (loess) of data points with 0.2 span.

also find that overlapping securities (i.e., those securities making up the validated overlaps) represent a larger average share of institutional portfolios, a configuration which would exacerbate the effect of fire sales. Additionally, we show that the presence of a validated link between two institutions is a good indicator of portfolio losses for these institutions in times of bearish markets, and of portfolio growth in times of bullish markets: validated links should indeed represent self-reinforcing channels for the propagation of financial distress or euphoria. More in general, we find that market trends tend to be amplified in the portfolios identified by the algorithm. Finally, we apply the validation procedure to the overlapping ownerships of securities to identify contagion channels between securities themselves, and observe a stable growth

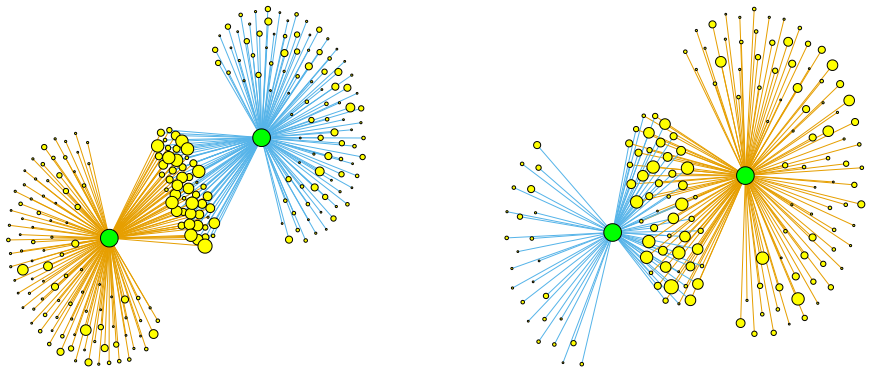

Figure 3.2: Two examples of institutions pairs (green nodes) with the securities they own (yellow nodes). The composition of each portfolio is denoted by different colors (blue and orange links). The symbol size of a security is proportional to the total number of its investors. Although both pairs in the plot have an overlap of 50 securities, the right pair is validated by the algorithm whereas the left pair is not. This is due to the fact that both the blue and orange portfolios on the right are smaller (the blue one in particular) and therefore under the BiCM null model the chance of having the same overlap of the pair on the left is considerably smaller.

of validated securities over the considered time span. This signals an ongoing, deep structural change of the financial market and, more importantly, that there are more and more stocks that can be involved in a potential fire sale. The presence of local maxima within this trend correspond to all periods of financial turmoil covered by the database: the *dot-com* bubble of 2001, the 2007-2008 global financial crisis and the 2010-2011 European sovereign debt crisis.

3.1 Results and Discussion

In order to properly understand the results of our validation method for overlapping portfolios, it is useful to provide a specific example. Fig. 3.2 shows two similar situations: two pairs of portfolios both owning 50 securities in common. Only the right pair is validated by our method, whereas the left pair is not. This happens because the portfolios in the right pair are of smaller size (especially the blue one) and the same overlap is therefore less likely to happen by chance. Hence, although the algorithm cannot directly take into account how much each institution is investing (particularly with respect to the total asset managed by the institution), it does so indirectly by taking into account the diversification of different portfolios (i.e., the degree of institutions). Validated pairs of portfolios indeed correspond to overlaps which constitute a

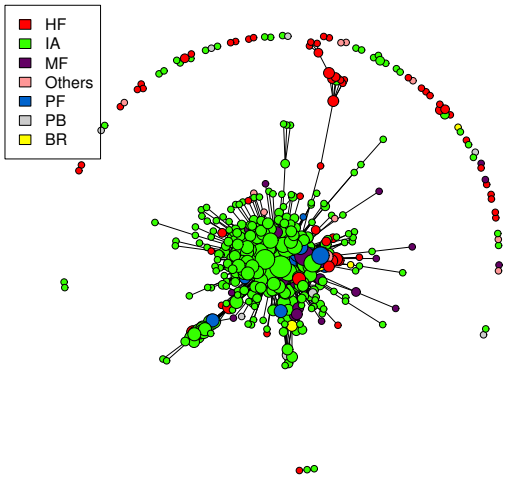

Figure 3.3: Validated networks of institutions at 2006Q4 (1293 institutions and 93602 validated links). Node colors label institution type, while their size is given by the logarithm of their degrees in the validated network: $d_i^V(t) = \sum_j V_{ij}(t)$. An institution is classified either as Broker (BR), Hedge Fund (HF), Investment Adviser (IA), Mutual Fund (MF), Pension Fund (PF), Private Banking (PB), or Other (i.e., without classification or belonging to a minor category).

considerable fraction of the total portfolio value of the pair. In short, pairs are validated when neither the diversification of the investments nor the degree of the securities are sufficient to explain the observed overlap. As we shall see later, the same mechanism is at play when we project the bipartite network on the securities side. In this case, since the degree of a security is a good approximation of its capitalization and of the dilution of its ownership (Zumbach, 2004; Eisler and Kertesz, 2006), the method will tend to validated links among securities whose ownership is relatively concentrated. Fig. 3.3 gives an overall picture of how the validated network looks like. In general, we observe the presence of multiple small clusters of institutions, together with a significantly larger cluster composed by many institutions linked by a complex pattern of significant overlaps.

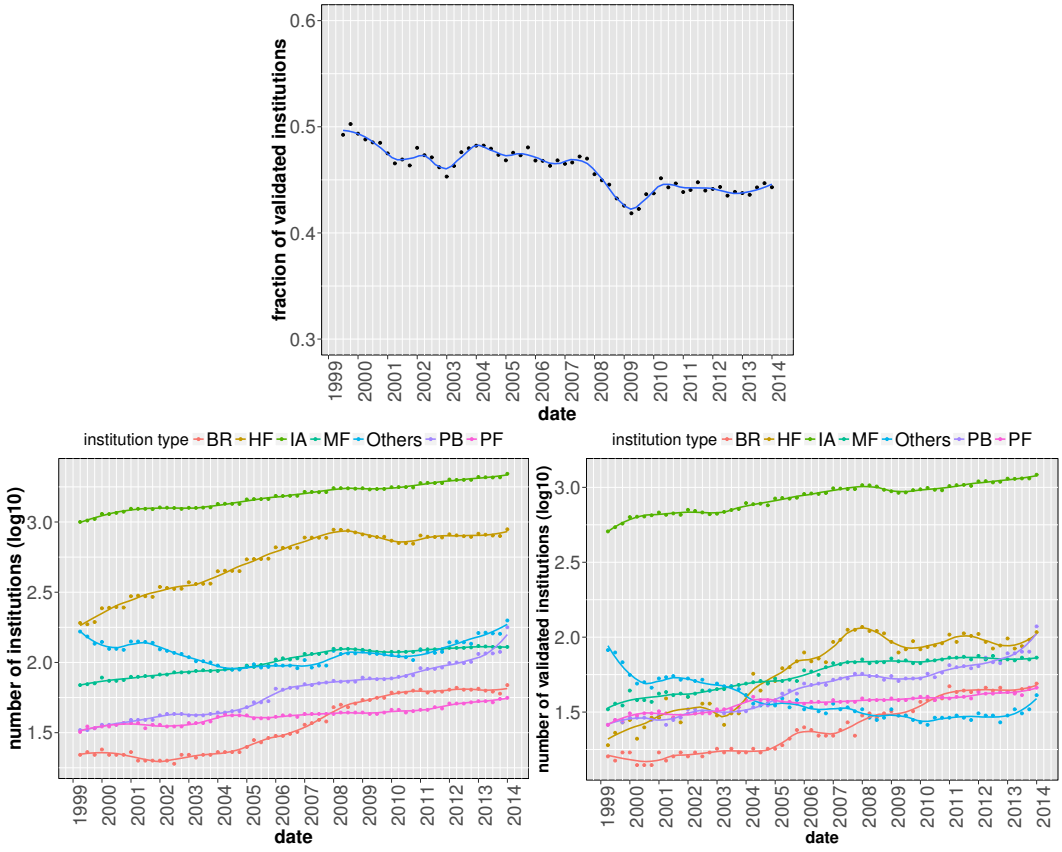

Figure 3.4: Fraction of institutions appearing in the validated network as a function of time (top panel); total number of institutions in the original bipartite network (bottom left panel) and number of validated institutions (bottom right panel) for the different institution types. Solid lines correspond to a locally weighted least squares regression (loess) of data points with 0.2 span.

3.1.1 Temporal evolution of the validated network of institutions

After these preliminary observations, we move to the temporal analysis of the structural properties of the whole validated network of institutions. In Fig. 3.4 we show the fraction of validated institutions (defined as the number of institutions having at least one validated link over the total number of institutions appearing in the ownership network) as a function of time. We also disaggregate data according to the type of institution and plot in this case both the number of validated institutions and the original number of institutions (we avoid to use directly their ratio for a better visualization). One sees that there is no particular pattern and the fraction of validated institutions is almost constant in time. By looking at disaggregated data, a few interesting things emerge. Investment Advisors account for the largest percentage of institutions and,

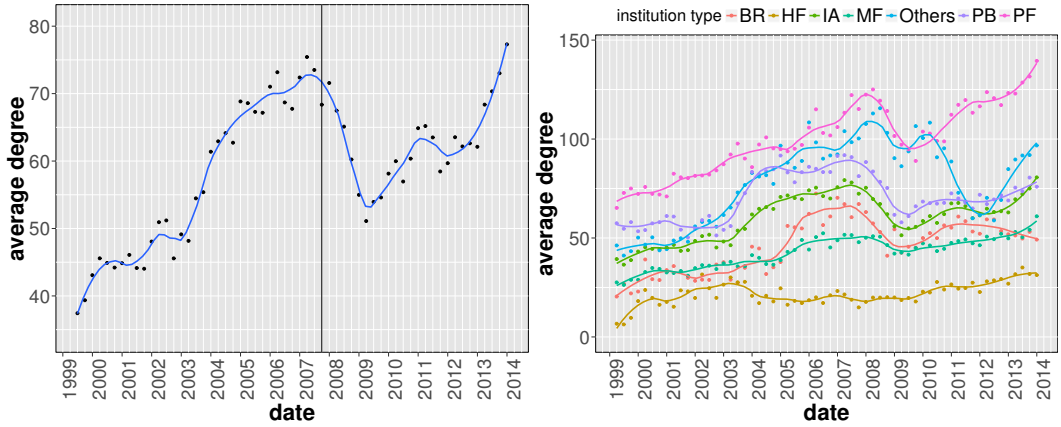

Figure 3.5: Average degree of institutions in the validated network as a function of time, aggregated (left panel) and separated for the different institutions type (right panel). The vertical line correspond to the date in which we observe the maximum total market value in the dataset just before prices started to fall during the financial crisis (see Fig. 3.1). Remarkably, we observe a slow but steady build-up of portfolios similarity with a clear acceleration in the years preceding the financial crisis and from 2009 onwards. Solid lines correspond to a locally weighted least squares regression (loess) of data points with 0.2 span.

more prominently, of validated institutions, followed by Hedge Funds and Mutual Funds. The most interesting behavior is however that of Hedge Funds in the validated networks: they are relatively under-represented until 2004, but after that their number displays a steep increase.

Fig. 3.5 displays the temporal evolution of the average degree in the validated network, which measures how much validated institutions are connected to each other. One clearly sees an overall increasing trend with a strong acceleration during the years preceding the financial crisis. In particular, the average degree reaches a maximum few months before prices started to fall. Furthermore, our results suggest that a similar process is taking place after 2009, a fact that might question the stability of financial markets nowadays. The right-hand side plot of Fig. 3.5 reports the same quantity for each category of institutions, which also has peaks just before the 2008 crash. The notable exception is Hedge Funds, whose average degree is roughly constant in time. In addition, the peak for Investment Advisors, Private Banking funds and Brokers occurs roughly 1-2 quarters before the global peak.

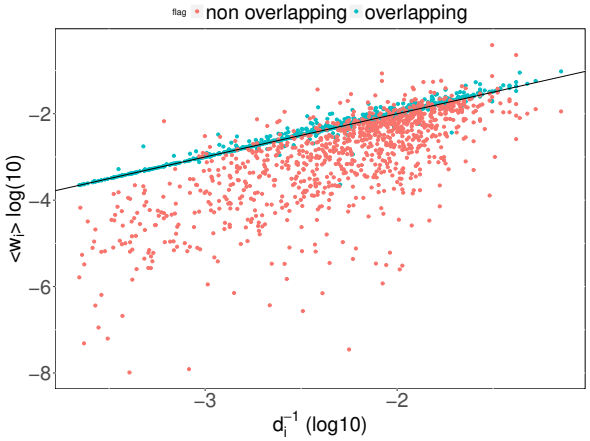

Figure 3.6: Scatter plot of the average share of securities market value in a portfolio $\langle w_i(t) \rangle = \sum_\alpha w_{i\alpha}(t)$ versus the inverse of the portfolio diversification $1/d_i(t)$ for each institution i. The average over all securities in a portfolio gives, by construction, the inverse of the institution's degree (corresponding to the straight line in the plot). Here we divide the average share over overlapping securities, securities in the portfolio belonging to the overlap with a validated neighbor) and non-overlapping securities (the complementary set). We clearly see that overlapping positions correspond to larger shares in the portfolio. The plot refers to 2006Q4, yet the same qualitative behavior is observed for other dates.

3.1.2 Validated overlaps vs portfolio size and security capitalization

A seemingly major shortcoming of using a binary holding matrix $\mathcal{A}(t)$ for validation purposes is that of not taking into account neither the concentration of ownership of a given security (i.e., which fraction of the outstanding shares a given institution is holding) nor the relative importance of different securities in a portfolio (*i.e.*, which percentage of the total portfolio market value a security is representing). These are clearly important types of information, since one would expect a mild price impact following the liquidation of the asset by an institution if the latter owns only a small fraction of that security's outstanding shares. Conversely, if the asset represents a considerable fraction of the portfolio market value, a price drop will have a stronger impact on the balance sheet of the institution. However, despite validating weighted overlaps $o_{ij}^W(t) = \sum_\alpha W_{i\alpha}(t) W_{j\alpha}(t)$ is more relevant than binary overlaps to identifying fire sales propagation channels, we cannot use the original weighted matrix $\mathcal{W}(t)$ in the validation procedure, as in this case it would be impossible to build an analytical null model—which would make the validation procedure extremely involving. Thus, we are forced to rely on binary overlaps. However, the dataset at our disposal allows us to check a posteriori the features of the portfolio positions which contributed to the

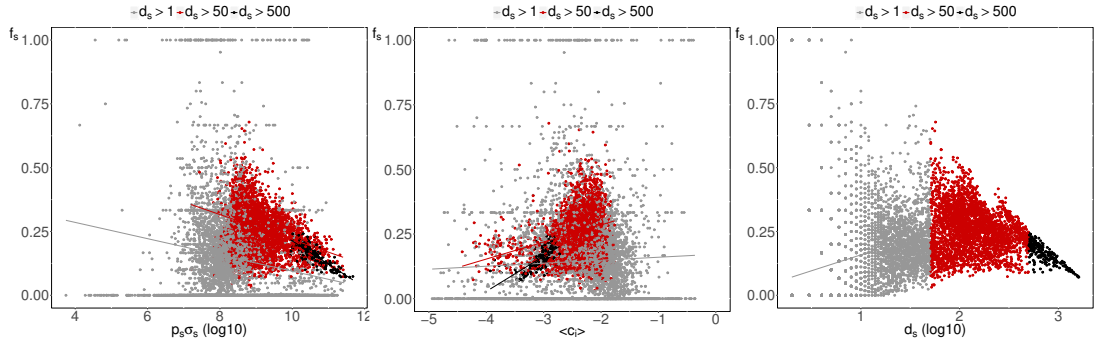

Figure 3.7: Scatter plots of the fraction $f_\alpha(t)$ of validated pairs of institutions owning a security α versus the security capitalization (left plot), its concentration (central plot) and number of owners (right plot). The straight lines show log-linear regressions of the data, divided according to securities degrees. Plots refer to 2006Q4, yet the same qualitative behavior is observed for other dates.

formation of validated links.

To this end, using the information about the price $p_\alpha(t)$ and outstanding shares $\sigma_\alpha(t)$ of different securities α at time t, we compute the fraction of the total market value of portfolio i represented by security α, namely $w_{i\alpha}(t) = p_\alpha(t)W_{i\alpha}(t)/\sum_x p_x(t)W_{ix}(t)$, and the fraction of outstanding shares of α held by institution i, namely $c_{i\alpha}(t) = W_{i\alpha}(t)/\sigma_\alpha(t)$. We apply this procedure to each position $W_{i\alpha}(t)$ of the bipartite ownership network in order to characterize the features of the positions belonging to validated overlaps. Fig. 3.6 shows that, on average, overlapping securities (*i.e.*, securities making up the validated overlaps) represent a larger share of the validated portfolio, namely 6% more than the average share given by the inverse of the degree.

In order to study the concentration of ownership of different securities we use the following procedure. Each security s belongs by construction to $d_\alpha(t)[d_\alpha(t)-1]/2$ pairs of overlapping portfolios, and we can compute which fraction $f_\alpha(t)$ of such pairs that are validated by the algorithm. We then compute for each security the total capitalization (as a proxy for the liquidity of the security) as well as the average ownership fraction per institution $\langle c_i(t)\rangle = \sum_i c_{i\alpha}(t)/d_\alpha(t)$ as a function of $f_\alpha(t)$. In Fig. 3.7 we show scatter plots of this quantities together with straight lines obtained from log-linear regressions. As one can see, the probability that any pair of institutions investing in the same asset are validated by the algorithm decreases as a function of the capitalization of the asset, increases as a function of the concentration (*i.e.*, with the average fraction of outstanding shares detained by an institution) and decreases as a function of the degree of the security. The relation is stronger for securities with higher degree, because of the larger number of available data points.

3.1.3 Distressed institutions in the validated networks

As a final test of the effectiveness of the validation procedure, we study the ability of the algorithm to retrieve (pairs of) institutions which are about to suffer significant losses. The dataset at our disposal indeed covers periods of financial distress (in particular the 2008 financial crisis) and it is in such periods that one would expect some institutions to incur fire sales. Then, if the algorithm does filter information in a useful way, the presence of a validated link between two institutions should represent a channel for the propagation of losses. Note that we do not attempt here to design a test for detecting self-reinforcing fire sales. Rather, we check if the presence of a validated link ultimately contains information on the occurrence of losses.

