UNIONE EUROPEA

REGIONE SICILIA

Ministero dell'Istruzione, dell'Università e della Ricerca

UNIVERSITA' DEGLI STUDI DI PALERMO

Dipartimento di Scienze Economiche, Aziendali e Statistiche

XXXII ciclo di dottorato in Scienze Economiche e Statistiche

Development of multivariate and network models for the analysis of Big Data

Applications in economics, insurance, and social sciences

Pietro Vassallo

SUPERVISOR
Andrea Consiglio

ADVISOR
Michele Tumminello

Academic year 2019/2020

"Audentes Fortuna Iuvat"
— Virgilio

Contents

List of Tables

List of Figures

Acronyms

AIA Antifraud Integrated Archive

ATT Average Treatment effect on Treated

CDS Credit Default Swap

COG Clusters of Ortholous Group

FEVD Forecast Error Variance Decomposition

GDF Generalized Dynamic Factor

GMM Gaussian Mixture Model

ISAAC Investigation System for Antifraud ACtivity

IVASS Istituto di Vigilanza sulle ASSicurazioni

KD Kristillisdemokraatit - Christian Democrats

KESK Keskusta - Centre Party

KOK Kokoomus - National Coalition Party

LASSO Least Absolute Shrinkage and Selection Operator

MA Moving Average

MCC Matthews Correlation Coefficient

MP Members of Parliament

OLS Ordinary Least Squares

PS Perussuomalaiset - Finns Party

RAE Research Assessment Exercise

REF Research Excellence Framework

RKP Ruotsalainen kansanpuolue - Swedish People's Party

RMSPE Root Mean Square Predictive Error

ROC Receiver Operating Characteristic

SCM Synthetic Control Method

SDP Sosialidemokraattinen puolue - Social Democratic Party

SVN Statistically Validated Network

VAR Vector AutoRegression

VAS Vasemmistoliitto - Left Alliance

VIHR Vihreä liitto - Green League

Introduction

Finding ways to explain, predict, and replicate behavioural patterns of the agents of a complex system has been the focus of scholars and policy makers in many areas of science and society, for example, biology, economics, engineering and sociology, to cite some of them. Due to the volume, velocity and variety of the data collected and managed by the increasingly powerful and capable IT technologies, there is a growing need to develop efficient mathematical and statistical methods to deal with the challenges arising from the complexity of real systems in the era of Big Data.

Ranging from biological molecules to economic and financial systems, across multiple scales, complex systems involve agents whose multiple micro-level interactions yield a macro-level behavior in a non-linear fashion.

In the last few decades, scientists have started to study complex systems by resorting to complex networks. The advantage of using complex networks is that they allow analysts or researchers to abstract the complexity that characterises real complex systems making very few assumptions on the type of interactions among their components. Moreover, networks provide a holistic approach to the comprehension of complex systems by focusing on the study of the system as a whole rather than on its separate parts.

The increasing complexity of societies suggests that there will be a growing need for the understanding of real complex systems. The insights of complex systems research and its methodologies may become pervasive in guiding research and policy decisions across disciplines. Indeed, national and international policies should be informed by the science of complex systems to undertake decisions with global effects.

In this thesis I will develop multivariate statistical and network methods for the study of complex systems. In particular, I will focus my analysis on the study of bipartite complex networks and their applications to (i) economics to understand the contagion effect between sovereign and financial institutions, (ii) to insurance surveillance to uncover fraudsters and (iii) to social science to study the effect of the politics of REF on research excellence of universities in the UK.

In what follows, I will discuss the content of each chapter in more detail by giving the reader a useful description of the context specific to each study.

Complex Systems and Complex Networks (Chapter 1)

In this chapter, I will introduce complex systems and I will highlight the differences between complexity and complicatedness in real life phenomena. A fundamental tool for my analysis is given by bipartite networks. I will give a mathematical definition of bipartite graphs and their main properties. Finally, I will introduce statistically validated networks, that are used to remove the intrinsic noise contained in the data, while putting in the foreground the systematic patterns of the observed network.

Emergent phenomena in bipartite complex systems with a double heterogeneity (Chapter 2)

Complex bipartite systems are studied in many application fields such as biology, physics, economics, and social sciences, and they can suitably be described as bipartite networks. Examples of bipartite networks are: criminals-crimes, actors-movies, people-accidents, authors-universities, General Practitioners (GP)-hospitals, etc. In general, when dealing with bipartite networks, we are interested in measuring the similarity between subject-nodes, given their linkage structure with the item-nodes. Although binary Pearson's correlation coefficient has proved effective to investigate the similarity structure of some real-world bipartite networks, when both node sides of the network are characterized by heterogeneity (high variability in the degree distributions), the sample covariance and correlation coefficients are biased.

In this chapter, I will introduce a weighted covariance and correlation estimator and show results that improve upon traditional similarity measures, when double-heterogeneity affects bipartite networks in real systems.

SVN to detect fraudsters' communities in the Italian car insurance sector (Chapter 3)

Accident claims are an example of heterogeneous and multidimensional data as they include—not being exhaustive—coded identity of all the subjects directly involved in an accident, such as, drivers, passengers, car owners, witnesses, and pedestrians; professionals, such as, doctors, lawyers, car repairs, as well as details about injuries, fatalities, requested amount, property damage, place and time of the accident, and all about the vehicles involved. Fraud is a social phenomenon and fraudsters often act in collaboration with players having different roles. Supervised methods, although they add value to the analysis, show two main drawbacks: first, their calibration is based on a set of known frauds which are very difficult to obtain, and that are a very small sample with respect to the total claims. Second, they miss a peculiar feature of frauds in motor insurance, i.e., the existence of "criminal infrastructures", which also encompass the professional profiles operating in this field.

In this chapter I will describe the development of an investigation system based on the application of bipartite networks to highlight the relationships between subjects and accidents or vehicles and accidents. This is a general approach that allows us to include the whole spectrum of actors around a claim: from the drivers to the legal professionals. Starting from the dense complex network, we will construct statistically validated networks to prune the connections that score a low likelihood level with respect to random chance. In this step only structures with very strong ties will appear, thus signalling potential group of fraudsters. I will also formalize the filtering rules through probability models and test specific methods to assess the existence of communities for very large networks and propose new alert metrics of suspicious structures. I will apply the above methodology to a real database—the Antifraud Integrated Archive (AIA)—and compare results to out-of-sample fraud scams assessed by the judicial authorities.

Impact Evaluation of the REF in the UK (Chapter 4)

The REF is the main UK government policy on public research in the last 30 years. It aims at promoting and rewarding research excellence through competition for limited resources. Despite the national interest and the severe criticisms about the effectiveness of the Research Assessment Exercise (RAE), very little has been done to assess its impact on research excellence outcomes. In this chapter I will exploit the publication and affiliation data contained in the Scopus database to empirically evaluate the impact of the REF on the research productivity of universities in terms of both quantity and quality of published scientific articles. To do so, I will rely on the Synthetic Control Method (SCM) [2, 3], which uses a time series of the outcome of the treated UK universities prior to the intervention and creates a counterfactual set of outcomes against which compare the outcomes of the treated group after the intervention. We take as a control the US academic system due to its strong ties with the UK one, such as their common language and the research productivity that is financially incentivised. I will compute both individual and ATT for each of the years amid the REF implementations of 2008 and 2014, eventually computing an overall ATT for the whole period as well.