To this end, we construct for each date t the set $\mathcal{L}_n(t)$ of the n institutions experiencing the highest drop in portfolio value between t and $t + dt$ (which we refer to as "distressed" institutions). We first consider drops in absolute terms (i.e., the total dollar amount) which we believe is of macroeconomic significance and check the relation with portfolio returns later on. We use here $n = 300$ (roughly corresponding to 10% of the total number of institutions) and omit in the following the n subscript. We then compute the fraction $l(t)$ of distressed institutions with respect to the total number of institutions $I(t)$ and compare it with the fraction of distressed institution $l_{\mathcal{V}}(t)$ in the validated network. The ratio $G_I(t) = P[i \in \mathcal{L}(t) \,|\, i \in \mathcal{V}(t)]/P[i \in \mathcal{L}(t)] = l_{\mathcal{V}}(t)/l(t)$ then indicates if distressed institutions are over-represented in the validated network. Indeed, if $G_I(t) = 1$ the algorithm is not doing anything better than putting distressed institutions at random in the validated network, whereas, if $G_I(t) > 1$ we effectively gain information by knowing that a institution belongs to $\mathcal{V}(t)$. Similarly, we compare the fraction of links in the validated network which connect institutions that are both distressed with the fraction of such links when all overlapping pairs of institutions (i.e., all pairs whose portfolios having at least one security in common) are considered. The ratio between these two quantities, namely $R_I(t) = P[i,j \in \mathcal{L}(t) \,|\, V_{ij}(t) = 1]/P[i,j \in \mathcal{L}(t) | o_{ij}(t) > 0]$, can then be used to assess the effectiveness of the algorithm to establish a link between two distressed institutions in the validated network. Since all the positions in our dataset are long positions, it makes sense to relate $G_I(t)$ and $R_I(t)$ to an index that encompasses many securities. Fig. 3.8 shows these quantities as a function of the market return $r(t)$ between t and $t + dt$ as measured by the Russell 2000 index. Indeed, both ratios are correlated with the total loss, and are significantly larger than 1 when $r(t) \ll 0$ (R_I in particular reaches values close to 8 in periods of major financial distress). Notably, both ratio are close to 1 when the market loss is close to 0, and decline afterwards. This could be interpreted as the fact that, in times of market euphoria, overlapping portfolios turn into self-reinforcing bubbles.

When we repeat the same procedure for portfolio returns (i.e., we use portfolio returns

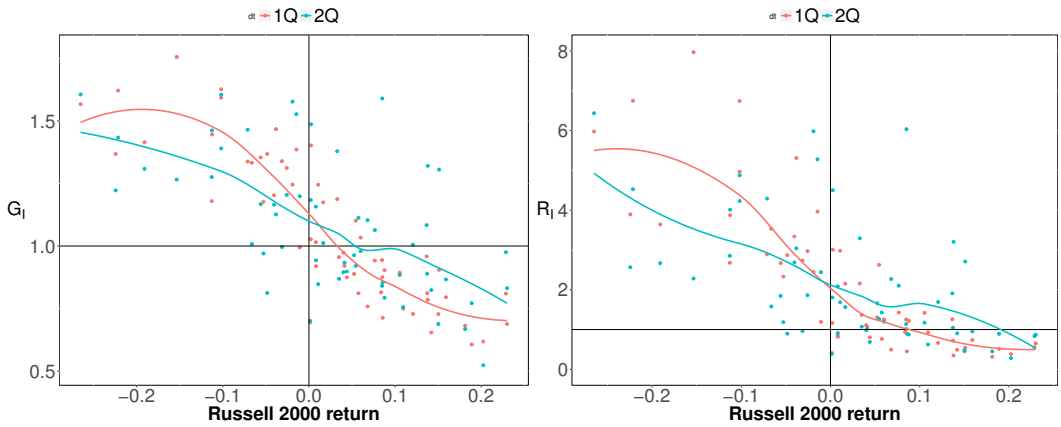

Figure 3.8: Scatter plots of the ratios G_I (left panel), i.e., the ratio between the probability of observing a distressed institution in the validated network and the a-priori probability of observing a distressed institution, and R_I (right panel), i.e., the ratio between the probability of observing a linked pair of distressed institutions in the validated network and the probability of observing a distressed pair of institutions when all overlapping portfolios are considered, versus the return $r(t)$ between t and $t + dt$ of the Russell 2000 index. Red points correspond to dt equal to one quarter, blue points to dt equal to two quarters; solid lines correspond to a locally weighted least squares regression (loess) of data points with 0.2 span. Panels are divided in four regions, corresponding to probabilities larger/smaller than one (i.e., distressed institutions over/under represented in the validated networks) and to $r(t)$ larger/smaller than zero (i.e., market contraction/growth).

to label institutions as distressed) we do not obtain meaningful results. This is however due to the fact that abnormal returns are in general observed for small portfolios for which we have few data points. Given the statistical nature of our method we cannot hope to correctly identify such situations for which a different (probably case by case) methodology is clearly needed. We can however take a simpler point of view and take for each time t all portfolios whose return is smaller (in absolute term) than a threshold r_{max} that we use as a parameter. We then use this subset to compute the average return of validated portfolios $< r >_{i \in \mathcal{V}(t)} (r_{max}, t)$ together with the average return of portfolios outside the the validated network $< r >_{i \notin \mathcal{V}(t)} (r_{max}, t)$. For a given value of r_{max}, we then have a scatter plot of these two quantities (one point for each date t) $< r >_{i \in \mathcal{V}(t)} (r_{max})$ versus $< r >_{i \notin \mathcal{V}(t)} (r_{max})$ which is well approximated by a straight line (see Fig. 3.9 left panel for an example). Note that with the significance threshold $P^*(t)$ used one has roughly half of the institutions in each set (see Fig. 3.4 left panel). Finally, we linearly regress $< r >_{i \in \mathcal{V}(t)} (r_{max}, t) = A < r >_{i \notin \mathcal{V}(t)} (r_{max}, t) + B$, and plot the value of the slope as a function of the threshold r_{max}. As one can see in the right panel of Fig. 3.9, the slope is significantly larger than 1 for values of the threshold up to roughly 30% in general, and up to 50% when we consider positive and nega-

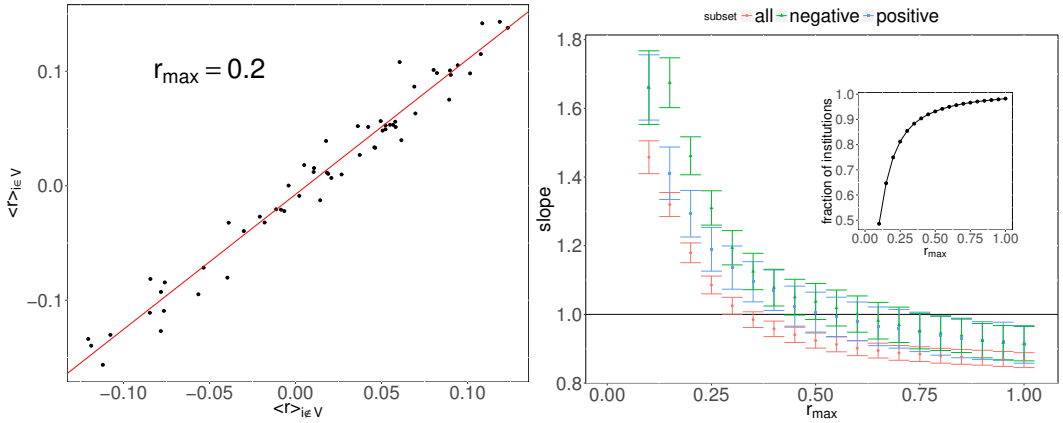

Figure 3.9: *Left panel:* average return of portfolios in the validated network vs average return in the complementary set. All returns for which $|r_i(t)| < r_{max}$ (here 0.2) are included. The straight line correspond to a linear regression of the data-points (one for each quarter). *Right panel:* value of the slope obtained as in the left panel as a function of r_{max} (see details in the text). When returns greater than roughly $30 - 40\%$ are excluded the slope is found significantly larger than 1. This indicates that portfolios in the validated network tend to have higher returns (in absolute terms) than their not validated counterpart. The inset shows the overall fraction of returns satisfying $|r_i(t)| < r_{max}$ as a function of r_{max}.

tive returns separately. In the latter case we first split for each date institutions with positive/negative returns and compute return averages in the validated and complementary set. The fact the the slope become slightly smaller than 1 for large values of r_{max} is putatively due to abnormal returns, most likely associated with small portfolios which tend to be outside the validated network. While this drawback is unavoidable given the statistical nature of our method, on the overall these results show that as long as abnormal returns are not considered, the returns of validated portfolios are on average greater (in absolute terms) than those of their not validated counterpart.

3.1.4 Buy and sell networks: the case of Hedge Funds

Before moving to the analysis of the validated network of securities, we illustrate another interesting application of our method. Our dataset allows us to build, for each date t, the buy (or sell) bipartite network, corresponding respectively to the positions acquired (or sold) by each institution between $t - dt$ and t: $A^{\text{BUY}}_{i\alpha}(t) = 1$ if $W_{i\alpha}(t) - W_{i\alpha}(t - dt) > 0$ and 0 otherwise; $A^{\text{SELL}}_{i\alpha}(t) = 1$ if $W_{i\alpha}(t) - W_{i\alpha}(t - dt) < 0$ and 0 otherwise. Validation of these bipartite networks then highlights the institutions that have updated their portfolios in a strikingly similar way. As a case study we consider

the Hedge Funds (HF) buy/sell networks, meaning that we only consider the positions bought or sold by HF (discarding all other links), and apply the validation procedure to these subnetworks. The focus on this particular subset of funds is motivated by the Great Quant Meltdown of August 2007, during which quantitative HF, in particular those with market neutral strategies, suffered great losses for a few days, before a remarkable (although incomplete) reversal (see, e.g., Khandani and Lo (2007)). In addition, we wish to investigate whether HF reacted in a synchronous way at the end of the 2000-2001 dot-com bubble.

As for Fig. 3.3, Fig. 3.10 shows that the fraction of HF validated in the buy/sell network is roughly constant in time, with however some more interesting local fluctuations (especially in the years around 2008). For what concerns the average number of neighbors in the validated network, one sees that the fluctuations of the sell networks lag by 3 months those of the buy networks: indeed, the cross-correlation is maximal at such a lag, and is quite high (0.8). This is possibly due to the fact that the typical position holding time of HF is smaller than 3 months: what has been bought will have been sold 3 months later. Notably, the right panel of Fig. 3.10 points to the fact that buy networks are more dense on average than sell networks. This is also reflected in the autocorrelation of the average number of neighbors, which decrease faster for sell networks. Since our dataset only contains long positions, we can only conclude that HF are more synchronized when they open long positions, and liquidate them in a less synchronized way.

Using as a first approximation the average number of validated neighbors per fund in order to assess the synchronicity of the HF actions, we clearly observe significant increasingly synchronized buying patterns after the top of the dot-com bubble. There may be two reasons for buying at this date: either the strategies of the HF were not aware of the bubble burst and were still using trend-following, or they took advantage of the burst to buy stocks at a discount. Noteworthy, synchronized selling lags on buying, and was overall less intense. Concerning the period of the global financial crisis, we observe one buy peak at 2007Q3, and one sell peak at 2007Q4. The first peak may indeed be related to the Big Quant Meltdown of August 2007. However, the so-called long-short market neutral funds that were forced to liquidate their positions should appear in the sell network, not the buy one. This would have been observed if that crisis had happened at the end of a trimester. Unfortunately, there is an almost two-month delay between the meltdown and the reporting, which probably hides the event. At all rates, the meltdown acted as a synchronization event, as the buy network density is clearly an outlier at the end of September 2007: HF have therefore acquired significantly similar long positions in their portfolios during the same quarter, and then, expectedly, liquidated them by the end of the next trimester.

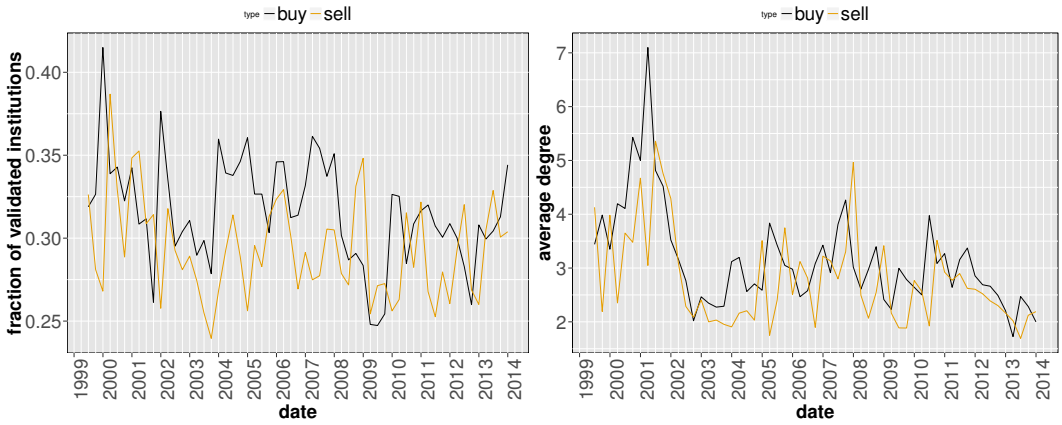

Figure 3.10: Fraction of validated institutions (left) and average degree in the validated network (right) for the buy/sell subnetworks of Hedge Funds. Here the original bipartite network is made up only of Hedge Funds and the positions they acquire/sell between $t - dt$ and t.

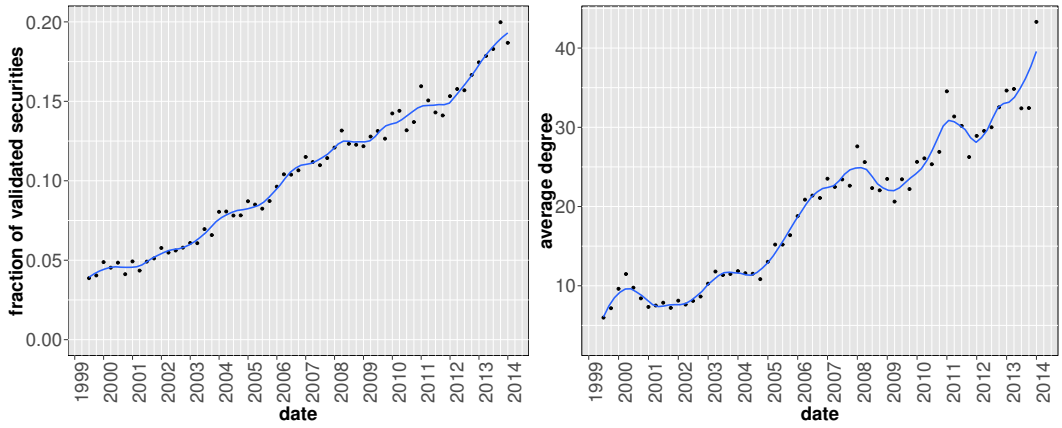

Figure 3.11: Fraction of securities appearing in the validated network (left panel) and their average degree in the validated network (right panel) as a function of time. Differently from the validated network of institutions, here the number of validated securities grows steadily in time. Yet, the number of validated links grows at a faster peace, as demonstrated by the increasing average degree. Solid lines correspond to a locally weighted least squares regression (loess) of data points with 0.2 span.

3.1.5 Temporal evolution of the validated network of securities

In this section we finally use our method to detect statistically significant common ownerships of securities, in order to identify contagion channels between securities themselves. Thus, we apply the validation procedure to the security ownership overlap $\tilde{o}_{\alpha q}(t) = \sum_i A_{i\alpha}(t)A_{iq}(t)$ (instead of the institution portfolios overlap $o_{ij}(t) =$

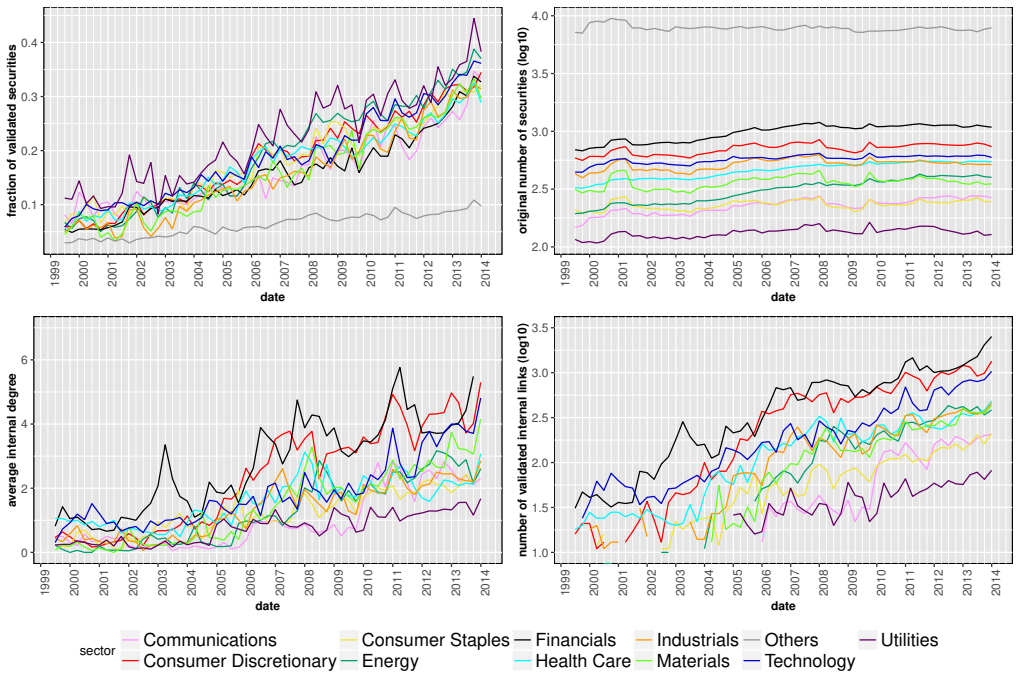

Figure 3.12: Statistics of the validated networks of securities, disaggregated by BICS category: fraction of validated securities (upper left panel), total number of securities in the original bipartite network (upper right panel), average internal degree (lower left panel) and internal links (lower right panel) in the validated network. The latter two quantities are obtained by considering only validated links connecting securities of the same category. Security categories that are more internally connected are Financials (which includes the following level 2 sectors: Banking, Commercial Finance, Consumer Finance, Financial Services, Life Insurance, Property & Casualty, Real Estate) and Consumer Discretionary (which includes: Airlines, Apparal & Textile Products, Automotive, Casinos & Gaming, Consumer Services, Distributors, Educational Services, Entertainment Resources, Home & Office Products, Home Builders, Home Improvements, Leisure Products, Restaurants, Travel & Lodging).

$\sum_\alpha A_{i\alpha}(t)A_{j\alpha}(t)$). The presence of a validated link between two securities then reflects the fact that they share a significantly similar set of owners, which again translates into a potential contagion channel through fire sales. Fig. 3.11 shows the temporal evolution of aggregate features of the validated network projection on securities. Contrarily to the case of the institutional projection (Fig. 3.4 and 3.5), here we observe a stable growth of validated securities: there are more and more stocks that can be involved in a potential fire sale (or closing down of similar institutions). Moreover, as testified by the growth of the average degree of validated securities, the validated network becomes denser, signaling the proliferation of contagion channels for fire sales. Note the presence of local maxima that correspond to all major financial crises covered by the

database: the *dot-com* bubble of 2001, the global financial crisis of 2007-2008 and the European sovereign debt crisis of 2010-2011. As for the case of institutions, the similarity pattern of securities ownerships is maximal at the end of the considered time span.