Spillover effects analysis in the Credit Default Swap (CDS) Market (Chapter 5)

Sovereigns are exposed to bank risk and, at the same time, banks are exposed to sovereign risk. During the euro-area sovereign debt crisis, this two-way risk exposure generated a "vicious circle", also known as the "doom loop" [66]. At a point when government bonds were considered risky assets, euro-area banks faced with both balance sheet and reputational risks, making it hard to compete with their non-euro area counterparts, forcing to tight their exposure to sovereign credit risk, thus igniting the most disruptive financial crisis has ever jeopardized the Euro currency system.

Over the years the failure of financial institutions has led to fears of system failure from domino effects of one failed entity bringing down others. Indeed, this way of thinking has given rise to concepts such as financial contagion and entities *too interconnected to fail*, and since then the interests have moved from the study of mechanisms of single entities towards the point where the interaction between entities has become crucial and more important than a single mechanism on its own. Financial distress and the consequences of risk propagation will depend on both the magnitude of external shocks and the position of hit entities in the system. The study of negative externalities cannot be done by using a perspective based on individuals but, rather, using a holistic approach to the problem, analysing the entire financial system as a whole.

This chapter is devoted to the application of SVN for the study of risk contagion among financial institutions such as banks and sovereigns in the CDS market. I will compare up-to-date econometric methods, that serve for the purpose of computing spillover effects based on regularized Vector AutoRegression (VAR) models, and forecast variance error decomposition. I will show that SVNs provide robust insights on how contagion transmits between sovereigns and financial institutions. I will also show that traditional approaches to compute the spillover effect can benefit when used in companion with SVNs.

Chapter 1

Complex Systems and Networks

1.1 Complex Systems

Most of the things happening around us are the result of a process of some kind: biological, e.g. when off-springs form from an organism, or when genetic characteristics of organisms change to allow the adaptation in the environments they may live in; physical or chemical, e.g. when a solid matter turns into liquid and then gas state, or the way to which planets and galaxies move in the universe; organizational, e.g. when individuals in a company specialize in specific tasks to optimize productivity and efficiency. Each one of these examples represents a *system*, made of elements with some degree of complexity that interact with each other, and that shows an evolution over time and space. Searching on the web, one can find the following definitions of system: "a set of things working together as parts of a mechanism or an interconnecting network; a complex whole", and, "a set of principles or procedures according to which something is done; an organized scheme or method". Of course, the way systems work are not random, at all. Moreover, the behaviour of any system strictly depends on the way its elements interacts with each other and also on the conditions of the outer environment they are involved in. Few examples of systems are: societies, cities, companies, markets, biological systems, financial markets, etc. A crucial property that is shared by all of these systems is *complexity*. The main assumption behind the idea of complex systems is that although their behaviours may seem random, they actually are governed by laws that determine specific patterns of evolution. Indeed, the seemingly chaotic behaviour does not lead to a total absence of order, but it mainly refers to an *ordered disorder* [138]. One can find many definitions of "complex system", but, as many complexity scientists point out, none of these represents a concise definition that manages to properly state what a complex system actually is, since they may depend on the context to be studied.

1.1.1 Complex versus complicated systems

People usually tend to erroneously interchange the adjective *complex* with the term *complicated* to describe a system. Indeed, there is a subtle difference between the notion of *complex system* and that of *complicated system*.

In general, a complicated system is pretty much related to the notion of *reductionism*, meaning that one can analyse and model the dynamics occurring within the system by considering all its parts one at a time and separately one from each other. They are viewed to have a large number of components that behave in a well-understood way leading to the resulting effect. Think of a clock as an example [170]: it has many heterogeneous components that have to work together as a network structure. Although clocks may be complicated systems, they cannot be considered complex, since a clock does not adapt to external conditions. Indeed, the requirement for a clock to work depends essentially on the fully functionality of each of its components, seen separately and individually: if just one of the components breaks down, then all the system won't work anymore.

In a complex system, by contrast, all individual parts are linked together, and their relations may change over time, adapting to both internal and external changes due to different scenarios of the outer environment. Moreover, the connections between the elements of a complex system are typically nonlinear, that implies that there is not a linear sequence of causes and effects in its behaviour [170]. Therefore, a crucial point when distinguishing a complex system from a complicated system regards the predictability of the system itself. In principle, as hard to understand as a complicated system can be, one can always know with certainty all the mechanism effects characterising the system. On the contrary, a complex system can be only predictable to some extent, and the level of uncertainty depends on many aspects characterising its complexity. Unlike complicated systems, the main property that characterizes complex systems is the presence of emergent phenomena, that take place at the macro level of the system and are very difficult to predict and to discern at small scales.

Nevertheless, a system could be really complex even if its elements are rather simple, e.g. the group of ants: while the behaviour of a single ant is assumed to be rather simple, when we consider the ants together as a whole entity, then their behaviour will result in a variety of interesting phenomena, such as foraging for food, bringing it to the anthill and leaving their pheromones on the route, so that other ants can follow the trail and find the food [52]. In a complex system even interactions of quite simple components can generate bewildering

behaviors [193]. Even if a complex system is locally really unordered, when observed into a higher level, it will exhibit ordered and structured patterns. In general, while the elements of a system behave according to their preferences, expectations and needs, their joint interactions could make the whole system have unexpected properties at the meso- and macro-level that are not directly known and induced by the single components. As Dekker once argued [52], "In a complex system, each component is ignorant of the behaviour of the system as a whole. This is a very important point. If each component "knew" what effects its actions had on the entire rest of the system, then all of the system's complexity would have to be present in that component. It isn't. This is the whole point of complexity and systems theory. Single elements do not contain all the complexity of the system. If they did, then reductionism could work as an analytic strategy: we could explain the whole simply by looking at the part".
When we ask ourselves "How does a cell phone work?", a mechanistic thinker, analysing the single components of the telephone's hardware, would reply that the device functioning is due to its internal mechanisms that give the possibility to make phone calls, take pictures, browse the internet, listen to music and so on, without focusing on the internal and external effects due to the other systems in its surroundings. Each functionality of the mobile phone corresponds to single components designed to make the device suitable for a specific activity. The reply of a system thinker, who uses a holistic approach to study a system, will be radically different and it will concern the multiple emergent phenomena related to mobile phone production at different levels and its implications to other systems. Starting from Coltan–a mineral used to improve the cell phone battery performances– Dekker explains how the system thinker will analyse the social, economic and environmental implications of Coltan extraction in Congo: such as the exploitation of miners that manually extract the mineral, the civil war to control the territories of extractions, the killing of gorillas in order to sell their meat to miners and rebels etc.