The fact that the average degree of the validated network of securities keeps growing boils down to the fact that institutions choose securities, not the opposite. While the number of institutions in our dataset has increased over the years, the number of securities has been roughly constant. If a new institution selects at random which assets to invest in, then the average degree of the securities network would stay constant. This is not the case, if only because of liquidity constraints. Therefore, on average, the portfolio of a new institution is correlated with the ones of pre-existing institutions.

In order to detect if the observed patterns concern peculiar classes of securities, we perform an analysis of the validated network distinguishing securities according to the Bloomberg Industry Classification Systems (BICS) —which rests on their primary business, as measured first by the source of revenue and second by operating income, assets and market perception. Each security thus belongs to one of the following sectors: Communications, Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Industrials, Materials, Technology, Utilities (or other). In particular, we try to detect whether securities of the same category tend to be connected together in the validated network. To this end, we denote as *internal* a validated link connecting two securities with the same BICS label, and we compute the internal degree as the degree of a security restricted to internal links. As Fig. 3.12 shows, the categories of securities that are more internally connected are (notably) Financials and, to a lesser extent, Consumer Discretionary. This does not mean that portfolio overlaps concentrate on these categories, but rather that relatively more contagion channels exist within securities belonging to them.

3.2 Discussion

In this work, we have proposed an exact method to infer statistically robust links between the portfolios of financial institutions based on similar patterns of investment. The method solves the problem of evaluating the probability that the overlap of two portfolios of very different size and diversification is due to random allocation of assets of very different market capitalization and number of owners. The use of an appropriate null hypothesis provided by the bipartite configuration model (Saracco et al., 2015) considerably improves the statistical significance of the detected features of the validated networks. Note that the method is general, and can be applied to any bipartite network representing a set of entities sharing common properties (e.g.,

membership, physical attributes, cultural and taste affinities, biological functions, to name a few) and where the presence of (unlikely) similar sets of neighbors is of interest

The present study then points to the conclusion that, just before financial crises or bubble bursts, the similarity of institutions holdings increases markedly. Perhaps worryingly for equity markets, the proposed proxy of fire sale risk, having reached a peak in 2008 and subsequently much decreased, has been increasing again from 2009 to the end of our dataset (2013) up to levels not seen since 2007. Despite our method relies on binary ownership information, we also found that on average overlapping securities correspond to larger shares of validated portfolios, potentially exacerbating fire sales losses. In addition, the proposed validation method can effectively retrieve the institutions which are about to suffer significant losses in times of market turmoil (when validated links are the channels for which liquidation losses propagate), as well as those with the highest growth in times of market euphoria (when overlapping portfolios turn into self-reinforcing bubbles). Finally we show that the number of securities that can be involved in a potential fire sale is steadily growing in time, with an even stronger proliferation of contagion channels.

In this work we have only investigated patterns of portfolio overlap, not the probability that they lead to fire sales. This is a more complicated problem for which other datasets and econometric techniques are needed. However, even if we cannot draw any strong implication from our findings, all the analysis we performed confirm the coherence of our method and suggest that overlapping portfolios do play a role in financial turmoils. Furthermore, the relationship between holdings and future portfolio changes must be better characterized. Indeed, even if two institutions with different strategies converge to a similar portfolio, this does not imply that they will update the latter in the same way and at the same time. However, it is likely that part of the institutions follow in fine) equivalent strategies, which implies portfolio overlap and subsequent increased risk of fire sales, which triggers further leverage adjustment, as pointed by Caccioli et al. (2014); Cont and Wagalath (2014). Finally, it will be useful to repeat our analysis on larger datasets so as to encompass other bubbles and crises, and to examine difference in investment patterns across various markets.

3.3 Methods

3.3.1 Dataset

We extracted 13-F SEC filings from the Factset Ownership database from 1999Q1 to 2013Q4, covering institutions valued more than 100 million dollars in qualifying assets which must report their long positions to the SEC at the end of each trimester. As

the 13-F dataset contains only positions greater than 10000 shares or $ 200000, very small positions are already filtered out. The dataset is composed of a set $I(t)$ of approximately 1500÷3500 institutions, holding positions from a set $M(t)$ of securities, whose size fluctuates around 12500 (see Fig. 3.1). Note that the portfolios of sub-funds are merged into a single report. In addition to the raw ownership data, our dataset is complemented by meta-data about both institutions and securities.

3.3.2 Significance level under multiple tests

In order to choose an appropriate threshold (the significance level) $P^*(t)$ to be used in the validation procedure, we have to account for the multiple hypothesis tested (corresponding to the number $n_{pairs}(t)$ of possible pairs of institutions having a nonzero overlap). Here we use the rather strict Bonferroni correction Miller (1981), meaning that we set the threshold to $P^*(t) = \epsilon / n_{pairs}(t)$. Note that the choice of the significance level still leaves some arbitrariness. While results presented in the paper are obtained with $\epsilon = 10^{-3}$, we have tested our method with various values of ϵ, and employed also the less-strict false discovery rate (FDR) criterion Benjamini and Hochberg (1995), without finding major qualitative differences. In fact, while the final size of the validated network clearly depends on the threshold, the relative temporal changes of the network statistics are much less affected by the particular value used.

3.3.3 Resolution problems for the hypergeometric distribution approach

As stated in the Introduction, the approach proposed in Tumminello et al. (2011) to divide the original bipartite network into homogeneous subnetworks of securities has some intrinsic limitations, especially when securities are characterized by a strongly heterogeneous number of investors (as it generally happens in stock market data). In this circumstance, in fact, the splitting procedure often translates into almost empty subsets—especially for securities held by a large number of investors. In these subsets, overlaps can assume only a few values, bounded by the limited number of securities considered, resulting in a handful, spaced-out possible outcomes for the p-values. The problem then arises with the use of a global threshold corrected for multiple hypothesis testing. In fact, since institutions are compared on the many subnetworks of securities with the same degrees, $n_{pairs}(t)$ scales as $I^2(t) d_\alpha^{max}(t) \equiv I^2(t) \max_\alpha d_\alpha(t)$: the validation threshold becomes extremely small for large and heterogeneous systems and vanishes in the infinite size limit. These issues lead to a serious problem of resolution, since $P^*(t)$ is too small to validate even the smallest non-zero p-value in most of the subnetworks. As a result, the validated network becomes almost empty by construction. Overall, while the method proposed in Tumminello et al. (2011) works well

for small networks with little degree heterogeneity, the same approach is not feasible in the case of large scale and highly heterogeneous networks.

3.3.4 p-values from the Bipartite Configuration Model

Determining the probability distributions used Eq. (3.1) requires to solve a technical problem caused by the heterogeneity of both institutions and securities. For example, it is hard *a priori* to compare a portfolio with very few assets and one with very many assets. However, the bipartite configuration model (BiCM) Saracco et al. (2015) provides a null network model suitable for these kind of situations. We remand the reader to Saracco et al. (2015); Squartini and Garlaschelli (2011); Park and Newman (2004) for more details on the method. In the following we will omit the explicit time dependence of the quantities considered, since the same procedure is repeated for each date.

In a nutshell, the BiCM prescribes to build the null model simply as the ensemble Ω of bipartite networks that are maximally random, under the constraints that their degree sequences of institutions and security is, on average, equal to the one of the original network. This is achieved through maximization of the Shannon entropy of the network subject to these constraints, that are imposed through a set of Lagrange multipliers $\{\theta_i\}_{i=1}^{I}$ and $\{\theta_s\}_{s=1}^{S}$ (one for each node of the network). Solving the BiCM means exactly to find these multipliers, that quantify the abilities of nodes to create links with other nodes. Thus, importantly, nodes with the same degree have by construction identical values of their Lagrange multipliers. Once these multipliers are found, the BiCM prescribes that the expectation values within the ensemble of the network matrix element $\langle A_{i\alpha}\rangle_\Omega$, i.e., the ensemble probability $Q_{i\alpha}$ of connection between nodes i and s, is given by:

$$\langle A_{i\alpha}\rangle_\Omega \equiv Q_{i\alpha} = \frac{\theta_i\theta_\alpha}{1+\theta_i\theta_\alpha}, \tag{3.2}$$

and the probability of occurrence $\mathcal{Q}(\mathcal{A})$ of a network $\mathcal{A}$ in Ω is obtained as the product of these linking probabilities $Q_{i\alpha}$ over all the possible $I \times M$ pairs of nodes. In other words, links are treated as independent random variables, by defining a probability measure where links correlations are discarded. The key feature of the BiCM model is that the probabilities $\{Q_{i\alpha}\}$ can be used to directly sample the ensemble of bipartite graphs and to compute the quantities of interest analytically. We can thus use the matrix $\mathcal{Q}$ to compute the expectation values of portfolios overlap between two institutions i and j as:

$$\langle o_{ij}\rangle_\Omega = \sum_{\alpha\in\mathcal{S}} Q_{i\alpha}Q_{j\alpha}, \tag{3.3}$$

or to compute the probability distribution $\pi(\cdot|d_i, d_j)$ of the expected overlap under the null hypothesis of random connections in the bipartite network—which, according to the BiCM prescription, only depends on the degrees of institutions i and j. Indeed, $\pi(\cdot|d_i, d_j)$ is actually the distribution of the sum of M independent Bernoulli trials, each with probability $Q_{i\alpha}Q_{j\alpha}$. This distribution can be computed analytically using a Normal approximation of the Poisson-Binomial distribution Hong (2013). This approach has been developed by Saracco et al. (2016) in parallel with our research. Here we discuss instead an exact and optimized numerical technique to compute $\pi(\cdot|d_i, d_j)$. Indeed, the computational complexity of the numerics can be substantially reduced by recalling, again, that $Q_{i\alpha} \equiv Q_{i\alpha'}$ if $d_\alpha \equiv d_{\alpha'}$ $\forall i$: connection probabilities only depends on nodes degree values. This is an important observation, which translates into the following statement: the expected overlap between any two institutions i and j restricted to the set of securities with a given degree follows a binomial distribution with probability $Q_{i\alpha}Q_{j\alpha}$ (where α is one of these securities) and number of trials equal to the cardinality of such set. More formally, if $\{\tilde{d}_h\}_{h=1}^{d_s^{max}}$ denotes the set of different degrees within securities, $\tilde{n}_h$ is the number of securities having degree $\tilde{d}_h$, h is any security having degree $\tilde{d}_h$, and if we define $q_{ij}^h = Q_{ih}Q_{jh}$, then the expected overlap $\langle o_{ij}^h \rangle_\Omega$ between institutions i and j restricted to securities having degree $\tilde{d}_h$ follows the binomial distribution

$$\pi_h(x|\tilde{n}_h, q_{ij}^h) = \binom{\tilde{n}_h}{x} [q_{ij}^h]^x [1 - q_{ij}^h]^{\tilde{n}_h - x}. \tag{3.4}$$

The overall distribution $\pi(\cdot|d_i, d_j)$ can now be more easily obtained as the sum of (much fewer than S) binomial random variables Butler and Stephens (1993): if $\pi_{\leq h}(\cdot|d_i, d_j)$ is the distribution of the overlap restricted to securities with degree smaller or equal than h, we have

$$\pi_{\leq h}(x|d_i, d_j) = \sum_{k=0}^{x} \pi_{\leq h-1}(x - k|d_i, d_j)\pi_h(x|\tilde{n}_h, q_{ij}^h) \tag{3.5}$$

and $\pi(\cdot|d_i, d_j) = \pi_{\leq d_s^{max}}(\cdot|d_i, d_j)$. For this computation, it is useful to recall the peculiar recurrence relation of the binomial distribution: starting from $\pi_h(0|\tilde{n}_h, q_{ij}^h) = [1 - q_{ij}^h]^{\tilde{n}_h}$, each subsequent probability is obtained through:

$$\pi_h(x|\tilde{n}_h, q_{ij}^h) = \frac{\tilde{n}_h - x + 1}{x} \frac{q_{ij}^h}{1 - q_{ij}^h} \pi_h(x - 1|\tilde{n}_h, q_{ij}^h). \tag{3.6}$$

Once the distribution $\pi(\cdot|d_i, d_j)$ is obtained, the p-value $P(o_{ij})$ can be associate to the overlap o_{ij} using Eq. (3.1), and the corresponding link can be placed on the validated monopartite network provided that $P(o_{ij}) \leq P^*$. Note that since this computation is made on the whole network, i.e., considering all the securities, we have a fairly large

spectrum of possible p-values. Thus, also if we still use a threshold depending on the number of hypothesis tested (which however now scales just as I^2), we have a much higher resolution than in Tumminello et al. (2011), and can obtain non-empty and denser validated networks.

Chapter 4

Collective rationality and functional Wisdom of the Crowd in far-from-rational institutional investors

4.1 Introduction

The collective ability of a non-expert crowd to accurately estimate an unknown quantity is known as the "Wisdom of the Crowd" (Surowiecki, 2005) (WoC thereafter). In many situations, the median or average estimate of a group of unrelated individuals is surprisingly close to the true value, sometimes significantly better than those of experts Galton 1907; Hill and Ready-Campbell 2011; Landemore 2012; Nofer and Hinz 2014. Understanding how, why and when WoC works is thus an important research topic. By far the most important condition under which WoC may emerge is diversity (Hong and Page, 2004; Surowiecki, 2005; Davis-Stober et al., 2014). On the other hand, social imitation is detrimental as herding may significantly bias the collective estimate (Lorenz et al., 2011; Muchnik et al., 2013), while averaging several guesses of each individual (known as the crowd within) may lead to sharper estimations (Vul and Pashler, 2008; Steegen et al., 2014).

In many economic systems, human beings must act according to their estimates of some quantities (such as the value of an item, of retirement plans, etc.), often each in his own way, with his own tool, experience and knowledge. Mainstream Economics takes a radical theoretical short-cut by assuming perfect individual rationality, i.e. that their estimates are the perfect, hence the same ones. This is the so-called the representative agent approach (Hartley and Hartley, 2002). We argue here that in these

systems, WoC applies, i.e., that the average estimate of economic agents may be remarkably close to the rational, optimal individual choice. In other words, the optimal individual choice may only be found in some economic systems when the diversity of individual decisions is averaged out. This is what we call collective rationality in the following and this is why it is worth extending the reach of WoC to economic theory. When it applies, it is a consistent aggregation of possibly inconsistent individual estimates (Hogarth, 1978). Aggregation of quite diverse individual actions, especially in a dynamic context where expectations are continuously revised, is still an open problem (Kirman, 1992).

Although almost all known examples of WoC are about a single number or coordinate, there is no reason why WoC could not be found for whole functional relationships between several quantities, which abound in economic systems. For example, Haerdle and Kirman analyse the prices and volume of many transactions in Marseille fish market: while the relationship between these two quantities is rather noisy, the market self-organises so that when more fish are sold, prices are lower, as revealed by a local average (Härdle and Kirman, 1995). More generically, many simple relationships found in Economics textbooks may only hold on average, but not for each agent or each transaction.

Finance is an interesting candidate for WoC because their competitive nature has two important consequences which are known to be two pre-conditions of existence of WoC (Hong and Page, 2004): competition forces market participants to be heterogeneous (Arthur, 1994), which ensures a large diversity amongst them, and it is a strong incentive for market participants not to behave randomly and try to do their best; some of them even try to build optimal portfolios from incomplete information.

Asset price efficiency is an obvious instance of WoC in Finance: it means that current prices, determined by the actions of many traders, are the best possible estimates on average and fully reflect all available information (Malkiel and Fama, 1970; Malkiel, 2003; Fama, 1998). If one departs from the estimation of a single number, thus, from the usual WoC, more complex relationships may be related to rational benchmarks. Since large funds do perform portfolio optimization according to their own cost function, we shall focus on their portfolio structure. Diversity of opinions, a crucial requirement for WoC, is also related to the fact that market participants have diverse constraints, some of them externally imposed (laws, regulations, exchange rules, etc.) and self-imposed (computational and mathematical methods, performance and risk objectives, benchmarks, etc.). Thus, the question here is whether their collective behaviour may be related to an unbeknownst or unreachable optimal portfolio; in practice, we perform a local average of the amount under management of funds as a function of the number of assets in their portfolio. Fortunately, this implies that we do not need to understand the minute details of the composition of all portfolios and can focus on average of

simple quantities instead. Our findings imply first that the concept of WoC holds on a more generic level, intuitive to many economists. which has profound implications for the role of rationality in economics modelling. On the one hand they vindicate the importance of rationality as a collective outcome, on the other hand our results clearly illustrate how individual rationality is not only unneeded, but also a fairy tale when making complex decisions with imperfect tools.

Another important result comes from the respective importance of two types of constraints. On the one hand, as explained above, these institutional investors face many regulatory, legal, and self-inflicted constraints, whose diversity explains in part the large individual deviations from the rational benchmark; however, these constraints have no effect on the ability of the population to approximate the rational benchmark. On the other hand, we find that some funds must invest into many more assets than the rational benchmark would recommend; this is due to the real-life constraint that some assets are not traded enough for a fund to have negligible impact on their prices. We propose a minimal model of how these funds manage to circumvent this practical problem.

4.2 Wisdom of the Crowd

Let us define some necessary quantities to be more precise. At quarter q, fund i has capital $W_i(q)$ which is invested into $n_i(q)$ securities among $M(q)$ existing ones. As a result, each security α, whose capitalization is denoted by $C_\alpha(q)$, is found in $m_\alpha(q)$ portfolios. The explicit time dependence is dropped hereafter unless needed for the sake of clarity.

The only quantity defined above which depends on asset allocation strategies of fund i is n_i, the number of securities it chooses to invest in. Our main hypothesis is thus that WoC is found in the average relationship between n_i and W_i. A simple rational benchmark is proposed by de Lachapelle and Challet (2010) : when a fund with capital W_i is able to invest the same amount in each of the n_i chosen securities and if the transaction cost does not depend on the security, then the optimal n_i is such that

$$W_i \propto n_i^\mu. \tag{4.1}$$

where the exponent μ is determined by the transaction costs fee structure: reference de Lachapelle and Challet (2010) shows that if the cost for a transaction of value W is proportional to W^δ, then $\mu = (2-\delta)/(1-\delta)$. For example, proportional transaction costs ($\delta = 1$) lead to $\mu = 1$, while a fixed cost per transaction ($\delta = 0$) corresponds to $\mu = 2$, which was found for individual investors and asset managers (with much

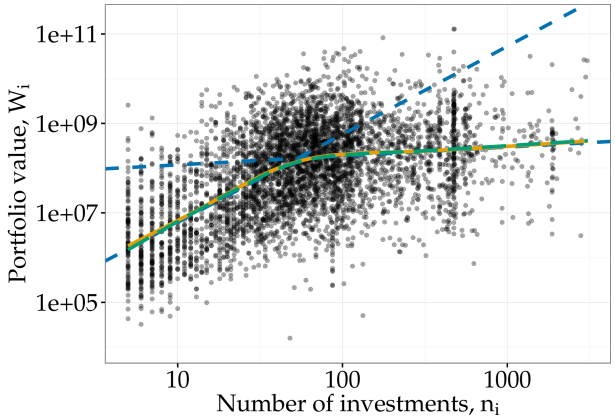

Figure 4.1: Scatter plot of the mark-to-market value W_i of fund i as a function of its number of investments n_i. Each black dot represents a single fund. Orange curve: robust locally weighted regression fit. Dashed blue lines: power-law fits of the small and large n regions. Green curve: robust locally weighted regression fit of the simulated model.

smaller portfolios than the large funds studied here). Allowing for individual fluctuations, Eq. (4.1) becomes $\log W_i = \mu \log n_i + \epsilon_i$, where ϵ_i has zero average. Denoting local average of x_i by x, the local average of Eq. (4.1) yields

$$W \propto n^{\mu}. \tag{4.2}$$

We will test the occurrence of WoC from the validity of Eq. (4.2). More precisely, our hypothesis is that if the effective transaction cost per transaction is the same for all assets, then (i) funds are able (and strive to) to build equally-weighted portfolios and (ii) then WoC is equivalent to the fact that Eq. (4.2) holds; if in addition the effective transaction cost is a flat fee, then (iii) $\mu = 2$.

The effective transaction cost for a given asset includes one's impact on the price, which becomes non-negligible if the value of the transaction is not a small fraction of the value exchanged daily. To avoid too large an impact, larger funds, on average, tend to spread their investments on a greater number of assets, which violates the equally-weighted portfolio hypothesis. As a result, the local average W is expected to increase more slowly as a function of n in the large n region; equivalently, the exponent μ is expected to be smaller than 2 in this region. In summary, two different regimes should emerge: one with $\mu_<$ for small n and one with $\mu_> < \mu_<$ for large n.

Figure 4.1 plots W_i versus n_i in logarithmic scales: a cloud of point emerges, with a roughly increasing trend. The large amount of noise confirms the great diversity of fund allocation strategies. WoC may only appear in a local average: we computed

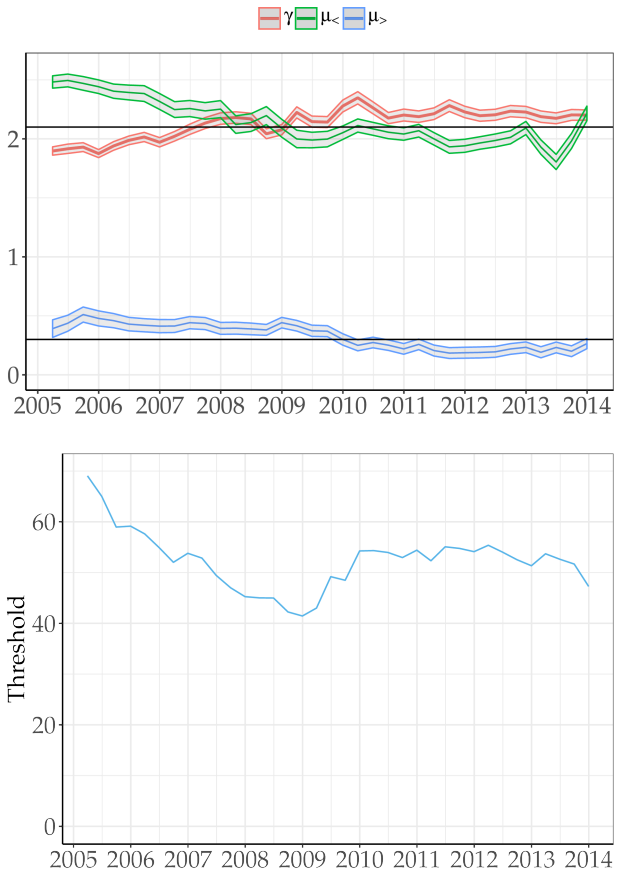

Figure 4.2: Top: temporal evolution of the coefficients $\mu_<$, $\mu_>$ and γ. Bottom: evolution of the value of the cross-over point n^* between the two regions as a function of time.

a locally weighted polynomial regression Cleveland et al. (1992) (refered to as local average henceforth) with the `stats::loess` function of `R`. As expected, two distinct regions appear. In each of them, the local regression follows a roughly linear behaviour.

The cross-over point n^* between the two regions is algorithmically determined for each quarterly snapshot (see S.I.); it is relatively stable as time goes on (see Fig. 4.2). The two exponents $\mu_<$ and $\mu_>$ are also quite stable as a function of time as well (see Fig. 4.2); their time-averages $\overline{\mu_<} \simeq 2.1 \pm 0.2$ and $\overline{\mu_>} \simeq 0.3 \pm 0.1$ are markedly different, which points to distinct collective ways of building portfolios in these two regions.

Let us now check that the funds have or strive to have equally-weighted portfolios. In that case, the investment fractions $p_{i\alpha} = W_{i\alpha}/W_i$ are well approximated by $1/n_i$ when $W_{i\alpha} > 0$, or, equivalently, the diversity of $p_{i\alpha}$ (as a function of α) is small. The latter may be summarized by the scaled Shannon Entropy $S_i = -\frac{1}{\log_2 n_i} \sum_\alpha p_{i\alpha} \log_2 p_{i\alpha}$, which

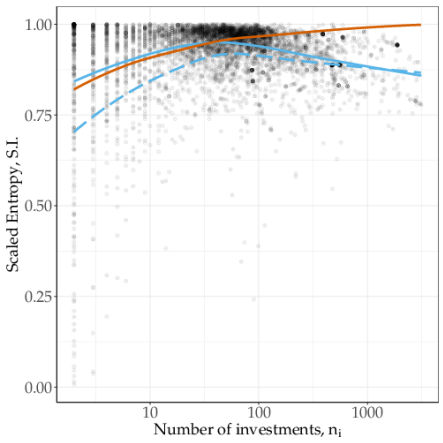

Figure 4.3: Scaled Shannon entropy S_i as a function of the number of investments n_i for all the funds on 2013-03-31 (circles). Blue line: local average S. Red line: local average S_{MC} from Monte-Carlo simulations reproducing the effect of asset price fluctuations over three months on the entropy of initially equally-weighted portfolios. Dashed blue line: local average of the scaled entropy $S_{restricted}$ restricted to unchanged portfolio positions from the previous time step properly rescaled to account for the difference in the number of asset (see text).

equals 1 and is maximal when all the non-null $p_{i\alpha}$ are equal. Figure 4.3 reports the scaled entropy S_i of all the funds for a given time snapshot, together with the local average S. The latter increases up to about $n \simeq n^*$ and then decreases. The fact that $S < 1$ is due in part to price fluctuations: even if fund i builds an equally-weighted portfolio at time q (thus $S_{i,q} = 1$), $S_{i,q+\delta q} < 1$ at a later date $q + \delta q$ because the relative values of its investments are not constant as a function of time. The importance of this mechanism is confirmed by Monte-Carlo simulations: for each fund i, compute the average entropy $S_{i,MC}$ of its portfolio after 3 months by simulating 20 price paths for each asset independently, using their respective volatility computed over the 60 last days. The red curve of Fig. 4.3 shows the effect of natural asset price evolution on perfectly equally-weighted portfolios after three months: the resulting (locally averaged) scaled entropy S_{MC} increases as a function of n, mirroring the local average of S_i in the same figure for $n < n^*$. In other words, the influence of asset price fluctuations on S_i decreases as n_i increases. Thus, the decrease of S for $n > n^*$ is a strong clue that the funds cannot build equally-weighted portfolios anymore.

We checked that funds attempt to make their portfolio more equally-weighted when they update them, or equivalently that the entropy of $p_{i\alpha}$ of the positions kept between $q - 1$ and q is smaller than the entropy of the full portfolio at time q, which implies that new positions are more equally-weighted than old ones. This is fully confirmed in Fig. 4.3: the dash blue line is the local average of the entropy restricted

to the set of common positions between two consecutive snapshots, multiplied by $S_{MC}(n_i)/S_{MC}(n_{i,\text{common positions}})$ in order to account for the dependence of S on n: the entropy of old positions is clearly smaller than the entropy of the new portfolio, hence new positions purposefully bring S_i closer to equally-weighted portfolios.

In summary, funds attempt to spread their investment evenly between the assets they invest in, which means that Eq. (4.2) may hold. Next, the fact that the local average leads to a well-defined exponent, constant over of a substantial range of $n < n^*$ implies that this equation holds, i.e. that WoC is present in this data set.

The value $\mu = 2$ corresponds to effective flat-fee transaction costs but this does not imply that market participants really do pay a flat fee per transaction, only that their population acts as if they do. Quite remarkably, the same exponent was found for private investors and asset managers (with portfolios smaller by several orders of magnitude). Thus the collective behaviour of large professional investment funds is essentially the same one as individual amateur investors. Since Eq. (4.2) holds over many decades of portfolio values for a wide spectrum of market participants, we argue that WoC is a plausible explanation of the average portfolio structure. Finally, we emphasis that our results much extend the validity of WoC, as it may hold for whole functional relationships over many decades of n and W, not only for a single number. This suggests to try to find and understand WoC in new ways and new types of data.

4.3 Asset selection model

So far, bringing to light WoC in the $n < n^*$ region only required to focus on the number of securities in a portfolio, not on how funds select securities. This implicitly assumed that funds could invest in all securities they wished, which is clearly not the case in the large diversification region: the fact that the exponent μ is much smaller in this region implies that funds need on average to split their investments into many more securities in a non-equally weighted way, as shown by scaled entropy. This is most likely due to liquidity constraints: in this region, funds could invest as much as they wish in some assets because there were simply not enough shares to build positions larger than certain sizes without having too large an impact on their prices. Each fund has its own way to determine the maximal amount to invest in a given security α; a common criterion is to limit the amount invested with respect to the capitalization, i.e. $W_{i\alpha}/C_\alpha$. Fig. A.4 in S.I. strongly suggests that each fund fixes its own upper bound

$$f_i^{(\max)} \geq \max_\alpha f_{i\alpha} \quad \text{where } f_{i\alpha} = \frac{W_{i\alpha}}{C_\alpha}. \tag{4.3}$$

It turns out that $f_i^{\max}$ is highly heterogeneous among funds $\log_{10}(f_i^{\max}) \simeq -3.0 \pm 1.0$ (see Fig. A.5), which reflects both the heterogeneous ways of portfolio construction and also the confidence of a fund in its abilities to execute large trades without too much price impact. The existence of such limits implies that portfolios are less likely to be equally-weighted in the large diversification region, as seen indeed in the decrease of the average portfolio weights scaled entropy for $n \geq 70$ (blue line in Fig. 4.3).

Funds, however, do not invest in a randomly chosen security, even in the low diversification region. Figure 4.4 displays a scatter plot of the capitalization C_α of each security α versus m_α, the number of funds which have invested in this security, together with a local non-linear fit. Similarly to W vs n, one finds a new power-law relationship

$$\log C_\alpha = \gamma \log m_\alpha + \epsilon_\alpha \tag{4.4}$$

for large enough m_α. Hence in local average notations, $C \propto m^\gamma$. Exponent γ is stable during the period 2007-2014 (see Fig. 4.2) and its average $\bar{\gamma} \simeq 2.2 \pm 0.1$.

In short, one needs to introduce a model of how funds choose to invest in securities to reproduce the average behaviour of both Eqs (4.4) and (4.1). Since one sees a cross-over between two types of behaviour rather than an abrupt change, one needs to investigate both regions together. Because $P(n_i)$ has an approximate power-law tail, we use logarithmic binning of the axis n_i, which ensures that the expected number of points per bin is approximately constant. We denote the bin number of fund i by $[n_i]$. Two mechanisms must be specified: how a fund selects security α and how much it invests in it. The latter point is dictated by Fig. A.4 in the large n_i region where fund i invests at most $W_{i\alpha} = f_i^{(\max)} C_\alpha$; for the sake of simplicity, we approximate $f_i^{(\max)}$ by the median value of $f_i^{(\max)}$ in the bin $[n_i]$, denoted by $f_{[n_i]}^{(\max)}$. In the small diversification region, we assume that $n_i = n_i^{\text{opt}}$, thus $W_{i\alpha} = W_i / n_i^{\text{opt}}$ to be consistent with our previous results. We choose a security selection mechanism that rests on the market capitalization C_α of a security α (see S.I.) which is a good proxy of the liquidity (Fig. A.6). We perform Monte-Carlo simulations from the empirical selection probabilities and $f_{[n_i]}^{(\max)}$ and display the resulting W vs n and C vs m in Figs 4.1 and 4.4 (continuous green lines), in good agreement with the local averages (continuous orange lines). One notices a discrepancy in the relationship C vs m for large n, which mainly comes from funds in the large diversification region. (See Fig. A.7 S.I).

The large diversification region illustrates how constraints may considerably modify the rational benchmark. While the above mechanism of security selection is able to reproduce adequately the behaviour of well diversified funds, we could not find a rational benchmark for the dependence of $f_i^{\max}$ and n_i. Thus, the case for WoC in the large diversification region is not entirely closed.

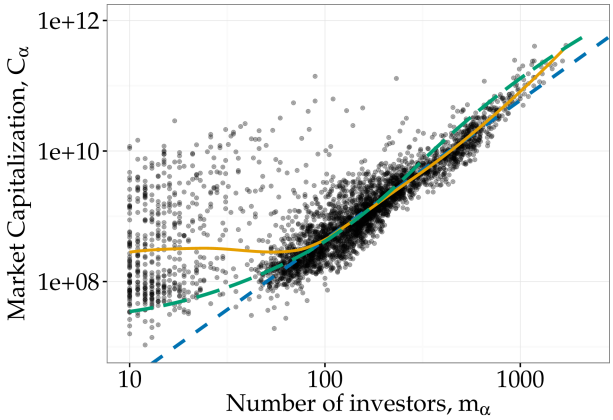

Figure 4.4: Market capitalization of securities as a function of the number of investors in logarithmic scale. From the local non-linear robust fit (orange line) we observe a linear relationship for assets with more than about 100 investors. The blue dashed line corresponds to a linear fit on that group of asset, which yields $W_\alpha \propto m_\alpha^\gamma$, with $\gamma \simeq 2.1$.
Green curve: local average from the simulated asset selection model.

Data

Our dataset consists of an aggregation of the following publicly available reports (in order of reliability): the SEC Form 13F, the SEC's EDGAR system forms N-Q and N-CSR and (occasionally) the form 485BPOS. Our work focuses on the period starting from the first quarter of 2005 to the last quarter of 2013.

These forms contain of some mistakes. We partially fix them by cross-checking different sources (which often contains overlapping information) and by filtering data before processing (see details in S.I.).

The main limitation of this dataset is that it provides accurate figures for long positions only. The other positions (short, bonds, ...) are most of the time only partially known. The frequency of the dataset is also inhomogeneous: data for most of the funds are quarterly updated (depending on regulations), hence we decided to restrict ourselves to 4 points in a year only. Such frequency is probably too low for investigating the dynamics of individual behaviour but is not a problem for we focus on an aggregate and static representation of the investment structure.

Discussion and conclusion

While Wisdom of the Crowd is commonly applied to a population collectively guessing a single number, we provide evidence that it holds much more generally when a population tries to guess an optimal answer imperfectly. We thus argue that one should look for WoC in potentially much more complex situations.

In economic systems, the reference function is dictated by individual optimality arguments. We have focused here on financial market participants, who satisfy two of the conditions understood to underlie the existence of WoC: diversity of opinions, and better than random behaviour (Hong and Page, 2004). Imitation, known to be detrimental to WoC (Lorenz et al., 2011; Muchnik et al., 2013), depends on publicly available data. For example asset price bubbles arise because of positive feedback on past prices (Lux, 1995). The fact that SEC data studied here is publicly available opens the way to imitation as well. This is however unlikely to have a sizeable effect in this study. First, while imitation is possible, it would show up in data with much more frequent (e.g. daily) updates. Second, large funds know that their positions may be copied just after publication (which does not occur immediately after filing). Thus, they may anticipate short-term imitation and may adjust their portfolios just before reporting dates accordingly. In short, imitation is likely to have a small effect in the data we study.

At a higher level, our results suggest that, while individuals may deviate much from the rational expectation theory, standard economic theory may hold at a collective level without need for micro-founded optimal individual decisions: the average decision may in some cases approximate that of a rational, representative agent. Our results however only hold on a snapshot of the system, for which individual fluctuations may be averaged out. In a dynamic setting, the very large deviations from the rational benchmark may not be neglected in the presence of feedback loops (Gualdi et al., 2015). In other words, the dynamics of these fluctuations are worth investigating in their own right.

Chapter 5

Large large-trader activity weakens long memory of limit order markets

5.1 Introduction

Financial market dynamics is complex in part because of the very large variety of timescales at play. Both traders and volatility feedback loops are known to have widely distributed timescales (Lynch and Zumbach, 2003; Lillo, 2007; Zhou et al., 2011; Tumminello et al., 2012; Challet et al., 2016). Accordingly, investigating how timescales interact reveals some of the fundamental dynamical ingredients of price dynamics. For example, the asymmetric relationship between historical and realized volatility shows that price dynamics is not symmetric with respect to time reversal (Lynch and Zumbach, 2003; Zumbach, 2009), which imposes a strong constraint on realistic stochastic volatility models (Blanc et al., 2017).

The long memory of the signs of market orders is a well-established stylized fact of limit order books (Bouchaud et al., 2004; Lillo and Farmer, 2004) that passes the most stringent statistical tests. Lillo et al. (2005) propose a mathematical framework that links the long memory of these signs to the way very large orders are split into a series of smaller market orders (thereby creating a meta-order) and is able to reproduce the empirical auto-correlation function if the distribution of the meta-order size has a Pareto-like tail. In other words, the shape of the sign auto-correlation function reflects that of the distribution of the size of meta-orders.

Here, we use two large databases of almost maximally different timescales, namely quarterly filings by large investment funds and a comprehensive tick-by-tick database, which allow us to investigate the influence of large funds on the memory properties of the limit order book. We first show that the memory length of market order signs

(buy/sell) of a given asset is markedly weaker when a large fraction of its capitalization is exchanged by large funds over a quarter. Reciprocally, we test if assets with the weakest market order sign memory are likely to being much traded by large funds. Finally, we use the theoretical framework of Lillo et al. (2005) to put forward a coherent picture of our findings.