1.1.2 Emergent properties of complex adaptive systems

Suppose you are driving your car and along the path you are told that a specific street has been closed for some reason. This new piece of information will affect the behaviour of drivers that received it. These changes in the behaviours and interactions of people might have a systemic effect resulting in traffic congestion [111]. "Most people most of the time act iteratively in terms of local information, knowing almost nothing about the global connections or implications of what they are doing. However, these local actions do not remain simply local since they are captured, represented, marketed, circulated and generalized elsewhere [...]. The consequences for the global level are non-linear, large-scale, unpredictable and partially ungovernable. Small causes at certain places produce massive consequences elsewhere" [186].

In order to predict the behaviour of a complex system, complexity scientists have to find a way to extract the systematic pattern suggested by the system over time and space. It's worth to note that the system and its components will behave according to the information which flows within the system as well as to the state conditions of the outer environment. Also, it is possible to distinguish the system from the environment that surrounds it, allowing one to infer how the system responds to the external inputs–its adaptation behaviour and resilience–without knowing all its internal self-organizing rules.

An important property of most complex systems is that they can be viewed as having a hierarchical structure [170], where every layer of the system produces outputs that influence the way other layers work. Nevertheless, single elements will interact without knowing the effects that they could bring to the whole system. Moreover, a complex system is *resilient*, meaning that when a local shock takes place and a specific critical point is reached towards a new phase transition, it may progressively modify its behaviours, independently of the components at lower levels, showing *self-organized criticality* [17]. This aspect is not observed in complicated systems. It is in these phase transitions that the system shows *emergent phenomena*, new behaviours that can't be described as just the sum of the effects of individual components and that could reveal in multiple ways. A trivial example that gives the idea of emergent phenomenon is the formation of the so-called *Mexican Wave*, occurred for the first time during the 1986 FIFA World Cup held in Mexico, and for which spectators in a stadium stand and then sit in groups until every section in the stadium has participated in turn. The local coordination of individuals will eventually form a macro behaviour of the whole, where the crowd will look like a rolling ocean wave when seen from a distance. Moreover, complex systems could be very sensible to small perturbation: a small perturbation in a system could potentially cause a catastrophic modification in the future dynamics of the system (this assumption has been demonstrated by [125], and is known as the "butterfly effects", where only the flapping of a butterfly could be determinant for the formation of a tornado).

So, there is a difference between complexity and emergence of a complex system: complexity refers to the set of properties that characterize both the inner and outer environment of a system and all potential final actions it can take under certain future scenarios. Instead, emergence refers to the actual behaviours or actions eventually undertaken by the system. Some systems may exhibit a sort of quite stable behaviour "followed by a sudden shift to disequilibrium or to another, quite different equilibrium" [170]. The sensitivity of a system to initial conditions, which can lead to a ripple effect, is, for example, at the root of the sociological analysis of the transformations in modern societies, and in particular relating globalization, that introduced new perceptions of risk and vulnerability–e.g. the consequences of nuclear disasters, the spread of diseases and terroristic attacks–[21], the "glocalization" concept–that highlights the interplay between local interactions and global effects–[163], or climate change perception and collective action [128, 174].

It is possible to investigate the "mechanisms that create and sustain complexity" of real complex systems using an empirical approach. Unfortunately, in many cases, an unsupervised and direct inspection of all the interactions among elements of a system is impossible to do. Wolfram [197], with his principle of *computational irreducibility*, states that it is impossible to predict what a complex system will do, except by going through as many steps in the computation as the evolution of the system itself. This is why it's much more efficient to describe a complex system as a phenomenon in its own right, rather than regarding its individual components. Complex behaviour features can be captured with models that have simple underlying structures. This certainly makes research much easier, but this resistance to simplification is also a fundamental feature of complex systems.

1.1.3 Cascade Phenomena and Herd behaviours

Cascade phenomena are really common in complex social and economic systems. They are the consequences that the actions of one or few elements have on the collective response of an entire system. In particular, the actions of the few are spread in a sequential fashion across the system. A *positive feedback* refers to the influence of the elements of a system that, eventually, will cause an emergent phenomenon to take place, breaking the current equilibrium of the system itself. On the other hand, a *negative feedback (or balancing feedback)* is the ability of a system to contrast internal or external shocks in order to maintain its equilibrium over time and space. An example of negative feedbacks are the homeostatic processes of organisms, that enables various measures (e.g. body temperature, or blood sugar level) to be maintained within a desired range.

Instead, some examples of positive feedbacks are: the process leading to the applause of an audience, for which just few people clapping their hands can sequentially induce to the applause of the whole audience; also, the same logic applies to the phenomenon of standing ovation; or again, suppose there are two restaurants, say A and B, that are opposed along the two sides of a street and a group of people has to choose either one. Moreover, assume that restaurant A is crowded while restaurant B is not. Even if the group of people possess some private information about the good quality of the food served in restaurant B, they will tend to follow the choices of people that arrived before them, eventually going to restaurant A. Social scientists refer also to *information cascade* that influence people's behaviours, that dominate their private beliefs and make them even act irrationally (e.g., against what they think is optimal). A person can't directly observe the outside information that other people possess, but he or she makes inferences about this information from what they do. A very similar but slightly different concept is the *herd behaviour*, an uncoordinated behaviour of self-serving individuals. This type of phenomenon is for example observed in flocks of migrating birds or people that are in danger and panic following the way out from a building: their uncoordinated but self-serving movements cause emergent phenomena to occur. While informational cascades are more stable, herd behaviour is more easily disrupted, since in the latter private information possessed by individuals is not dominated by the behaviours of other individuals.

1.1.4 Non-stationarity of complex systems

The dynamics of real complex systems involve a certain degree of non-linearity in the interactions between elements. Also, complex systems could involve non-stationary processes in time and space domains. Since real-world systems evolve under transient conditions, the signals obtained from there tend to exhibit very many forms of non-stationarity. Indeed, the non-linear and non-stationary dynamics of the underlying processes pose a major challenge for accurate forecasting of space and time series. Recently, a review of the advancements made so far has been presented in [39].

1.1.5 Heterogeneity

Most real complex systems consist of elements that are very different from each other, both qualitatively and quantitatively speaking. Let's consider the financial system as an example: the elements of the system can be single individuals, families, small, medium and large companies, or even sovereigns.

Therefore, each element within the system will have a specific role with specific objectives, and a different importance for the stability of the system itself. Many real complex systems are heterogeneous, where, from the viewpoint of individual interactions, elements follow a power-law distribution and their network representation is said to be *scale-free* [154, 155]. In a scale-free network most of the elements show a low number of interactions while only a small proportion of them does show many interactions. Eventually, the topological structure of the network will not depend on its size, i.e. number of nodes in the network. Modelling the source of heterogeneity of a system can be crucial to draw accurate conclusions. Nevertheless, dealing with heterogeneity can often be a challenging task.

1.1.6 Motifs in complex networks

The functional properties of complex networks may be highlighted by the so-called motifs. A motif refers to a local and persistent structural pattern that occurs across a network. Motifs may be useful to study the functioning of a system as they may reflect a framework in which particular functions are achieved efficiently. They attract much attention because they allow to uncover structural design principles of complex networks. Although their relative simplicity, motifs are challenging to discover. Many methods have been proposed for motifs detection, under essentially two different paradigms such as exact counting methods [139], [85] (computationally heavy) and sampling methods [116] (faster but may be unreliable), pattern growth methods and so on ([166], [194], [153], [42]), all of them relying on the frequency concepts of sub-graphs and their statistical significance.

Many efforts have been made to analyse the data coming from complex systems, several of them focusing on cross-correlation between elements. This approach presents some drawbacks: i) it needs large statistics, in most cases requiring the assumption of quasi-stationarity of the process underlying the system; ii) it superimposes the model of dynamics [48]; iii) it disregards non-linear correlations (a way of taking into account non-linear correlations by estimating the mutual information has been proposed by Kraskov et al. (2004) [119]). An effective approach that manages to abstract complex systems and that relaxes the aforementioned assumptions is given by complex network theory, that started to attract complexity scientists, in particular following the two papers by Paul Erdős and Alfréd Rényi in 1959 [64] and by Mark Granovetter in 1973 [84], whose works marked the beginning of the application of network theories to the study of complex systems. Resorting to complex networks allows one to represent the interactions among the elements of a system without specifying the nature of their relationships, that can be either linear or non-linear, symmetric (correlation) or asymmetric (causality). A clear introduction and an exhaustive review of the major concepts and results achieved in the study of the structure and dynamics of complex networks can be found in [28, 124, 148, 126]. Eventually, networks give a flexible way to study the structures and dynamics of complex systems' phenomena with a holistic approach, and, because of these aspects, it is continuously engaging the interest of many scientists working on very different application fields.

1.2 Bipartite Complex Networks

Complex phenomena can be described through the relationships shared by their actors. A bipartite network is a useful tool to represent interactions occurring among the entities of a system involving two different groups of nodes. In Fig. 1.1 we display a bipartite network where the entities of the system are partitioned in two sets, U and S, and the relation between two any *nodes* of each set is reproduced through a *link* connecting the two nodes. There is an extensive literature on (bipartite) network methodology and its application to the analysis of social systems. An illustrative, but not exhaustive, list of papers includes: movies and actors [192, 18, 173], authors and scientific papers [90, 19, 152], email accounts and emails [134], mobile phones and phone calls [155], the criminal-crime relationship to assess generalist vs specialist behaviour in crime [184], the GOTCHA! system which is based on a bipartite graph relating companies and resources [187]. In graph theory, a bipartite network is a *graph* with two disjoint set of nodes. We provide in the next section the basic notation and definitions we will use throughout the paper.

We denote by $\mathcal{G}(V, E)$ a *graph* where V is the set of *vertices* and E is the set of *edges* connecting any couple of vertexes $v_i, v_j \in V$, where $i, j = 1, 2, \ldots, |V|$ and $(v_i, v_j) \in E$. The *neighborhood* of a vertex $v_i \in V$ is the sub-graph of $\mathcal{G}$ composed of the vertexes $v_j \in V$ and the edges $(v_i, v_j) \in E$. We denote by $N(v_i)$ the neighborhood of v_i and by $\deg(v_i)$ the *degree* of v_i, i.e., the number of edges incident to the vertex v_i. Here, we assume the graph is undirected. If we deal with directed graph, then a distinction between in-degree and out-degree must be done. Moreover, If there are no loops, $\deg(v_i)$ coincides with the number of vertexes of $N(v_i)$, excluding v_i itself.

A *bipartite graph* is characterized by two sets $U, S \subset V$, such that $V = U \cup S$ and $U \cap S = \emptyset$; moreover, $\forall i = 1, 2, \ldots, |U|$ and $\forall i = 1, 2, \ldots, |S|$ the edge $(u_i, s_j) \in E$ cannot have both vertex in the same set. We usually denote by

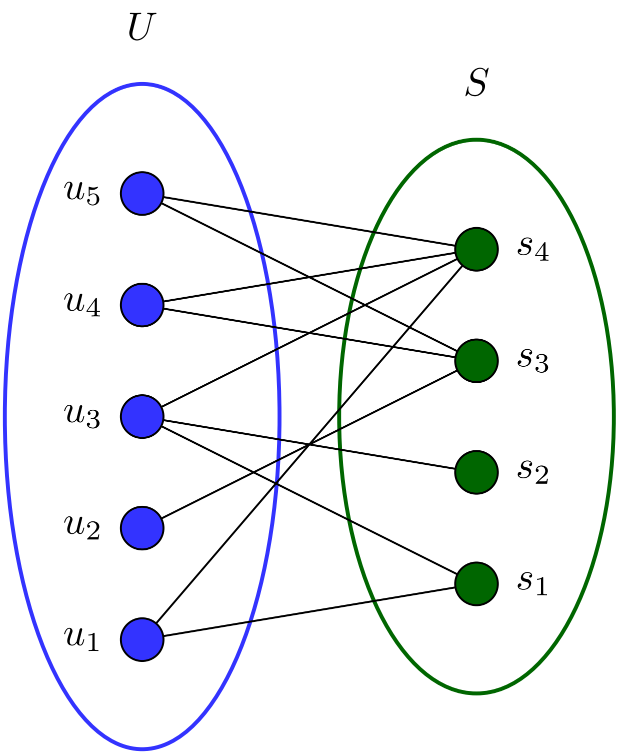


Figure 1.1: Bipartite network

$\mathcal{G}(U, S, E)$ a bipartite graph and we can represent it by a $|U| \times |S|$ matrix known as *bi-adjacency* matrix A, where the element a_{ij} is one when there is an edge from vertex u_i to vertex s_j, and zero otherwise,

$$(A)_{ij} = \begin{cases} 1, & \text{if} \quad (u_i, s_j) \in E \\ 0, & \text{otherwise.} \end{cases} \tag{1.1}$$

The properties of bipartite networks are typically investigated by analyzing the so-called *one-mode network* or *co-occurrence network*. This is a new graph in which there is a link between two vertices of the set U if they share one or more vertices of the set S. Analogously, elements of the set S can be *"projected"* onto the set U, thus producing a new unipartite network.

1.2.1 Projected networks

The one-mode network is a weighted network, where the weight of a link is set according to a specific weighing function $l : U \times U \to \mathbb{R}$. Formally, given the bipartite graph $\mathcal{G}(U, S, E)$, the one-mode graph of U with respect to S is the weighted graph denoted by $\mathcal{P}(U, F)$, where U is the set of vertexes and F is the set of edges. Likewise, we can project the bipartite network with respect to set S, constructing the one-mode graph of S with respect to U. For