5.2 The data

Our dataset consists of two databases: quarterly snapshots of large investment ownership, from the corresponding FactSet database in the 2007-2013 period (32 reports), which contains data 10845 funds. We filter out funds with less than USD 100'000 invested into securities. The remaining funds are invested in 12531 securities. We focus on the 2480 assets continuously recorded in FactSet database after their first quarter of appearance and with at least one full year of record. Using automated methods, we link assets found in both FactSet and the Thomson Reuters Tick History databases. The latter provides an event-by-event history of limit order books. For each asset traded on the NASDAQ and each day, we extracted all the trade prices together with the best bid and ask prices just before the trades. Finally, we keep assets traded for at least 200 days and with more than 200 trades per day on average. This leaves 846 stocks and more than 6.7 billion trades. Details about the procedure is provided in Chapter 2.

5.3 Methods

In order to link trade-by-trade data with quarterly fund filings, we define suitable quantities in each dataset and investigate how they are related. For each asset α, we compute the mid-price just before the n-th trade, denoted by $m_{\alpha,n}$ as the average between the best bid and ask prices. Then we define the sign of the n-th market order of asset α as

$$\epsilon_{\alpha,n} = \text{sign}(p_{\alpha,n} - m_{\alpha,n}).$$

We drop trades that occur exactly at the previous mid-price. As suggested by the above notation, we define the time as the number of market orders since the beginning of the time-series.

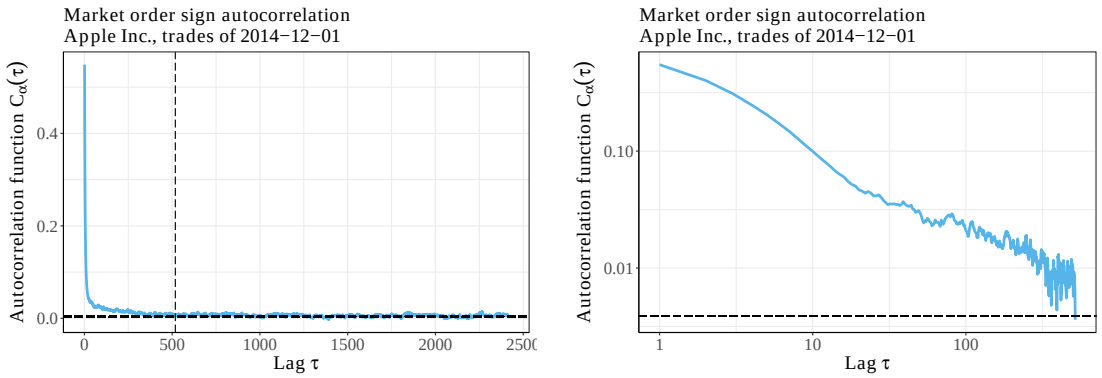

Figure 5.1: Example of market order sign autocorrelation functions with linear axis scales (left plot) and log-log axis scales (right plot). The black dashed horizontal line is the noise level $2/\sqrt{N}$ where N is the length of the market order sign time series. The black dashed vertical line corresponds to the maximum lag τ^*. Apple Inc., trades of 2014-12-01.

5.3.1 Microstructure: memory length of market order sign auto-correlation

We define several simple ways to characterize the memory of market order signs. Unless specified otherwise, all market microstructure quantities are measured over a full trading week. The first one consists in measuring the probability of the occurrence of κ consecutive trades of a given sign in a random contiguous subset of $\{s_n\}$. Mathematically, for a generic κ, this amounts to measuring the conditional event frequency

$$\pi_\alpha^{(s\kappa)} = P(s_n = s_{n+1} = \cdots = s_{n+\kappa} = s) \tag{5.1}$$

for both $s \in \{-1, 1\}$.

Another way to characterize the memory of order signs is the market order sign autocorrelation at lag τ (in unit of market orders), denoted by $C_\alpha(\tau)$. Process ϵ_n^α has a long memory if the integral of its autocorrelation function C_α diverges. Many references find that $C_\alpha(\tau) \propto a\tau^{-b}$ where $b < 1$, in which case the integral of $C_\alpha(\tau)$ is infinite (see e.g. Lillo et al. (2005); Toth et al. (2015)) (we omit the α index for a and b in order to avoid too heavy notations). This is indeed a good approximation for very long time series. For finite time series of length N, one can define the effective memory length as the lag τ_α^* after which C_α reaches for the first time the noise level of autocorrelation functions $2/\sqrt{N}$, i.e., τ_α^* is such that $C_\alpha(\tau) > 2/\sqrt{N}\ \ \forall \tau \leq \tau_\alpha^*$. We will also consider the scaled maximum lag τ_α^*/N.

5.3.2 Macro-dynamics: directional fund activity ratio

First, we introduce a quantity that measures, for a given asset α, the rescaled global directional change of ownership averaged over all the funds between the quarter ends $q-1$ and q. We will call it the directional fund activity ratio and define it as

$$r_\alpha(q) = \frac{\sum_i [W_{i\alpha}(q) - W_{i\alpha}(q-1)]}{V_\alpha(q)}, \tag{5.2}$$

where $W_{i\alpha}(q)$ is the position in dollars of fund i on security α at the end of quarter q, and V_α the total volume-dollar of security α exchanged between $q-1$ and q. If $r_\alpha(t) > 0$ (resp. $r_\alpha(t) < 0$) then the security α is more bought (resp. sold) than sold (resp. bought) by the large funds in our database. We will focus on $R_\alpha(q) = |r_\alpha(q)|$.

5.3.3 Macro-dynamics: absolute fund activity ratio

Another important measure of aggregate fund behaviour consists in quantifying how much the investment has changed in absolute terms. We thus define $S_\alpha(q)$ as the rescaled absolute difference of invested amounts between quarter ends $q-1$ and q, i.e.,

$$S_\alpha(q) = \frac{\sum_i |W_{i\alpha}(q) - W_{i\alpha}(q-1)|}{V_\alpha(q)}. \tag{5.3}$$

This quantity cannot account for round trips of funds over a quarter, which are fortunately very unlikely for the largest values of S , i.e., the values relevant to the present work. In addition, when the relative influence on S of large fund round-trips is negligible, S is a good approximation of large fund participation ratio.

5.4 Results

5.4.1 From large fund behaviour to microstructure dynamics

The premise of this paper is that relating tick-by-tick order book properties to the fund ownership database is easiest when the aggregate behaviour of large investment funds is the most extreme, which corresponds to large values of either R_α or S_α. Thus

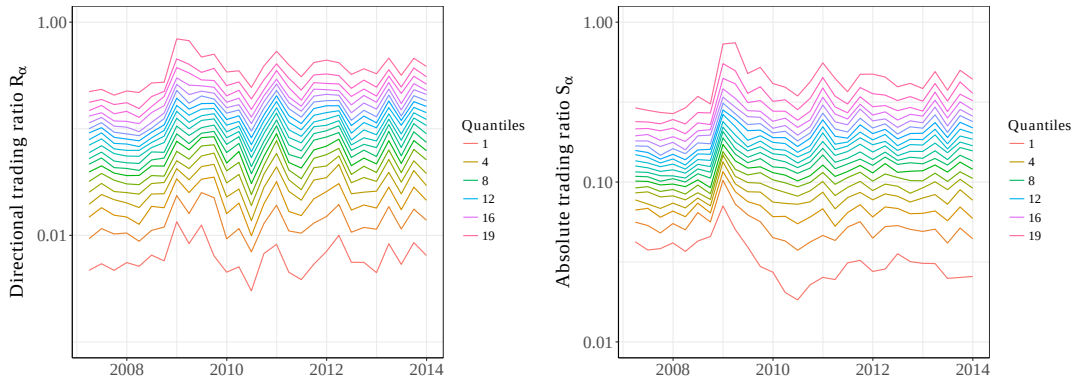

Figure 5.2: Time evolution of the quantiles of the directional fund activity ratio $R_\alpha(q)$ (left plot) and of the absolute fund activity ratio $S_\alpha(q)$ (right plot).

for each quarter q, we divide the assets into 20 groups of $R_\alpha(q)$ by computing the quantiles $k_R \in \{1, \cdots, 19\}$; we do the same for $S_\alpha(q)$, yielding $k_S \in \{1, \cdots, 19\}$. We first compare the microstructural dynamics of the top and bottom groups of both quantities. By convention, the bottom groups $g_X = 1$ correspond to small values of $X \in \{R, S\}$, i.e., to securities that are bought and sold equally (R) or not much traded by large funds (S). Figure 5.2 shows the time evolution of the quantiles of the ratios $R_\alpha(q)$ and $S_\alpha(q)$. The large-R quantile are clearly correlated with the large-S quantiles. This should be expected, as a large R_α implies a large S_α.

Focusing on the assets belonging to the top and bottom groups determined by the quantiles of R and S, one now assess the influence of trading by large funds on the market order sign memory length measures. Starting with the two extreme groups determined by the quantiles of R, Fig. 5.3 reports that the frequency of κ consecutive trades of the same sign $\pi_\alpha^{(s\kappa)}$ for $\kappa = 2$, once averaged over all assets belonging to a given group, is consistently different between the two groups of assets as time goes on; one also sees that the difference is larger for $\kappa = 10$ than for $\kappa = 2$; in fact, it is an increasing function of κ, at least for $2 \leq \kappa \leq 10$. The difference is larger when the assets are grouped according to the quantiles of S, as illustrated in Fig.5.4. In short, being actively traded by large funds decreases the probability of occurrence of consecutive market orders of the same kind, which thus is a sign of weakening of market order sign memory.

The other measures of memory length lead to the same conclusion. For example, fitting the trade sign autocorrelation $C_\alpha(\tau)$ with $a\tau^{-b}$ for α in the top and bottom groups of assets consistently yields smaller values of the prefactor a for the quantiles of either R or S (top plots of Fig. 5.5), except in 2008-2009 with respect to the quantiles of R: during this period, a was roughly the same in both groups. The similar behaviour of $\pi_\alpha^{(s\kappa)}$ and a is to be expected: $C_\alpha(\tau)$ being a function of $\{\pi_\alpha^{(s\kappa)}\}_\kappa$, the prefactor a

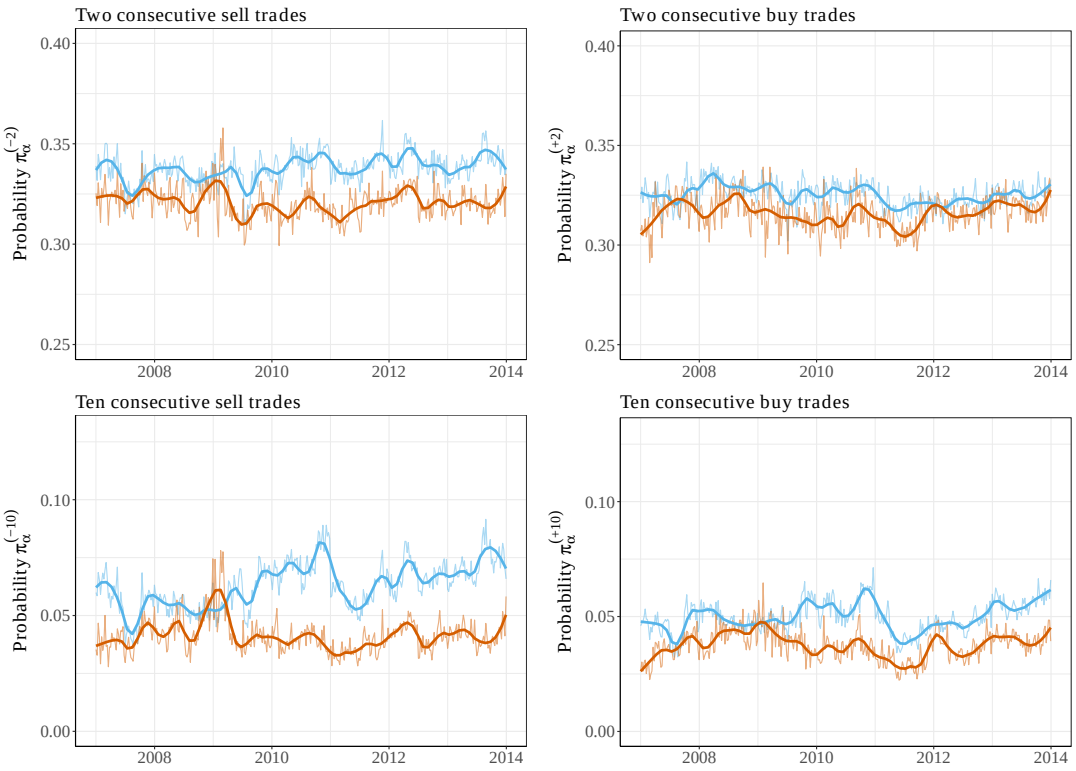

Figure 5.3: Time evolution of $\pi_\alpha^{(\kappa)}$, the probability of observing κ consecutive negative trade signs (left plots) and κ consecutive positive trade signs (right plots) for the top and bottom quantiles of $R_\alpha(q)$ (orange and blue lines, respectively).

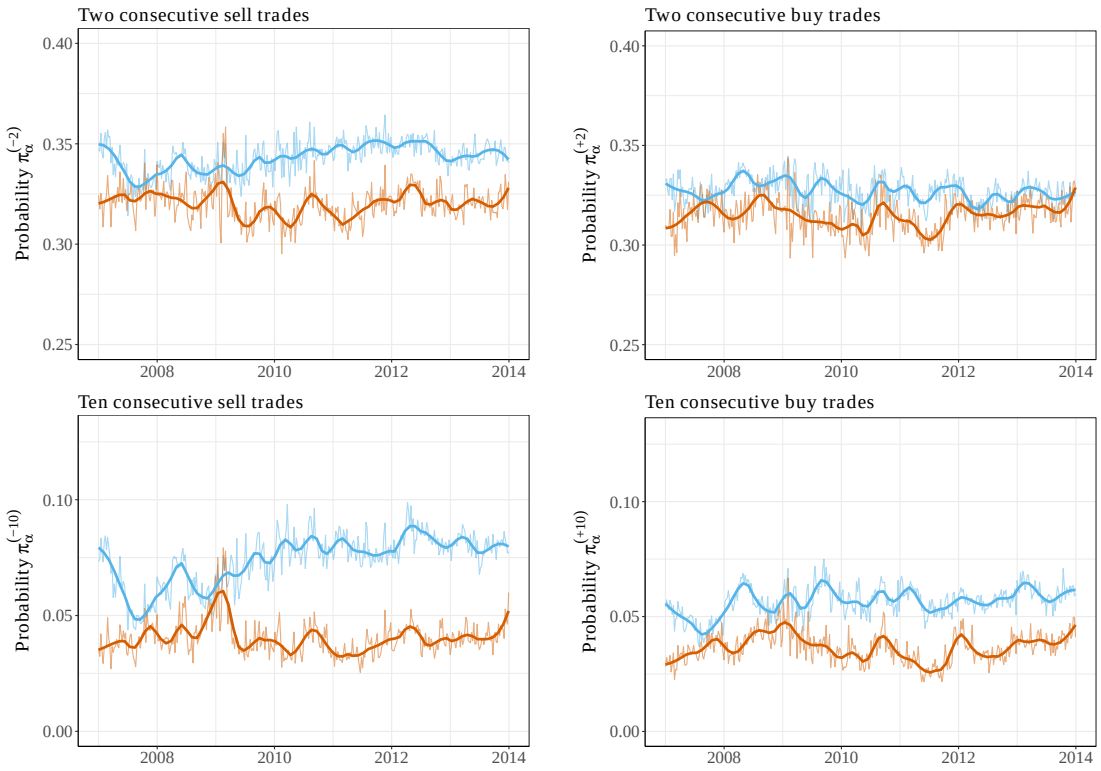

Figure 5.4: Time evolution of $\pi_\alpha^{(\kappa)}$, the probability of observing κ consecutive negative trade signs (left plots) and κ consecutive positive trade signs (right plots) for the top and bottom quantiles of $S_\alpha(q)$ (orange and blue lines, respectively).

is mostly related to small-κ probabilities $\pi_\alpha^{(s\kappa)}$. The case of the exponent b is more nuanced and revealingly so (bottom-left plot of Fig. 5.5): large directional trading by large funds has no clear influence on b, except in times of crisis, as e.g. in 2008-2009 when assets with large R a smaller b than the assets in the small-R group. The 2011 crisis also lead to a significant and similar influence of R on b; while the typical value of a of assets with a large R plunged, it did not reach that of assets with small R, contrarily to what happened during the 2008-2009 period. The absolute activity ratio S has always been discriminant for a. Regarding b, assets with a large S also had a smaller b (but a large a) in 2008-2009, while in the 2012-2014 period, the reverse is true. Thus these fitting parameters provide a more dynamic picture on the influence of the activity of large funds.

The overall picture is nevertheless the same: for a given number of trades, on average, the difference of a and b between the top and bottom quantiles of R and S contribute to shorten the length of the memory as inferred by $C_\alpha(\tau)$ because the noise level is reached at smaller lags for assets in the top quantiles. This is indeed confirmed by Fig. 5.6 which shows the time evolution of the τ_α^* averaged over the top and bottom quantiles of R and S. One notes that τ_α^* of the top and bottom groups are clearly separated, while this ceases to be the case for and τ_α^*/N since 2012. The effect of R or S is opposite on τ_α^* and τ_α^*/N, which is due to the fact that the number of transactions N of assets with large R or S is typically smaller.

5.4.2 Large fund directional and absolute trading detection

A more relevant question in practice is whether one can detect trading by large investment funds from quantities measured from tick-by-tick data. In the context of this paper, the question may be rephrased as how to guess in which quantile of R or S a given asset may be from the knowledge of a, b, τ^* or τ^*/N. Since the later quantities are measured of a week, we compute with their averages over a given quarter.

Here, we focus on the following simple classification problem: for each quarter, we split the 20 groups into two categories according to quantile k_{cut}: assets belonging to the quantiles $k \leq k_{\text{cut}}$ form the first category and the remaining ones the other one. Choosing a memory length measure as the variable according to which one classifies the assets during a given quarter, it is then straightforward to compute the parametric Receiver Operating Characteristic (ROC) curve and its associated area under curve (AUC) for a given quarter. Figure 5.8 reports the AUC associated with $\pi^{(s10)}$, a, b, τ^* and τ^*/N for the quantiles of both R and S, averaged over the 32 quarters. One sees that $\pi^{(s10)}$ and a are roughly equivalent and are the best variables to discriminate the large values of R and S. Even more, their detection power with respect to the quantiles of S does not depend much on k_{cut}.