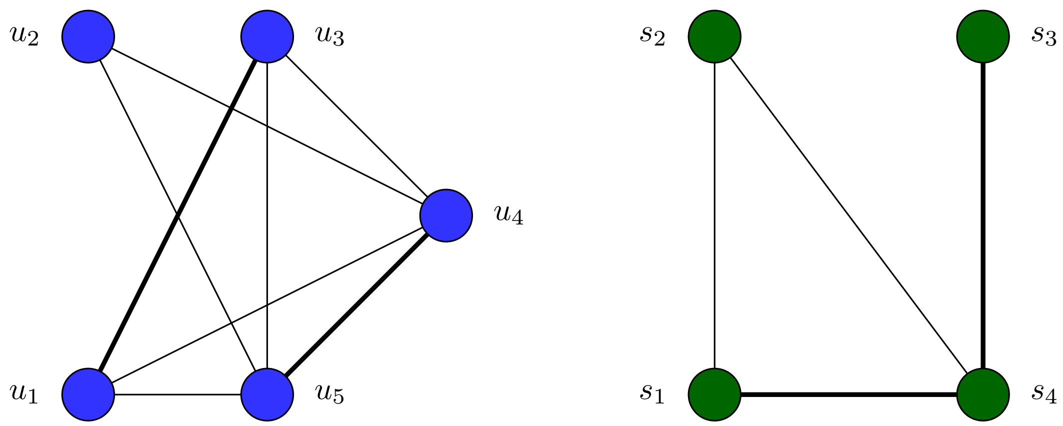


Figure 1.2: One-mode network

any $i, j = 1, 2, \dots |U|$ and $i \neq j$, a link (u_i, u_j) is set and included in F, if $l(u_i, u_j) > \xi$, where $\xi \in \mathbb{R}$.

The simplest weighing function assigns to each element of the matrix W a value corresponding to the number of co-occurrences between u_i and u_j, i.e., $l(u_i, u_j) = |N(u_i) \cap N(u_j)|$ and $\xi = 0$:

$$(W)_{ij} = \begin{cases} |N(u_i) \cap N(u_j)|, & \text{if} \quad N(u_i) \cap N(u_j) \neq \emptyset \\ 0, & \text{otherwise.} \end{cases} \tag{1.2}$$

Mappings like $l(u_i, u_j)$ are also known as *similarity functions*. Many filtering techniques use similarity functions to assign weights that become crucial in reducing the connection density of the projected network by filtering out those links that are considered not significant according to given criteria (see Section 1.2.2).

One-mode networks can be obtained through the projection of both sides of the bipartite network onto the respective sets. In Figure 1.2 we show the one-mode projections extracted from the bipartite network given in Figure 1.1 when using the co-occurrences similarity function described above, and where edges with weights higher than one are marked by a bold line. Depending on the characteristics of the original system, a projected network can also take the form of a *directed graph*, i.e. a graph where all the edges are directed from one vertex to another. For our purposes, however, it will suffice to focus only on *undirected graph*.

1.2.2 Statistically validated networks

In several real-world applications, the projected network turns out to be dense, that is, it has a high number of edges given the number of nodes n

contained in the network (the closer the number of edges to $n(n-1)/2$, the denser the network). Such a density may hide the topological properties of the system, e.g. the presence of communities and other emergent properties.

Reducing the number of edges, by keeping those which carry the essential information about the structure of the system, is therefore a crucial aspect of our analysis. Indeed, by setting a lower value of the threshold, ξ, can result to a poor representation of the important information contained in the network and the analysis of its topological properties can be misleading ([118], [122]).

We very often deal with bipartite networks that are characterized by a high level of heterogeneity in terms of vertex degree. In this respect, a validation process where co-occurrences are tested against a unique threshold will lead to filtered networks where nodes (and their respective links) are validated just because they have high degrees, and, therefore, it is likely that they display sizeable intersections with other nodes. In the insurance specific case, introduced in Chapter 3, that would mean that, for example, subjects like car repairers would be over-represented in the validated network because of their *"natural"* activity within the claim process. Conversely, nodes with lower degrees (e.g., drivers) will be excluded *a priori*, thus removing interactions which can disclose hidden anomalous behaviours.

To this purpose, we describe the co-occurrence between two nodes as a conditional event where the conditioning evidences are the degrees of both nodes and the total number of elements in the projecting set of the bipartite network. Formally, given the bipartite graph $\mathcal{G}(U, S, E)$, $\forall u_i, u_j \in U$, we define by

$$(n_{ij}|n_i, n_j, N), \tag{1.3}$$

the *conditional co-occurrence*, where $n_{ij} = l(u_i, u_j)$ is the unconditional co-occurrence, n_i and n_j are, respectively, the degree of u_i and u_j, and $N = |S|$ is the total number of nodes of the projecting set S.

Observe that, the conditional co-occurrence (1.3) has just a symbolic meaning, however, its introduction allows comparisons with the—more substantial—*conditional threshold* that is defined as follows:

$$(\xi_{ij}|n_i, n_j, N) = Q(\alpha), \tag{1.4}$$

where $Q(\alpha)$ is the right-tail α-quantile of the hypergeometric distribution,

$$Q(\alpha) = \inf\left\{q \in \mathbb{Z}_{>0} : \alpha \geq \sum_{x=q}^{\min(n_i,n_j)} \text{Hyper}(x|n_i, n_j, N)\right\}, \tag{1.5}$$

and,

$$\text{Hyper}(x|n_i, n_j, N) = \frac{\binom{n_i}{x}\binom{N-n_i}{n_j-x}}{\binom{N}{n_j}}. \tag{1.6}$$

Armed with the conditional threshold ξ_{ij}, which is inferred from the null distribution $\text{Hyper}(x|n_i, n_j, N)$, the link (u_i, u_j) is statistically significant if

$$(n_{ij}|n_i, n_j, N) \geq (\xi_{ij}|n_i, n_j, N). \tag{1.7}$$

It is worth noticing that the validation rule in (1.7) is possible because both elements are conditioned to the same set of events, which, eventually, simply turns to verifying that $n_{ij} \geq \xi_{ij}$.

Remark (1). Similarity measures that account for the marginal distributions of u_i and u_j, i.e. that explicitly make use of n_i and n_j in their formulas, are not effective in dealing with the heterogeneity of the bipartite network. For example, given the bipartite graph $\mathcal{G}(U, S, E)$ and the associated adjacency matrix $A_{M\times N}$, where $M = |U|$ and $N = |S|$, the Pearson correlation coefficient between any two binary row vectors of A, $\rho(A_i, A_j)$, is a measure of the similarity between nodes u_i and u_j, where n_{ij} is *"adjusted by"* the degree of the two nodes, n_i and n_j. The conditional co-occurrence (1.3) is explicitly given by

$$(n_{ij}|n_i, n_j, N) = \rho(A_i, A_j) = \frac{n_{ij} - \frac{n_i\, n_j}{N}}{\sqrt{n_i\, n_j \left(1 - \frac{n_i}{N}\right)\left(1 - \frac{n_j}{N}\right)}}. \tag{1.8}$$

If we consider real instances where $N \gg n_i$, n_j, for classes of nodes with almost the same vertex degree, $n_i \simeq n_j = K$, we can approximate the relation (1.8) as follows:

$$\rho(A_i, A_j) \approx \frac{n_{ij}}{K}. \tag{1.9}$$

Equation (1.9) clearly shows that if we set the threshold ξ to a high level in order to reduce the complexity of the network, we will exclude with very high probability (unless n_{ij} grows with almost the same pace of the vertex degree K) all those nodes which characterize as *hubs*, i.e. nodes with a high vertex degree K. In fraud investigation, that would imply the exclusion, *a priori*, of subjects like lawyers or car repairers. Conversely, a low level of the threshold ξ, calibrated to include node hubs of peculiar interest, it would yield a very dense and uninformative network, since even drivers sharing a single accident will be deemed as significant and included in the projected network.

Remark (2). The distribution function $\text{Hyper}(k|n_i, n_j, N)$ exactly computes the probability that k co-occurrences take place when n_j links depart from node u_j and n_i links depart from node u_i. This is easily assessed if we describe the event using an urn model where, $n = n_j$ marbles are extracted without replacement from an urn with a total of $N = |S|$ marbles, and the the urn contains $n_i = K$ marbles with a given property. In this respect, $P(X = k)$ is the probability that the sample n, drawn without replacement from the urn, shows exactly k marbles with the chosen attribute. It is worth noticing that, the Hypergeometric distribution implicitly accounts for the heterogeneity of the set U. Indeed, the probability of a given intersection depends on the marginal distribution of the set U through the vertex degree n_i and n_j.

1.2.3 Multiple hypothesis testing: family-wise error rate vs false discovery rate

The introduced Hypergeometric null hypothesis can be used to test the presence of an excess of co-occurrences between any pair of nodes u_i and u_j of either sets of a real bipartite network. Indeed, again assuming without loss of generality that nodes u_i, with degree n_i, and u_j with degree n_j belong to set U in a real bipartite network (U, S, E), and that the actual co-occurrences of these nodes in set S is $\hat{n}_{ij}$, then the probability that a value larger than or equal to $\hat{n}_{ij}$ is observed by chance, according to the null hypothesis, is

$$\text{p-value}(\hat{n}_{ij}|n_i, n_j, N) = \sum_{n_{ij}=\hat{n}_{ij}}^{\min\{n_i,n_j\}} \frac{\binom{n_i}{n_{ij}}\binom{|S|-n_i}{n_j-n_{ij}}}{\binom{|S|}{n_j}} \tag{1.10}$$

Eq. 1.10 can be used to test the excess of co-occurrences between any pair of nodes linked in the projected network, and the test fully takes into account the heterogeneity of nodes u_i and u_j, since degree n_i and n_j correspond to the actual values observed in the real bipartite network. To claim that the number of co-occurrences $\hat{n}_{ij}$ between nodes u_i and u_j is too large to be consistent with the null hypothesis of random co-occurrences, one should introduce a threshold α of statistical significance to be compared with the p-value. It could be $\alpha = 0.01$ for instance. However, the value of α does not take into account the fact that, for a given projected network, the total number of tests that one should run equals the total number of edges in the projected network, $|F|$. Therefore, α should be corrected for multiple hypothesis testing, in order to control for the family-wise error rate, that is, for errors of type I. Among the several ways to control type I errors, we consider the Bonferroni correction. The Bonferroni correction indicates that, given a univariate threshold of statistical significance, α, then the statistical threshold corrected in presence of $|F|$ tests is $\alpha_M = \frac{\alpha}{|F|}$. The Bonferroni correction is the most appropriate because it is the most restrictive and it's not affected by the fact that tests are dependent.

The Bonferroni Statistically Validated Network, or simply Bonferroni Network (BN) is obtained by filtering a given real projected network, in order to only keep links that display a statistically significant number of co-occurrences. Specifically, given a bipartite network (U, S, E), and the associated network projected on set U, (U, F), the Bonferroni network is the network that only includes all the links in the projected network such that

$$\text{p-value}(\hat{n}_{ij}|n_i, n_j, |S|) < \alpha_M = \frac{\alpha}{|F|}. \quad (1.11)$$

1.2.4 Properties of the Bonferroni SVN

The properties of the Bonferroni correction for multiple hypothesis tests and those of the hypergeometric distribution induce some interesting properties of the Bonferroni network, which are summarized in the following propositions.

Proposition 2: Given a bipartite network (U, S, E) and its projection on set U, (U, F), then the probability that one link in the Bonferroni network filtered from (U, F) is a false positive, according to the null hypothesis of random co-occurrences, is smaller than α.

Proof: let's indicate with T_0 the (unknown) number of true negative links, that is, those links that are observed by chance, because of the intrinsic heterogeneity of the bipartite network. Of course $T_0 < |F|$ by construction. Then the probability that one true negative link is included in the Bonferroni network is equal to the probability that the event $E_{ij} = p-value(\hat{n}_{ij}|n_i, n_j, |S|) < \alpha_M = \frac{\alpha}{|F|}$ occurs, where the link between u_i and u_j is a true negative. Therefore, if one lists all of the T_0 events E_{ij} as $\{E_k = p_k < \alpha_M, \forall k = 1, \dots T_0\}$ it turns out that

$$P\left(\bigcup_{k=1}^{T_0} E_k\right) = P\left(\bigcup_{k=1}^{T_0} p_k < \alpha_M\right) = P\left(\bigcup_{k=1}^{T_0} p_k \frac{\alpha}{|F|}\right) \leq$$
$$\sum_{k=1}^{T_0} P\left(p_k < \frac{\alpha}{|F|}\right) < \sum_{k=1}^{T_0} \frac{\alpha}{|F|} = \alpha \frac{T_0}{|F|} < \alpha. \quad (1.12)$$

Proposition 3: Given a bipartite network (U, S, E) and its projection on set U, (U, F), a co-occurrence $\hat{n}_{ij} = 1$ between elements u_i and u_j, with degree n_i and n_j, respectively, does not induce a link in the Bonferroni network if

$$|F| \geq \alpha|S| \quad (1.13)$$

Proof: according to the hypergeometric distribution, we have that

$$\text{p-value}(\hat{n}_{ij} = 1|n_i, n_j, |S|) > \text{p-value}(\hat{n}_{ij} = 1|1, 1, |S|) = \frac{1}{|S|}. \tag{1.14}$$

A link with co-occurrence $n_{ij} = 1$ is included in the Bonferroni network if $p - value(\hat{n}_{ij} = 1|n_i, n_j, |S|) < \frac{\alpha}{|F|}$, which, in light of Eq. 1.14, requires that

$$\frac{\alpha}{|F|} > \frac{1}{|S|} \Leftrightarrow |F| < \alpha|S|. \tag{1.15}$$

Therefore, if $|F| \geq \alpha|S|$ any link with co-occurrence $\hat{n}_{ij} = 1$ is not included in the Bonferroni network.

Chapter 2

Bipartite complex systems with a double heterogeneity: a new measure of similarity

Abstract

Complex bipartite systems are studied in Biology, Physics, Economics, and Social Sciences, and they can suitably be described as bipartite networks. The heterogeneity of elements in those systems makes it very difficult to perform a statistical analysis of similarity starting from empirical data. Though binary Pearson's correlation coefficient has proved effective to investigate the similarity structure of some real-world bipartite networks, here we show that both the usual sample covariance and correlation coefficient are affected by a bias, which is due to the aforementioned heterogeneity. Such a bias affects real bipartite systems, and, for example, we report its effects on empirical data from two bipartite systems. Therefore, we introduce weighted estimators of covariance and correlation in bipartite complex systems with a double layer of heterogeneity. The advantage provided by the weighted estimators is that they are unbiased and, therefore, better suited to investigate the similarity structure of bipartite systems with a double layer of heterogeneity. We apply the introduced estimators to two bipartite systems, one social and the other biological. Such an analysis shows that weighted estimators better reveal emergent properties of these systems than unweighted ones.