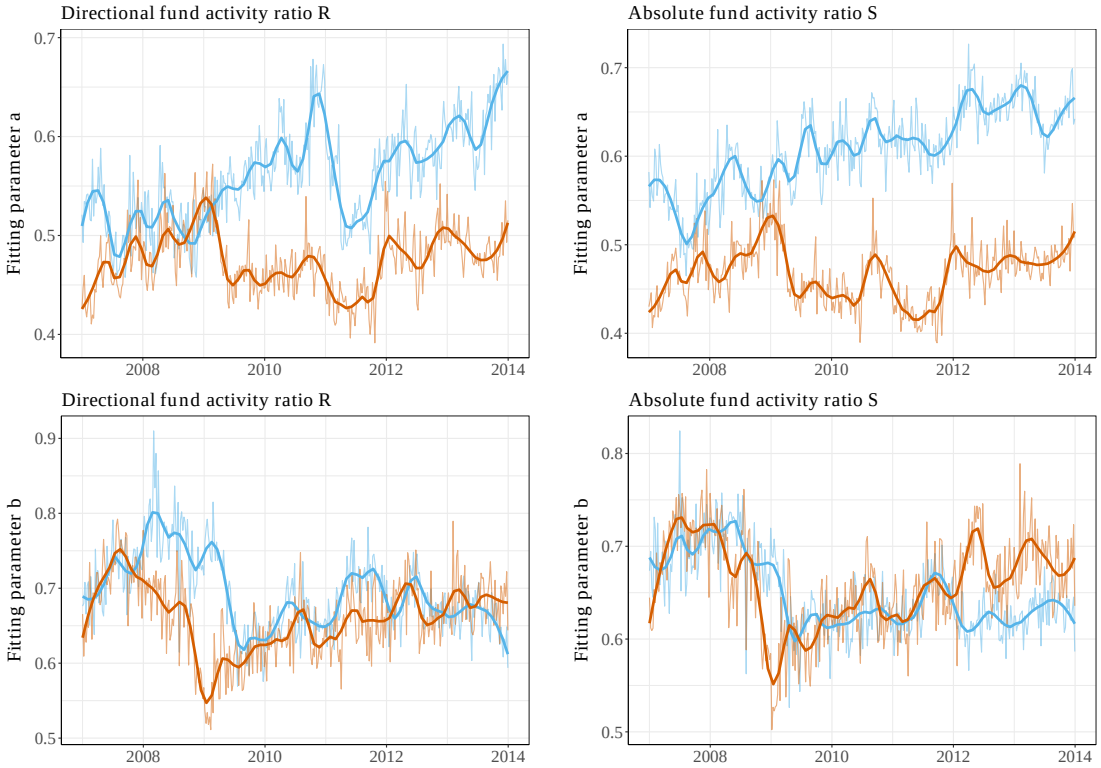

Figure 5.5: Time evolution of the fitting parameters characterizing the long memory of the market order sign autocorrelation $C_\alpha(\tau) = a\tau^b$. Top plots: evolution of a averaged over all the members of the bottom and top quantile of $R_\alpha(q)$ (left plot) and $S_\alpha(q)$ (right plot). Bottom plots: evolution of b averaged over all the members of the bottom and top quantiles (orange and blue lines, respectively) of $R_\alpha(q)$ (left plot) and $S_\alpha(q)$ (right plot).

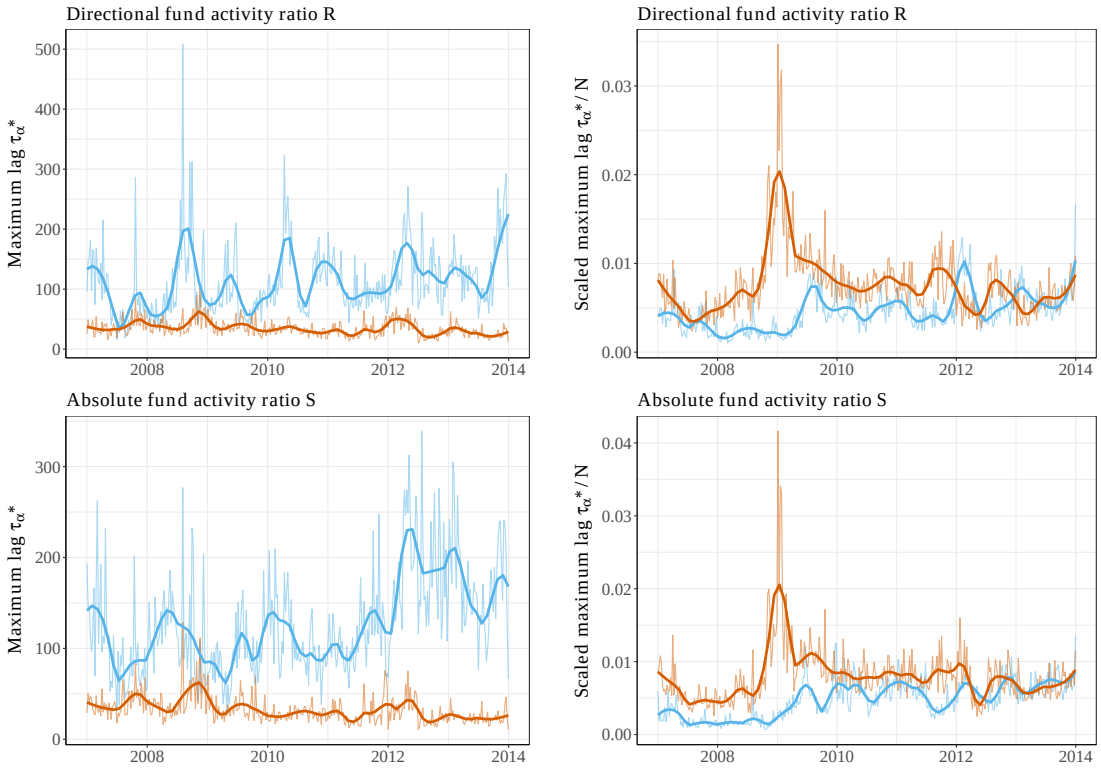

Figure 5.6: Time evolution of τ_α^*, the scaled maximum lag before the autocorrelation function of the trade signs reaches the noise level, averaged over the top and bottom quantiles (orange and blue lines, respectively) of $R_\alpha(q)$ (upper plots) and $S_\alpha(q)$ (bottom plots).

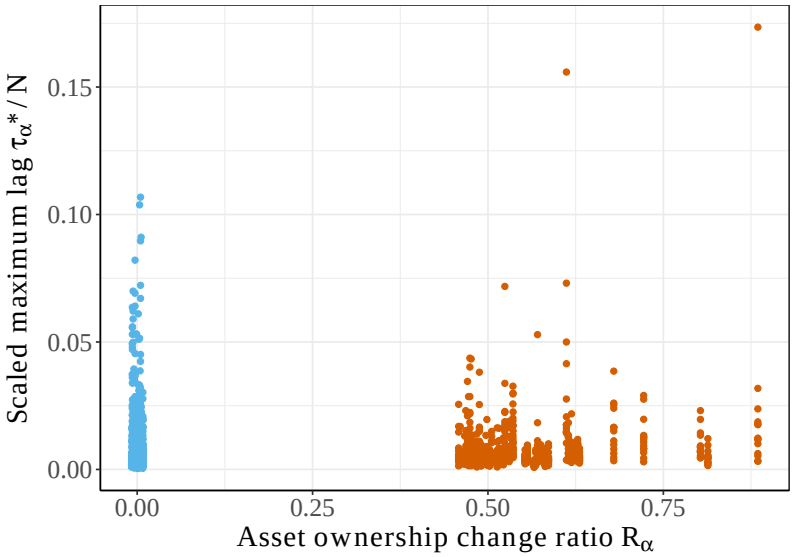

Figure 5.7: Memory length of stocks in $R_\alpha^1(q)$ (blue dots) and in $R_\alpha^{20}(q)$ (orange dots) at quarter q=26 (2013-09-30). One point corresponds to the memory length computed for one trading week for a given stock. Points with the same R_α correspond to the same stock (One quarter is about 12 trading weeks).

5.5 A theoretical approach

Lillo et al. (2005) establish the link between the size of meta-orders and the order sign auto-correlation. In particular, they find that if the distribution of the size L of meta-orders, denoted by $P(L)$, has a power-law tail $P(L) \propto L^{-(\beta+1)}$, then, assuming that the sign order of each meta-order is equiprobably -1 or $+1$ and that there are exactly M active meta-orders at any time for a given asset, the auto-correlation function is given by $C(\tau) \simeq a\tau^{-b} = \frac{M^{\beta-2}}{\beta}\tau^{-(\beta-1)}$ for large τ. This simple relationship gives two insights relevant to our results.

First, the pre-factor a is a decreasing function of M when $\beta < 2$, which is generally the case as $\beta \simeq 1.5$ on average for example in the London Stock Exchange (Lillo et al., 2005). Identifying $\beta - 1$ with b here shows that in our case $b < 1$, hence that $\beta < 2$. This implies that the pre-factor a is a proxy for the number of meta-orders present in the market, *ceteribus paribus*, which should then be strongly related to S. This is why the pre-factor a (and thus $\pi^{(10)}$) are among the best predictors of the quantile range of S (Fig. 5.5).

Second, b being a proxy for an effective β, it allows to gain some insight on the tail of $P(L)$. Since the measured b decreases, the probability of very large meta-orders

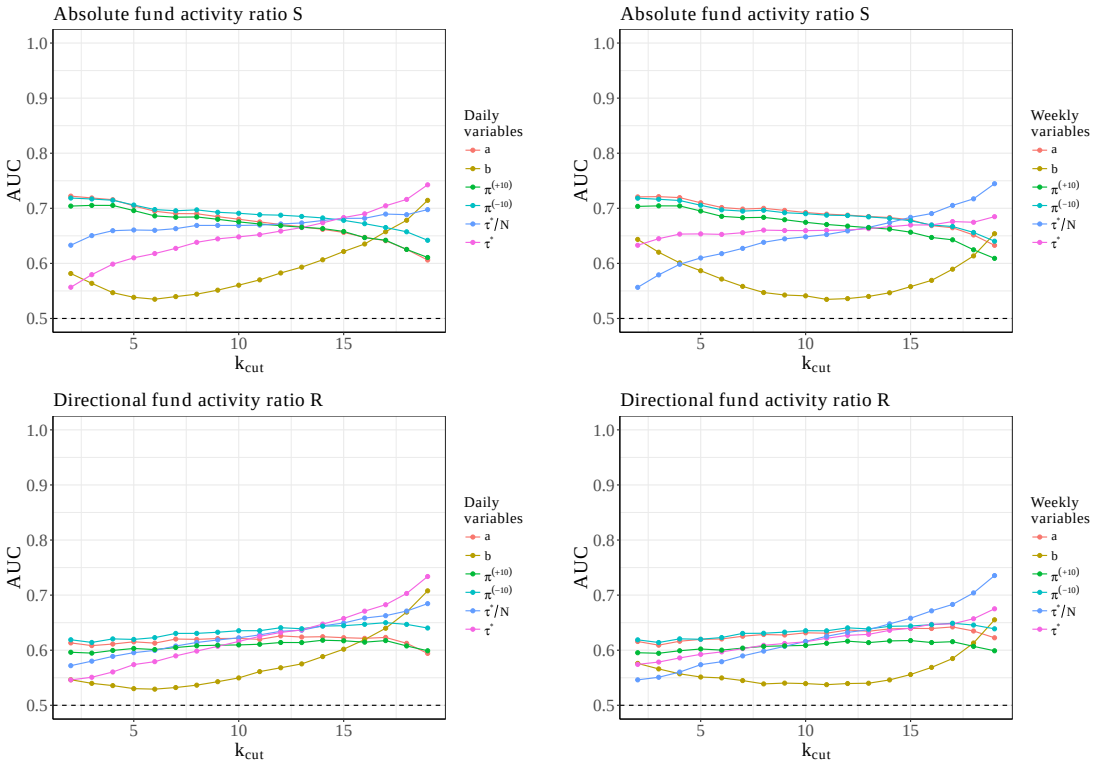

Figure 5.8: Average quarterly Area Under Curve (AUC) corresponding to the classification of funds belonging to quantile $k \geq k_{\text{cut}}$ for each market order sign memory length measures. Left plot: averages of daily microstructural quantities; right plot: averages of weekly microstructural quantities. AUC averaged over the whole history.

increases, which is coherent with directional trading from large funds after large price changes. Another explanation for the change of b resides in the the way meta-orders are split: there must be feedback loops between the exponent β and the available liquidity which in turns most probably depends on the current order flow, determined in part by meta-orders.

5.6 Concluding remarks

This work has brought to light the fact that the influence of the trading of large funds on the memory of market order signs is far from negligible: large funds do not suppress long memory, but may weaken it. When one knows *ex postfacto* how large funds have traded, there is a clear difference between limit order book dynamics in which large fund took a substantial part and those barely touched by them (relatively speaking). Reversely, even when averaging the market order sign memory length measures over a quarter allows to some extend to predict if large funds have been much involved in the trading of a given asset.

The main limitation of the present work is the use of quarterly data to characterise fund behaviour. Using labelled trades would open the way to relate the properties of order book day by day and and to improve our understanding of the link between the composition of meta-orders and the memory length of market order signs.

Conclusion and outlooks

In this conclusion I want to give a general overview of the topics studied in this thesis in light of recent developments in the literature, and I will suggest different directions for possible extension.

In chapter 4 we propose that the average portfolio structure of institutional investors reproduces the structure which is optimally accounting for transaction costs. Chapter 4 also discusses that, while individuals may deviate much from the rational expectation theory, standard economic theory may hold at a collective level without need for micro-founded optimal individual decisions. If similar results were to be found on portfolio selection then symmetric information properties would emerge from an aggregation of private, diverse and asymmetric information of a crowd of informed and noise traders.

We also reminded that, for Wisdom of the Crowds to be valid, the effect of imitation has to be limited, which can be in contradiction with the fact that the SEC data, for example, is publicly available. Therefore it appears critical to estimate the degree of imitation that exists in financial markets. In Chapter 3 we quantified portfolio overlaps between institutional investors, a temporal analysis of portfolio overlaps lead-lag relationships could be used as a proxy for imitation and allow to quantify in what extent institutional investors imitate their (more informed?) peers after publication. Although that effect may be altered by the fact that large funds know that their positions may be copied just after publication (which does not occur immediately after filing). Thus, they may anticipate short-term imitation and may adjust their portfolios just before reporting dates accordingly.

In chapter 5 we studied the influence that large-traders have on the market dynamics. We found that large directional traders shortened the memory length of order signs however we did not explore the origin of these behaviours. Were large trades triggered by the arrival of new information? Friedman (1953) and Fama (1965) argued that

irrational investors are met in the market by rational arbitrageurs who trade against them, driving the prices close to a fundamental values. Identifying informed traders from noise traders could help a better understanding of price formation.

Even though the scope of this thesis was limited to large-traders (institutional investors and large funds) behaviour and limit order book dynamic, we developed tools (in Chapter 2) that allows to tackle most of the technical challenges one can encounter when working with data from diverse sources and nature. These tools can be used for future development using more macro-economic related datasets.

From a macro-economic perspective, the role of financial markets is to efficiently allocate capital among competing projects by allowing market participants to meet and exchange securities, and, consequently, to enable risk sharing for market participants. Adam Smith believed in self regulated free-market force, giving everyone freedom to produce and exchange goods, that became known as the "invisible hand" (Smith, 1759, 1776).

Firms are required to publicly report their accounting information (e.g. income statement and balance sheet). Not only public information release has been shown to improve the risk sharing mechanism described by Adam Smith (Diamond, 1985), it has also lead to a lot of work, referred to as capital markets research. For example, the accounting measurement theory seminal work by Ball and Brown (1968), which documents the relationship between prices and accounting numbers using traditional fundamental analysis, involves the determination of the value of securities by examination of key value-drivers (such as earnings, risk, growth and competitive position) reported by firms.

The demand of capital markets research in accounting is strong, numbers of applications explain its popularity. Fundamental analysis and valuation have been used to identify mispriced securities (Frankel and Lee, 1998), short-sellers have been shown to position themselves in stocks with low fundamental-to-price ratios (Dechow et al., 2001), Lev and Thiagarajan (1993) validated that some investors use the fundamentals to assess the extent of earnings persistence and growth, future earnings prediction from fundamental signals (Abarbanell and Bushee, 1997; Penman, 1992), and an indirect approach which is based on an examination of the relations between fundamental signals and stock returns (Ou and Penman, 1989; Greig, 1992).

In this thesis we considered firms as passive agents in the funds portfolio selection process. However firms actively influence this process and evidences show financial reporting to be important means for management to communicate firm performance to outside investors (Healy and Palepu, 2001). The consequences are important: managers have been found to manipulate real activities (e.g. price discounts, overproduction, reduction of discretionary expenditures) and sacrifice economic value, as avoid

initiating a very positive net present value (NPV) project if it meant falling short of the current quarter's consensus earnings, in order to avoid reporting losses and, even though these activities enable managers to meet short-run earnings targets, they are unlikely to increase long-run firm value (Graham et al., 2005; Roychowdhury, 2006).

These facts raise numbers of still open questions. It is instructive to enumerate some of them:

- do some investors imitate peers using ownership public disclosure?
- do some investors manipulate their portfolio before publication?
- how do investors respond to corporate disclosures?
- how does disclosure affect resource allocation in the economy?
- how much to pay for a stock and what is the role of accounting in that assessment? In other words, does accounting information allow to predict long term stock prices?

Answering these questions requires a global approach and to the economy as a whole. A tri-partite study of capital markets, large-traders' portfolio and firms' accounting information would open the way to a better understanding of the mechanisms at play.

Appendix A

Wisdom of the institutional crowd - Supporting information

A.1 Determination of the crossover point n^*

For each date t, we define the cross-over point n^* between the two regions which appear in the local polynomial regression. We determine this point value with a likelihood maximization of the model

$$\log W = \mu_< \log n + (\mu_> - \mu_<)(\log n - \log n^*)\theta(\log n - \log n^*), \tag{A.1}$$

where $\theta(x)$ is the Heaviside function which encodes an if clause so that $\mu = \mu_<$ if $n < n^*$ and $\mu = \mu_>$ otherwise. We use the method introduced by Muggeo (2003) to find parameters $\mu_<$, $\mu_>$ and n^*, which is implemented by its author in the `R` package `segmented`. In essence, this method consists in linearising Eq. (A.1) and determining the relevant parameters recursively. Figure 4.2 shows that n^* is stable as a function of time.

A.2 Asset selection: a model

The framework we introduce in this paper follows a series of a few elementary steps described below. The aim is for the model to be sensitive to the different constraints which dominates the portfolio selection of a fund.

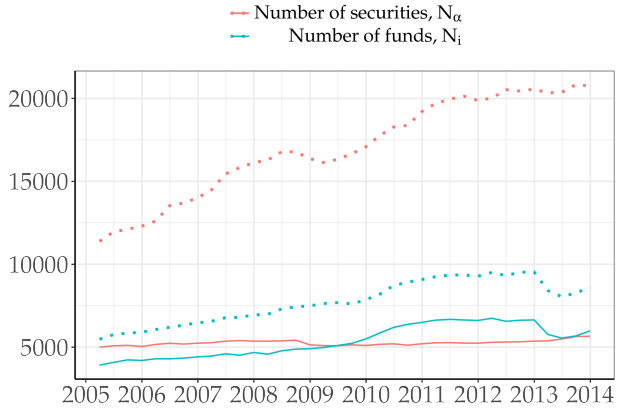

Figure A.1: Temporal evolution of the number of funds N_i and securities N_α in the database. Unfiltered in dashed lines and US based only in solid lines.