2.1 Introduction and literature review

Bipartite systems consist of two sets of elements in which elements of one set directly relate to elements of the other set only. Often these systems are described as networks. Complete information about bipartite systems can usually be incorporated in bipartite networks, however, many studies use the bipartite structure of the system only to set relationships between the elements of one of the two sets. For instance, the scientific collaboration network in [149], [150] can be seen as the projection of the bipartite system of authors and papers, where co-authored papers are only used to set a relationship between any pair of authors.

Bipartite networks and their projections are widely used to study complex systems such as mobile communication [154, 155], criminal activity [182], interbank credit markets [108, 97], investors activity [185], and recommendation systems for users and objects [127, 67]. A common feature of complex bipartite systems is heterogeneity, which typically characterizes both sides of the system and makes the statistical analysis of the various properties a challenging task. Here we focus on the heterogeneity of nodes, and, specifically, on the fact that the distribution of the number of connections of nodes from both sets, i.e. the degree, is eventually scale-free. This phenomenon is apparent in all of the systems mentioned above. For instance, in the criminal-crime bipartite system analysed in [182], there are criminals involved in more than a thousand events, while most of criminals have been found guilty of only one crime, as well as there are crimes committed by hundreds of thousands of people (like crimes against the traffic law in Sweden) and very brutal crimes, such as omicide of children, which are very rare–a few events over a decade. Such an heterogeneity of degree in the bipartite network makes it very difficult to quantify the similarity between two elements of the same set, e.g., between two criminals, in order to elicit the similarity of criminal patterns from historical data series, or between crimes, in order to investigate the association between them, and, eventually, determine the specificities they share. Another example of a system with such features is the scientific collaboration network, where there is heterogeneity of authors in terms of the number of papers they authored, and heterogeneity of papers in terms of the number of co-authors. Indeed, Newman [150] – to account for such heterogeneity in the construction of the weighted collaboration network of scientists – weighted a link between two coauthors by not just counting the number of papers in common, but weighting each one of such papers inversely according to the number of co-authors [150]. The heuristic reasoning behind such a choice is that two scientists participating in a very large collaboration are less likely to know very well each other than two scientists being the only authors of a specific paper. In systems as sparse as the collaboration network, the weight introduced by Newman can be considered as a good measure of the acquaintance between scientists, since the probability that two scientists end up authoring the same paper “by chance" is negligible. However, there are other bipartite systems where such a probability is not negligible at all. A clear example of such systems is the one of users and movies of a streaming OTT media provider, such as Netflix. Suppose that one is interested in measuring the similarity between two users based on their watching profile over a certain period of time, which is a key step to develop recommendation systems [127, 67]. The probability that two users have watched the same n movies just by chance is not negligible, and it depends on their heterogeneity, i.e., the number of movies each one of them has watched in the past. This is due to the finite number of movies available to stream, which is small if compared to number of users in the system. Such an evidence suggests that a better measure of similarity between users could be obtained by considering the difference between the number of movies two users have both watched and the expectation of such a number under an hypothesis of random selection of movies [127, 67], i.e., a sample covariance. To account for the heterogeneity of users, that is, their degree, the Pearson’s correlation coefficient might be used in place of the covariance [67, 102, 41].

However, when one is interested in covariance and correlation coefficients to estimate the connectivity between two nodes in the projected network, we show that even Newman’s solution is not sufficient to account for the double heterogeneity present in complex bipartite systems. In general, the presence of such heterogeneity of degree may induce a bias in covariance and correlation coefficient estimates, which, in turn, would make the task of discriminating information from noise in covariance/correlation matrices even more impervious [122], [158], [130].

To remove such a bias from covariance and correlation coefficients we introduce weighted estimators that take into account, at once, the heterogeneity on both sides of a bipartite network. Moreover, we also quantify the improvement of the new estimators compared to unweighted ones and demonstrate the power of the introduced methodology with applications to two real social and biological datasets. From a conceptual point of view, the newly proposed estimators are such that the covariance/correlation between any two given elements in the system depends on all the others, in such a way that adding or removing even a single element influences the value of the estimator. To prove the stability of the weighted estimators against such a change in the system, we ran a robustness analysis and show that the proposed estimators are rather robust to changes in the system composition up to 30%.

The paper is structured in the following way. Section 2.2.1 discusses the problem of a bias in the sample covariance and correlation of bipartite systems and in Section 2.2.2 we propose a model of the rewiring process which demonstrates that the expected value of the covariance is different from zero. In Section 2.2.3 we define the new weighted covariance estimator in the multivariate case and show that its expected value is indeed null. In Section 2.2.4 we focus on the weighted correlation coefficient and show the improvement it offers over the unweighted one. Section 2.2.6 introduces the methodology used to estimate the parameters of the underlying model for the heterogeneity of the bipartite system. Section 2.3 displays the results of employing the weighted against the unweighted estimators in two empirical datasets.

2.2 Methods

2.2.1 Sample covariance and correlation in bipartite systems

In bipartite networks elements can be divided in two disjoint, independent sets, such that only links between the two sets are allowed, see Fig. 2.1.

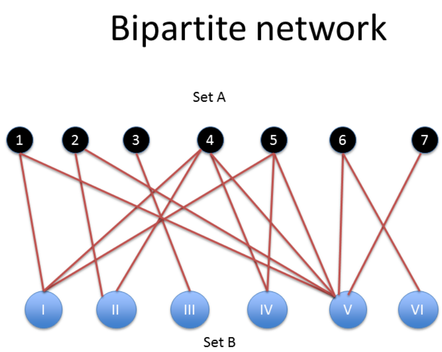


Figure 2.1: Schematic representation of a bipartite network with N nodes in set A (black), e. g., authors, and T nodes in set B (blue), e. g., papers. Links are only possible between the two sets and are shown in red. A projected network of nodes in set A is obtained by linking any two nodes in A that share one or more connections to nodes in set B of the bipartite network.