A.2.1 Asset selection in the small diversification region $n_i < n^*$

In this region, we assume that portfolios are equally weighted. Each position has a size $\frac{W_i}{n_i^{\text{opt}}}$, where n_i^{opt} is the optimal number of position computed with Eq. (4.1). The funds select their asset randomly with a probability proportional to C_α. Also, in order to build an equally-weighted portfolio, a position is valid only if it is of size $\frac{W_i}{n_i^{\text{opt}}}$.

A.2.2 Asset selection in the large diversification region $n_i \geq n^*$

In this region, the liquidity constraints make it harder for funds to keep an equally-weighted portfolio and portfolio values are thus spread on a larger number of assets. We propose here a stochastic model of asset selection based on two main ingredients: first that the selection probability of asset α by fund i depends on the diversification of a fund n_i and on the scaled rank of the capitalization of asset α, and that the investment is bounded by an hard constraint on the fraction of market capitalization of asset α.

We chose a security selection mechanism which rests on the scaled rank of capitalization of security α, defined as $\rho_\alpha = \frac{r_\alpha}{M}$ where r_α is the rank of capitalization C_α and M the number of securities at a given time. The selection probability $P(W_{i\alpha} > 0|\rho_\alpha)$ is then obtained by parametric fit to a beta distribution in each logarithmic bin. Note that we do not use the same rank-based selection mechanism in the low-diversification region because in this case it is harder to have a good fit with the beta distribution. This is however only a minor point since the capitalization is approximately power-law distributed and the two selection mechanisms are basically equivalent (the rank

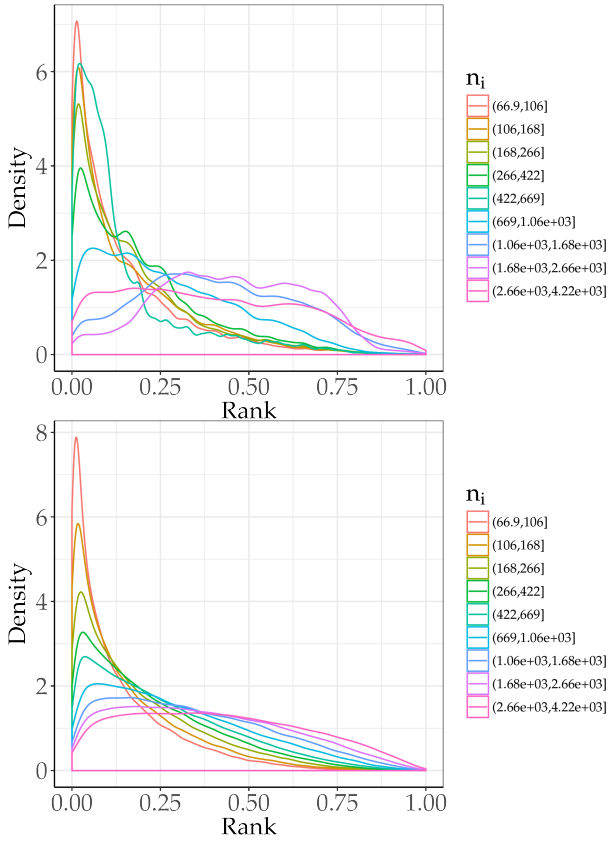

Figure A.2: Top: Empirical probability of investing in a security of scaled capitalization rank ρ for each fund diversification bin $[n_i]$. Bottom: Probability density function of investing in a security of scaled capitalization rank ρ given the diversification n_i of the fund, given by the model.

is proportional to a power of the capitalization) and indeed results are very similar in both cases.

Figure A.2 shows that the distribution of the ranks in which a fund is invested is sensitive to its diversification n_i (t =2013-03-31). The Beta distribution, defined as defined as

$$f(x;a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \tag{A.2}$$

where a and b are the shape parameters of the distribution and B is a normalization constant, is limited to a $[0,1]$ interval and is flexible enough to describe the asset selection mechanism of funds.

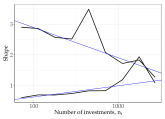

Figure A.3: Coefficients a and b of the Beta Distribution A.2 as a function of n_i . Linear fits are for eye-guidance only.

A.2.2.1 Maximum investment ratio

Funds limit their investment in a given asset by using a simple rule of thumb: defining the investment ratio $f_{i,\alpha} = \frac{W_{i\alpha}}{C_\alpha}$, one easily sees in Fig. A.4 that each fund has its iwn maximum investment ratio

$$f_i^{\max} = \max_\alpha \left(\frac{W_{i\alpha}}{C_\alpha} \right). \tag{A.3}$$

Since the average exchanged dollar-volume of an asset is proportional to its capitalization (Fig. A.6), the existence of $f_i^{\max}$ is a way to account for the available liquidity.

Although that limit is clear for an individual fund, the range of empirical values $f_i^{\max}$ is remarkably large (see Fig. A.5).

A.3 Simulation of asset selection

The simulation is done in a few simple steps:

1. For a given time t, compute n^* from the data using the segmented model Eq. (A.1).
2. Iterate over all the funds: for fund i, with a number of assets n_i,
   (a) If $n_i < n^*$:
      i. Compute its optimal portfolio value using Eq. (4.1). The fund will invest $\frac{W_i^{\text{opt}}}{n_i}$ for every position.
      ii. Select assets randomly with a probability proportional to C_α.
   (b) Else if $n_i \geq n^*$:
      i. Compute its $f_i^{\max}$, so that the fund i will invest $f_i^{\max}$ in n_i assets.
      ii. Select assets according to their capitalization rank following a Beta probability distribution Fig. A.2 with the parameters found in Fig. A.3.[1]

[1] We could use here the empirical probability of asset selection according to their capitalization rank. However, using a parametric distribution with only a few number of parameters reduces the number of degrees of freedom and thus makes the model more generic.

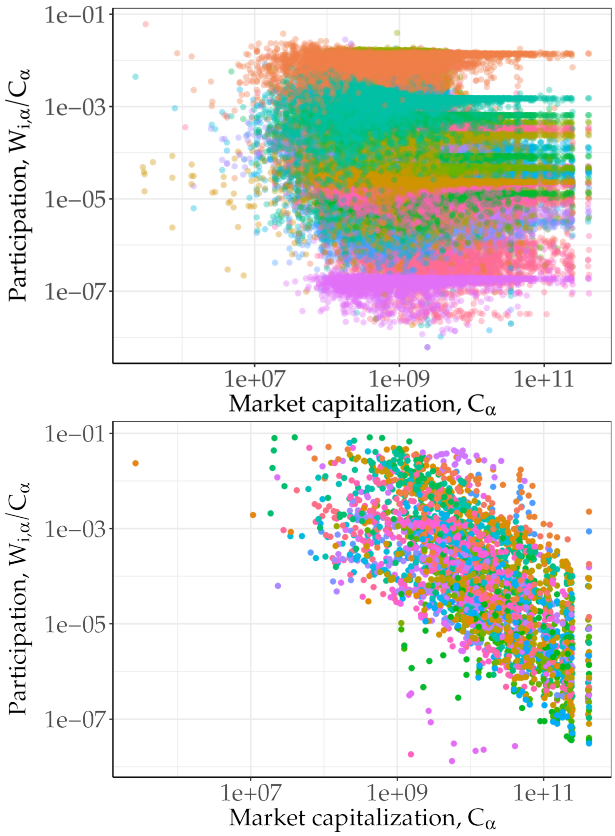

Figure A.4: Fraction of the market capitalization of a security held by a fund. Each color represent a different fund. Top: Funds with a large diversification ($n_i > 800$). We can clearly see a delimitation for most of the funds, which correspond to the maximum fraction $f_i^{\max}$. The value of $f_i^{\max}$ widely differs from one fund to another. Bottom: Funds with a low diversification (n_i<60), $f_i^{\max}$ doesn't appear.

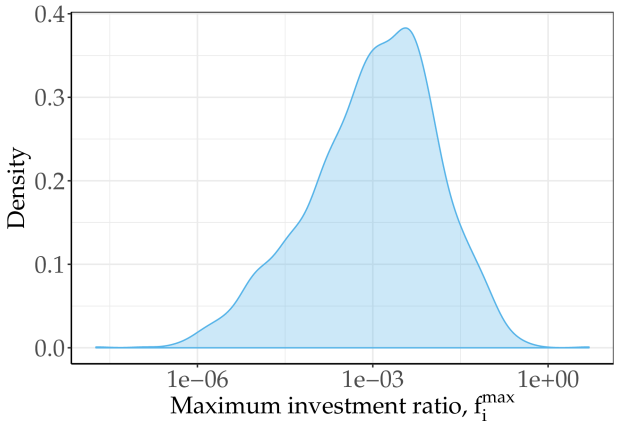

Figure A.5: Empirical probability density function of $f_i^{\max}$ for all the funds in the $n > n^*$ region.

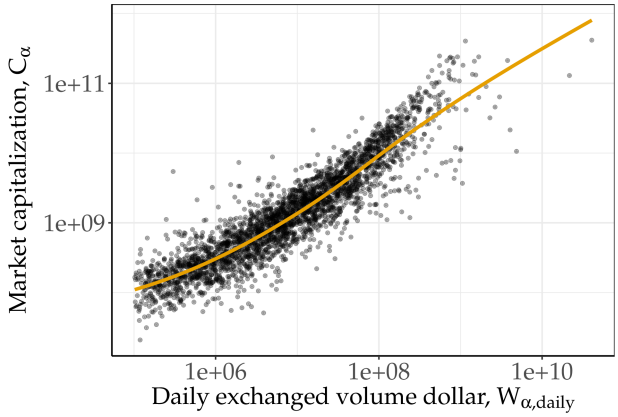

Figure A.6: Market capitalization as a function of the daily exchanged volume dollar averaged over the previous three months, for 2013-03-31. Fitting data to $C_\alpha = W^\eta_{\alpha,\text{daily}}$. We find $\eta \simeq 1$ for all the dates in our database, confirming the hypothesis that the daily exchanged volume dollar of an asset is approximately proportional to its market capitalization.

By iterating those steps we obtain Fig. 4.1.

Since the simulation outputs a portfolio for every fund, we can directly infer the number of investors m_α of every security.

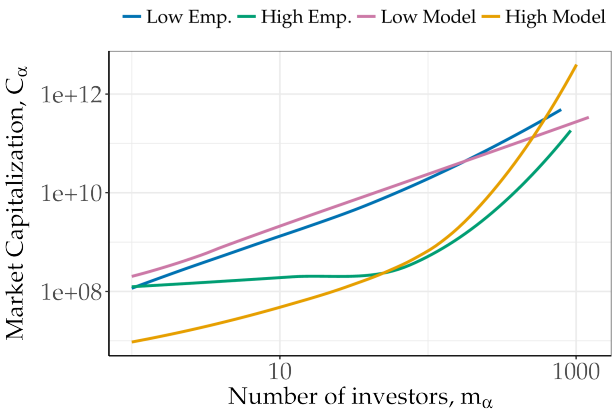

Figure A.7: We separate the contribution from the low and highly diversified region. The origin of the discrepancy observed in Fig. 4.4 appears to be mainly due to the highly diversified region (Green dots for the empirical data, and orange dots for the model).

References

Abarbanell, J. S. and Bushee, B. J., 1997. Fundamental analysis, future earnings, and stock prices. *Journal of Accounting Research*, 35, 1 (1997), 1–24. 80

Acemoglu, D.; Ozdaglar, A.; and Tahbaz-Salehi, A., 2015. Systemic risk and stability in financial networks. *American Economic Review*, 105, 2 (2015), 564–608. 31

Akyildirim, E. and Soner, H. M., 2014. A brief history of mathematics in finance. *Borsa Istanbul Review*, 14, 1 (2014), 57–63. 3

Allen, F. and Gale, D., 2000. Financial contagion. *Journal of political economy*, 108, 1 (2000), 1–33. 31

Amaral, L. A. N.; Buldyrev, S. V.; Havlin, S.; Maass, P.; Salinger, M. A.; Stanley, H. E.; and Stanley, M. H., 1997. Scaling behavior in economics: the problem of quantifying company growth. *Physica A: Statistical Mechanics and its Applications*, 244, 1-4 (1997), 1–24. 27

Anton, M. and Polk, C., 2014. Connected stocks. *The Journal of Finance*, 69, 3 (2014), 1099–1127. 31

Aristotle, 350bc. *Politics*. 2

Arthur, B. W., 1994. Inductive reasoning and bounded rationality: the El Farol problem. *Am. Econ. Rev.*, 84 (1994), 406–411. 56

Axtell, R. L., 2001. Zipf distribution of us firm sizes. *science*, 293, 5536 (2001), 1818–1820. 27, 28

Bachelier, L., 1900. *Théorie de la spéculation*. Gauthier-Villars. 3

Ball, R. and Brown, P., 1968. An empirical evaluation of accounting income numbers. *Journal of accounting research*, (1968), 159–178. 80

Barucca, P.; Bardoscia, M.; Caccioli, F.; D'Errico, M.; Visentin, G.; Battiston, S.; and Caldarelli, G., 2016. Network valuation in financial systems. (2016). 31

Battiston, S.; Farmer, J. D.; Flache, A.; Garlaschelli, D.; Haldane, A. G.; Heesterbeek, H.; Hommes, C.; Jaeger, C.; May, R.; and Scheffer, M., 2016. Complexity theory and financial regulation. *Science*, 351, 6275 (2016), 818–819. 31

Benjamini, Y. and Hochberg, Y., 1995. Controlling the False Discovery Rate : a practical and powerful approach to multiple testing. *J. R. Statistic. Soc. B 57*, (1995). 50

Blanc, P.; Donier, J.; and Bouchaud, J.-P., 2017. Quadratic hawkes processes for financial prices. *Quantitative Finance*, 17, 2 (2017), 171–188. 65

Blanchet, J. and Stauffer, A., 2013. Characterizing optimal sampling of binary contingency tables via the configuration model. *Random Structures & Algorithms*, 42, 2 (2013), 159–184. 33

Bluhm, M. and Krahnen, J. P., 2011. Default risk in an interconnected banking system with endogeneous asset markets. (2011). 31

Borgatti, S. P. and Halgin, D. S., 2011. Analyzing affiliation networks. *The Sage handbook of social network analysis*, 1 (2011), 417–433. 33

Bottazzi, G. and Secchi, A., 2003. Common properties and sectoral specificities in the dynamics of us manufacturing companies. *Review of Industrial Organization*, 23, 3-4 (2003), 217–232. 28

Bouchaud, J.-P., 2001. Power laws in economics and finance: some ideas from physics. (2001). 27

Bouchaud, J.-P., 2002. An introduction to statistical finance. *Physica A: Statistical Mechanics and its Applications*, 313, 1 (2002), 238–251. 4

Bouchaud, J.-P.; Gefen, Y.; Potters, M.; and Wyart, M., 2004. Fluctuations and response in financial markets: the subtle nature of 'random' price changes. *Quantitative finance*, 4, 2 (2004), 176–190. 65

Brunnermeier, M. K., 2009. Deciphering the liquidity and credit crunch 2007-2008. *Journal of Economic perspectives*, 23, 1 (2009), 77–100. 31

Butler, K. and Stephens, M., 1993. The distribution of a sum of binomial random variables. Technical report, STANFORD UNIV CA DEPT OF STATISTICS. 52

Caccioli, F.; Shrestha, M.; Moore, C.; and Farmer, J. D., 2014. Stability analysis of financial contagion due to overlapping portfolios. *Journal of Banking & Finance*, 46 (2014), 233–245. 31, 49

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

Cesa, M., 2017. A brief history of quantitative finance. *Probability, Uncertainty and Quantitative Risk*, 2, 1 (2017), 6. 4

Challet, D., 2016. Regrets, learning and wisdom. *The European Physical Journal Special Topics*, 225, 17-18 (2016), 3137–3143. 2

Challet, D.; Chicheportiche, R.; Lallouache, M.; and Kassibrakis, S., 2016. Trader lead-lag networks and order flow prediction. (2016). 65

Chan-Lau, J. A.; Espinosa, M.; Giesecke, K.; and Solé, J. A., 2009. Assessing the systemic implications of financial linkages. (2009). 31

Cifuentes, R.; Ferrucci, G.; and Shin, H. S., 2005. Liquidity risk and contagion. *Journal of the European Economic Association*, 3, 2-3 (2005), 556–566. 31

Cimini, G. and Serri, M., 2016. Entangling credit and funding shocks in interbank markets. *PloS one*, 11, 8 (2016), e0161642. 31

Cimini, G.; Squartini, T.; Garlaschelli, D.; and Gabrielli, A., 2015. Systemic risk analysis on reconstructed economic and financial networks. *Scientific reports*, 5 (2015), 15758. 31

Clauset, A.; Shalizi, C. R.; and Newman, M. E., 2009. Power-law distributions in empirical data. *SIAM review*, 51, 4 (2009), 661–703. 27

Cleveland, W. S.; Grosse, E.; and Shyu, W. M., 1992. Local regression models. *Statistical models in S*, 2 (1992), 309–376. 59

Cont, R. and Wagalath, L., 2014. Fire sales forensics: measuring endogenous risk. *Mathematical Finance*, (2014). 31, 49

Corsi, F.; Marmi, S.; and Lillo, F., 2016. When micro prudence increases macro risk: The destabilizing effects of financial innovation, leverage, and diversification. *Operations Research*, 64, 5 (2016), 1073–1088. 32

Davis-Stober, C. P.; Budescu, D. V.; Dana, J.; and Broomell, S. B., 2014. When is a crowd wise? *Decision*, 1, 2 (2014), 79. 55

de Lachapelle, D. M. and Challet, D., 2010. Turnover, account value and diversification of real traders: evidence of collective portfolio optimizing behavior. *New Journal of Physics*, 12, 7 (2010), 075039. 57

Dechow, P. M.; Hutton, A. P.; Meulbroek, L.; and Sloan, R. G., 2001. Short-sellers, fundamental analysis, and stock returns. *Journal of Financial Economics*, 61, 1 (2001), 77–106. 80

Diamond, D. W., 1985. Optimal release of information by firms. *The journal of finance*, 40, 4 (1985), 1071–1094. 80

Dosi, G., 2005. Statistical regularities in the evolution of industries: a guide through some evidence and challenges for the theory. Technical report, LEM working paper series. 28

Easley, D.; de Prado, M. L.; and O'Hara, M., 2016. Discerning information from trade data. *Journal of Financial Economics*, 120, 2 (2016), 269–285. 25

Eisenberg, L. and Noe, T. H., 2001. Systemic risk in financial systems. *Management Science*, 47, 2 (2001), 236–249. 31

Eisler, Z. and Kertesz, J., 2006. Size matters: some stylized facts of the stock market revisited. *The European Physical Journal B-Condensed Matter and Complex Systems*, 51, 1 (2006), 145–154. 37

Fama, E. F., 1965. The behavior of stock-market prices. *The journal of Business*, 38, 1 (1965), 34–105. 79

Fama, E. F., 1998. Market efficiency, long-term returns, and behavioral finance. *Journal of financial economics*, 49, 3 (1998), 283–306. 56

Farmer, J. D., 1999. Physicists attempt to scale the ivory towers of finance. *Computing in Science & Engineering*, 1, 6 (1999), 26–39. 6, 27

Fox, G.; Qiu, J.; Jha, S.; Ekanayake, S.; and Kamburugamuve, S., 2015. Big data, simulations and hpc convergence. In *Big Data Benchmarking*, 3–17. Springer. 1

Frankel, R. and Lee, C. M., 1998. Accounting valuation, market expectation, and cross-sectional stock returns. *Journal of Accounting and economics*, 25, 3 (1998), 283–319. 80

Friedman, M., 1953. The case for flexible exchange rates. (1953). 79

Gabaix, X.; Gopikrishnan, P.; Plerou, V.; and Stanley, H. E., 2003a. A theory of power-law distributions in financial market fluctuations. *Nature*, 423, 6937 (2003), 267–270. 27, 28

Gabaix, X.; Gopikrishnan, P.; Plerou, V.; and Stanley, H. E., 2006. Institutional investors and stock market volatility. *The Quarterly Journal of Economics*, 121, 2 (2006), 461–504. 28

Gabaix, X.; Ramalho, R.; and Reuter, J., 2003b. Power laws and mutual fund dynamics. Technical report, MIT mimeo. 28

Gai, P. and Kapadia, S., 2010. Contagion in financial networks. *Proc. R. Soc. A*, 466, 2120 (2010), 2401–2423. 31

Galton, F., 1907. Vox populi (the wisdom of crowds). *Nature*, 75 (1907), 450–51. 55

Giannone, D.; Reichlin, L.; and Small, D., 2008. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55, 4 (2008), 665–676. 2

Gillespie, C., 2015. powerlaw: Analysis of heavy tailed distributions. *R package version 0.30. 0, URL http://CRAN. R-project. org/package= poweRlaw*, (2015). 27

Gionis, A.; Mannila, H.; Mielikäinen, T.; and Tsaparas, P., 2007. Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1, 3 (2007), 14. 33

Glasserman, P. and Young, H. P., 2015. How likely is contagion in financial networks? *Journal of Banking & Finance*, 50 (2015), 383–399. 31, 32

Goldberg, D. S. and Roth, F. P., 2003. Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences*, 100, 8 (2003), 4372–4376. 33

Graham, J. R.; Harvey, C. R.; and Rajgopal, S., 2005. The economic implications of corporate financial reporting. *Journal of accounting and economics*, 40, 1-3 (2005), 3–73. 81

Greenwood, R. and Thesmar, D., 2011. Stock price fragility. *Journal of Financial Economics*, 102, 3 (2011), 471–490. 31

Greig, A. C., 1992. Fundamental analysis and subsequent stock returns. *Journal of Accounting and Economics*, 15, 2-3 (1992), 413–442. 80

Gualdi, S.; Tarzia, M.; Zamponi, F.; and Bouchaud, J.-P., 2015. Tipping points in macroeconomic agent-based models. *Journal of Economic Dynamics and Control*, 50 (2015), 29–61. 64

Haldane, A. G. and May, R. M., 2011. Systemic risk in banking ecosystems. *Nature*, 469, 7330 (2011), 351. 31

Hamermesh, D. S., 2013. Six decades of top economics publishing: Who and how? *Journal of Economic Literature*, 51, 1 (2013), 162–72. 4

Härdle, W. and Kirman, A., 1995. Nonclassical demand: A model-free examination of price-quantity relations in the Marseille fish market. *Journal of Econometrics*, 67 (1995), 227–257. 56

Hartley, J. E. and Hartley, J. E., 2002. *The representative agent in macroeconomics*. Routledge. 55

Healy, P. M. and Palepu, K. G., 2001. Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of accounting and economics*, 31, 1-3 (2001), 405–440. 80

Hill, S. and Ready-Campbell, N., 2011. Expert stock picker: the wisdom of (experts in) crowds. *International Journal of Electronic Commerce*, 15, 3 (2011), 73–102. 55

Hogarth, R. M., 1978. A note on aggregating opinions. *Organizational Behavior and Human Performance*, 21, 1 (1978), 40–46. 56

Hong, L. and Page, S. E., 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 46 (2004), 16385–16389. 55, 56, 64

Hong, Y., 2013. On computing the distribution function for the poisson binomial distribution. *Computational Statistics & Data Analysis*, 59 (2013), 41–51. 52

Horvát, E.-Á. and Zweig, K. A., 2013. A fixed degree sequence model for the one-mode projection of multiplex bipartite graphs. *Social Network Analysis and Mining*, 3, 4 (2013), 1209–1224. 33

Iori, G.; Jafarey, S.; and Padilla, F. G., 2006. Systemic risk on the interbank market. *Journal of Economic Behavior & Organization*, 61, 4 (2006), 525–542. 31

Khandani, A. and Lo, A. W., 2007. What happened to the quants in august 2007? (2007). 45

Kirman, A. P., 1992. Whom or what does the representative individual represent? *The Journal of Economic Perspectives*, 6, 2 (1992), 117–136. 56

Krause, A. and Giansante, S., 2012. Interbank lending and the spread of bank failures: A network model of systemic risk. *Journal of Economic Behavior & Organization*, 83, 3 (2012), 583–608. 31

Landemore, H. E., 2012. Why the many are smarter than the few and why it matters. *Journal of public deliberation*, 8, 1 (2012). 55

Latapy, M.; Magnien, C.; and Del Vecchio, N., 2008. Basic notions for the analysis of large two-mode networks. *Social networks*, 30, 1 (2008), 31–48. 32

Lee, C. and Ready, M. J., 1991. Inferring trade direction from intraday data. *The Journal of Finance*, 46, 2 (1991), 733–746. 25

Lev, B. and Thiagarajan, S. R., 1993. Fundamental information analysis. *Journal of Accounting research*, (1993), 190–215. 80

Lillo, F., 2007. Limit order placement as an utility maximization problem and the origin of power law distribution of limit order prices. *The European Physical Journal B*, 55, 4 (2007), 453–459. 65

Lillo, F. and Farmer, J. D., 2004. The long memory of the efficient market. *Studies in nonlinear dynamics & econometrics*, 8, 3 (2004). 65

Lillo, F.; Mike, S.; and Farmer, J. D., 2005. Theory for long memory in supply and demand. *Physical review e*, 71, 6 (2005), 066122. 65, 66, 67, 75

Lorenz, J.; Rauhut, H.; Schweitzer, F.; and Helbing, D., 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108, 22 (2011), 9020–9025. 55, 64

Lux, T., 1995. Herd behaviour, bubbles and crashes. *The economic journal*, (1995), 881–896. 64

Lynch, P. and Zumbach, G., 2003. Market heterogeneities and the causal structure of the volatility. *Quantitative Finance*, 3 (2003), 320–331. 65

Malkiel, B. G., 2003. The efficient market hypothesis and its critics. *The Journal of Economic Perspectives*, 17, 1 (2003), 59–82. 56

Malkiel, B. G. and Fama, E. F., 1970. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25, 2 (1970), 383–417. 56

Mandelbrot, B., 1967. The variation of some other speculative prices. *The Journal of Business*, 40, 4 (1967), 393–413. 27

Mandelbrot, B. B., 1997. The variation of certain speculative prices. In *Fractals and scaling in finance*, 371–418. Springer. 27

Mantegna, R. N. and Stanley, H. E., 1999. *Introduction to econophysics: correlations and complexity in finance*. Cambridge university press. 4

Markowitz, H., 1952. Portfolio selection*. *The journal of finance*, 7, 1 (1952), 77–91. 3

May, R. M. and Arinaminpathy, N., 2010. Systemic risk: the dynamics of model banking systems. *Journal of the Royal Society Interface*, 7, 46 (2010), 823–838. 31

McCulloh, I.; Lospinoso, J.; and Carley, K., 2010. The link probability model: A network simulation alternative to the exponential random graph model. (2010). 33, 34

Miller, R. G. J., 1981. *Simultaneous Statistical Inference*. Springer Series in Statistics. Springer-Verlag, New York. 50

Mizuno, T.; Takayasu, M.; and Takayasu, H., 2004. The mean-field approximation model of company's income growth. *Physica A: Statistical Mechanics and its Applications*, 332 (2004), 403–411. 27

Muchnik, L.; Aral, S.; and Taylor, S. J., 2013. Social influence bias: A randomized experiment. *Science*, 341, 6146 (2013), 647–651. 55, 64

Muggeo, V. M., 2003. Estimating regression models with unknown break-points. *Statistics in medicine*, 22, 19 (2003), 3055–3071. 83

Neal, Z., 2013. Identifying statistically significant edges in one-mode projections. *Social Network Analysis and Mining*, 3, 4 (2013), 915–924. 32

Neal, Z., 2014. The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. *Social Networks*, 39 (2014), 84–97. 32, 33, 34

Nofer, M. and Hinz, O., 2014. Are crowds on the internet wiser than experts? the case of a stock prediction community. *Journal of Business Economics*, 84, 3 (2014), 303–338. 55

Ou, J. A. and Penman, S. H., 1989. Financial statement analysis and the prediction of stock returns. *Journal of accounting and economics*, 11, 4 (1989), 295–329. 80

Pareto, V., 1964. *Cours d'économie politique*, vol. 1. Librairie Droz. 27

Park, J. and Newman, M. E., 2004. Statistical mechanics of networks. *Physical Review E*, 70, 6 (2004), 066117. 34, 51

Penman, S. H., 1992. Return to fundamentals. *Journal of Accounting, Auditing & Finance*, 7, 4 (1992), 465–483. 80

Roychowdhury, S., 2006. Earnings management through real activities manipulation. *Journal of accounting and economics*, 42, 3 (2006), 335–370. 81

Saracco, F.; Clemente, R. D.; Gabrielli, A.; and Squartini, T., 2016. Grandcanonical projection of bipartite networks. *manuscript in preparation*, (2016). 52

Saracco, F.; Di Clemente, R.; Gabrielli, A.; and Squartini, T., 2015. Randomizing bipartite networks: the case of the world trade web. *Scientific Reports*, 5 (2015), 10595. 33, 48, 51

Schwarzkopf, Y. and Farmer, J. D., 2008. Time evolution of the mutual fund size distribution. *arXiv preprint arXiv:0807.3800*, (2008). 28

Schwarzkopf, Y. and Farmer, J. D., 2010. Empirical study of the tails of mutual fund size. *Physical Review E*, 81, 6 (2010), 066113. 28

Serrano, M. Á.; Boguná, M.; and Vespignani, A., 2009. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences*, 106, 16 (2009), 6483–6488. 33

Shleifer, A. and Vishny, R. W., 1992. Liquidation values and debt capacity: A market equilibrium approach. *The Journal of Finance*, 47, 4 (1992), 1343–1366. 31

Shleifer, A. and Vishny, R. W., 2010. Fire sales in finance and macroeconomics. Technical report, National Bureau of Economic Research. 31

Smith, A., 1759. *The theory of moral sentiments*. 80

Smith, A., 1776. *The Wealth of Nations*. London: Methuen & Co., Ltd. 80

Solomon, S. and Richmond, P., 2001. Power laws of wealth, market order volumes and market returns. *Physica A: Statistical Mechanics and its Applications*, 299, 1-2 (2001), 188–197. 27

Squartini, T. and Garlaschelli, D., 2011. Analytical maximum-likelihood method to detect patterns in real networks. *New Journal of Physics*, 13, 8 (2011), 083001. 51

Stanley, M. H.; Amaral, L. A.; Buldyrev, S. V.; Havlin, S.; Leschhorn, H.; Maass, P.; Salinger, M. A.; and Stanley, H. E., 1996. Scaling behaviour in the growth of companies. *Nature*, 379, 6568 (1996), 804–806. 27, 28

Stanley, M. H.; Buldyrev, S. V.; Havlin, S.; Mantegna, R. N.; Salinger, M. A.; and Stanley, H. E., 1995. Zipf plots and the size distribution of firms. *Economics letters*, 49, 4 (1995), 453–457. 28

Staum, J., 2012. Counterparty contagion in context: Contributions to systemic risk. (2012). 31

Steegen, S.; Dewitte, L.; Tuerlinckx, F.; and Vanpaemel, W., 2014. Measuring the crowd within again: a pre-registered replication study. *Frontiers in psychology*, 5 (2014). 55

Sudarsanam, P.; Pilpel, Y.; and Church, G. M., 2002. Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rrna transcription motifs in saccharomyces cerevisiae. *Genome research*, 12, 11 (2002), 1723–1731. 33

Surowiecki, J., 2005. *The wisdom of crowds*. Anchor. 55

Theissen, E., 2001. A test of the accuracy of the lee/ready trade classification algorithm. *Journal of International Financial Markets, Institutions and Money*, 11, 2 (2001), 147–165. 25

Toke, I. M., 2016. Reconstruction of order flows using aggregated data. *Market microstructure and liquidity*, 2, 02 (2016), 1650007. 25

Toth, B.; Palit, I.; Lillo, F.; and Farmer, J. D., 2015. Why is equity order flow so persistent? *Journal of Economic Dynamics and Control*, 51 (2015), 218–239. 67

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

Tumminello, M.; Miccichè, S.; Lillo, F.; Piilo, J.; and Mantegna, R. N., 2011. Statistically validated networks in bipartite complex systems. *PLOS ONE 6 (3) e17994*, (2011). 33, 34, 50, 53

Vul, E. and Pashler, H., 2008. Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19, 7 (2008), 645–647. 55

Vuong, Q. H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, (1989), 307–333. 28

Weatherall, J. O., 2013. *The physics of wall street: a brief history of predicting the unpredictable*. Houghton Mifflin Harcourt. 3

Xue, B.; Zhang, M.; Browne, W. N.; and Yao, X., 2016. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation*, 20, 4 (2016), 606–626. 2

Zhai, Y.; Ong, Y.-S.; and Tsang, I. W., 2014. The emerging" big dimensionality". *IEEE Computational Intelligence Magazine*, 9, 3 (2014), 14–26. 2, 5

Zhou, W.-X.; Mu, G.-H.; Chen, W.; and Sornette, D., 2011. Investment strategies used as spectroscopy of financial markets reveal new stylized facts. *PloS one*, 6, 9 (2011), e24391. 65

Zumbach, G., 2004. How trading activity scales with company size in the ftse 100. *Quantitative Finance*, 4, 4 (2004), 441–456. 37

Zumbach, G., 2009. Time reversal invariance in finance. *Quantitative Finance*, 9, 5 (2009), 505–515. 65

Zweig, K. A. and Kaufmann, M., 2011. A systematic approach to the one-mode projection of bipartite graphs. *Social Network Analysis and Mining*, 1, 3 (2011), 187–218. 33

université PARIS-SACLAY
ÉCOLE DOCTORALE
INTERFACES
Approches Interdisciplinaires : Fondements, Applications et Innovation

Titre : Comportement des traders institutionnels et microstructure des marchés : une approche big data

Mots clés : systèmes complexes, microstructure des marchés, marchés financiers

Résumé : Cette thèse est composée de quatre chapitres.
Le premier chapitre est une description préliminaire de la base de données Factset Ownership. Nous en donnons une description statistique et exposons quelques faits stylisés caractérisant notamment la structure du portefeuille des institutions financières et fonds d'investissement, ainsi que la capitalisation boursière des entreprises y étant recensées.
Le second chapitre propose une méthode d'évaluation statistique de la similarité entre des paires de portefeuilles d'institutions financières. Une paire statistiquement significative donnant lieu à la création d'un lien de similarité entre ces deux entités, nous sommes en mesure de projeter un réseau à l'origine bi-partite (entre institutions financières et entreprises) en un réseau mono-partite (entre institutions uniquement) afin d'en étudier l'évolution de sa structure au cours du temps. En effet, d'un point de vue économique, il est suspecté que les motifs d'investissements similaires constituent un facteur de risque important de contagion financière pouvant être à l'origine de banqueroutes aux conséquenc es systémiques significatives.
Le troisième chapitre s'intéresse aux comportements collectifs des gestionnaires de fonds d'investissement et, en particulier, à la manière dont la structure du portefeuille de ces fonds prend en compte, en moyenne, de façon optimale les frais de transaction en présence de faibles contraintes d'investissements. Ce phénomène où, dans de nombreuses situations, la médiane ou la moyenne des estimations d'un groupe de personnes est étonnamment proche de la valeur réelle, est connu sous le nom de sagesse de la foule
Le quatrième chapitre est consacré à l'étude simultanée de données de marché. Nous utilisons plus de 6.7 milliards de trades de la base de données Thomson-Reu ters Tick History, et de données de portefeuille de la base FactSet Ownership. Nous étudions la dynamique tick-à-tick du carnet d'ordres ainsi que l'action aggrégée, c'est-à-dire sur une échelle de temps bien plus grande, des fonds d'investissement. Nous montrons notamment que la mémoire longue du signe des ordres au marché est bien plus courte en présence de l'action, absolue ou directionnelle, des fonds d'investissement. Réciproquement nous expliquons dans quelle mesure une action caractérisée par une mémoire faible est sujette à du trading directionnel provenant de l'action des fonds d'investissement.

Title : Large-trader behaviour and market microstructure: a big data approach

Keywords : financial networks, market microstructure, complex systems, collective optimization

Abstract : The thesis is divided into four parts.
Part I introduces and provides a technical description of the FactSet Ownership dataset together with some preliminary statistics and a set of stylized facts emerging from the portfolio structure of large financial institutions, and from the capitalization of recorded securities.
Part II proposes a method to assess the statistical significance of the overlap between pairs of heterogeneously diversified portfolios. This method is then applied to public assets ownership data reported by financial institutions in order to infer statistically robust links between the portfolios of financial institutions based on similar patterns of investmen t. From an economic point of view, it is suspected that the overlapping holding of financial institution is an important channel for financial contagion with the potential to trigger fire sales and thus severe losses at a systemic level.
Part III investigates the collective behaviour of fund manager and, in particula r, how the average portfolio structure of institutional investors optimally accounts for transactions costs when investment constraints are weak. The collective ability of a crowd to accurately estimate an unknown quantity is known as the Wisdom of the Crowd . In many situation, the median or average estimate of a group of unrelated individuals is surprisingly close to the true value.
In Part IV, we use more than 6.7 billions of trades from the Thomson-Reuters Tick History database and the ownership data from FactSet. We show how the tick-by-tick dynamics of limit order book data depends on the aggregate actions of large funds acting on much larger time scale. In particular, we find that the well-established long memory of marker order signs is markedly weaker when large investment funds trade in a markedly directional way or when their aggregate participation ratio is large. Conversely, we investigate to what respect an asset with a weak memory experiences direction trading from large funds.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France