In the previous section, we discussed the importance of evaluating—within many applications—the similarity between two nodes, say i and j, which belong to one set of a bipartite system, according to their connections to elements of the other set. Such a similarity measure should have specific properties, typically depending on the nature of the applications. However, one desirable feature, which most of the similarity measures share, is that the similarity should suitably take into account the heterogeneity of nodes i and j, i.e., their degree. This is attained in different ways: for instance according to

Jaccard [110], this is done by taking the number of connections that i and j share, n_{ij}[1], divided by the total number of elements in the second set that are connected to i and j, that is, $K_i + K_j - n_{ij}$[2], where K_i (K_j) is the degree of node i (j). Another possibility is to consider the difference between the number n_{ij} and the expected value of n_{ij}, $E(n_{n_{ij}})$, according to a simple urn model. Here it is assumed that node i and node j independently and randomly select K_i, and K_j nodes, respectively, from the second set, the urn with T labeled marbles, without restitution. According to such a simple model, n_{ij} follows the Hypergeometric distribution (see for instance [183]), and therefore $E(n_{ij}) = K_i\,K_j/T$. In summary, the similarity between node i and j can be evaluated as $n_{ij} - K_i\,K_j/T$, and the method to attain this result is pretty similar to the one that brought Newman and Girvan to introduce and operationalize the contribution to "modularity" [151] of a community of nodes as the difference between number of links observed in that community and the expected number of links in the same community under an hypothesis of random connectivity that preserves the degree of each node. Therefore, typically, measures of similarity, such as those described above, make use of the observed value of n_{ij} and rescale and/or shift it according to a model in which the degree of each node is assumed as a constraint, or, in other words, as a conditioning quantity. Similarity $n_{ij} - K_i\,K_j/T$ can be interpreted, apart from a scaling constant, as a sample covariance, as discussed in the next paragraph, and it explicitly and suitably takes into account the heterogeneity of degree of the set of nodes i and j belong to, through the quantities K_i and K_j. However, such a measure totally disregards the heterogeneity of nodes belonging to the second set, and, as shown below, this absence of consideration determines a bias in the similarity.

Let's suppose we measure the sample covariance between two elements i and j in set A of a bipartite system, as the scalar product between the binary vectors $\mathbf{v_i}$ and $\mathbf{v_j}$. A component $v_{i,h}$ ($v_{j,h}$), with $h \in [1, ..., T]$, of vector $\mathbf{v_i}$ ($\mathbf{v_j}$) is equal to 1 if element i (j) is linked to node h in set B, and 0 otherwise. Therefore, the sample covariance estimator between two binary vectors can be written as [67]:

$$\hat{\text{cov}}(i,j) = \frac{1}{T}\left(\mathbf{v_i}\cdot\mathbf{v_j}\right) - \frac{1}{T^2}\left(\sum_{h=1}^{T} v_{i,h}\right)\left(\sum_{h=1}^{T} v_{j,h}\right) = \frac{1}{T}\left(\hat{n}_{ij} - \frac{K_i\,K_j}{T}\right), \quad (2.1)$$

the hat is henceforth used to denote an estimator. In Eq.(2.1) $\hat{n}_{ij}$ is the

[1] n_{ij} is the size of the intersection between the sets of first-neighbors of nodes i and j.
[2] $K_i + K_j - n_{ij}$ is the size of the union of the sets of first-neighbors of nodes i and j.

observed number of links in common between the pair of elements i and j, of degree $K_i = \sum_{h=1}^{T} v_{i,h}$ and $K_j = \sum_{h=1}^{T} v_{j,h}$. Degrees are parameters which are kept fixed throughout. For example, looking at Fig. 2.1, we have for the pair of nodes 4 and 5 in set A, of degree, respectively, $K_4 = 4$ and $K_5 = 3$, binary vectors $\mathbf{v_4} = \{1, 1, 0, 1, 1, 0\}$ and $\mathbf{v_5} = \{1, 0, 0, 1, 1, 0\}$, number of common links $n_{45} = 3$, a covariance of $\hat{\text{cov}}_{45} = \frac{1}{6}(3 - 2) = 1/6$.

From Eq.(2.1), the sample correlation coefficient estimator between two binary vectors becomes:

$$\hat{\rho}_{ij} = \frac{\hat{\text{cov}}(i,j)}{\hat{\sigma}_i\, \hat{\sigma}_j} = \frac{\hat{n}_{ij} - \frac{K_i K_j}{T}}{\sqrt{K_i \left(1 - \frac{K_i}{T}\right) K_j \left(1 - \frac{K_j}{T}\right)}}, \tag{2.2}$$

where $\hat{\sigma}_i$ and $\hat{\sigma}_j$ are standard deviation estimators of vector $\mathbf{v_i}$ and $\mathbf{v_j}$,

$$\hat{\sigma}_i = \sqrt{\frac{K_i}{T}\left(1 - \frac{K_i}{T}\right)}, \hat{\sigma}_j = \sqrt{\frac{K_j}{T}\left(1 - \frac{K_j}{T}\right)}. \tag{2.3}$$

An evaluation of the accuracy of an estimator, the covariance and correlation coefficient in the present case, represents a crucial aspect to assess the performance of the estimator itself. However, evaluating the accuracy of an estimator requires that the true value of the estimated quantity is known. In this study, the heterogeneity of both sets of nodes in the bipartite system is a feature that shall be considered in the assessment of estimators' accuracy, as heterogeneity represents a key feature of most real world (bipartite) complex systems. As far as we know, there is no way to simulate a bipartite network with a double heterogeneity and controlled connectivity of nodes. Therefore we started from real data describing a bipartite network, with both layers of heterogeneity, and performed a random rewiring of the network, in such a way to destroy any association between nodes' connectivity [46]. In this way the expected covariance between two nodes connectivity patterns is zero. Basically, one step in the rewiring procedure consists of randomly sampling a pair of links in the bipartite network, involving two nodes on each side, and a swap of the target nodes of the link in set B, if the latter newly formed links are not already present in the system. For example, from Fig. 2.1, one randomly selects the pair of links $4 - II$ and $6 - IV$ and swaps the target nodes in set B to obtain two new links $4 - IV$ and $6 - II$, since neither 4 nor 6 were already linked, respectively, to IV and II. To randomize the network, one needs to perform a great number of swaps, stopping when the overlapping between the original and rewired networks, evaluated with an appropriate measure, stabilizes around a minimum value (see Section 2.2.6 for details). However, when considering a randomly rewired bipartite network, we note that resulting covariance and correlation matrices still display a residual structure as detailed in section 7.1. The residual structure still present in matrices appears to depend on the degree distributions of both sets of nodes, that is, on the intrinsic double heterogeneity of the system. Thus, the sample covariance and correlation estimators reported in Eq. 2.1 and 2.2, respectively, appear to be biased in such systems, and the bias won't be uniform. Such a bias is evaluated and interpreted through a biased urn model in the next section.

2.2.2 Expected value of the covariance and correlation under a biased urn model: the Wallenius' non-central hypergeometric distribution

Here, we propose a model which approximately describes the statistical properties of the outcome of a random rewiring procedure. The model we propose is a simplification of the problem which, nonetheless, allows us to exactly preserve the degree distribution on one side of the bipartite network, and to keep the degree distribution on average on the other side. The underlying idea is to model the random rewiring as a sampling from a biased urn, followed by a sampling from an unbiased urn, both without replacement (to preserve degrees).

Our aim is to show the origin of the bias in the covariance and correlation coefficient in Eqs. (2.1) and (2.2) of the randomized network, by calculating their expected values and showing that they are different from zero.

To show the presence of a bias we describe a simplified situation, where nodes in set B only have either a high degree, which we'll formalize as a heavy weight w_2, or a low degree w_1 (a "light" weight). If we now look at how random links form between a node i in set A and a number K_i of nodes in set B, such a process can be modeled as a sampling of exactly K_i marbles (node's i degree), from the total of T marbles in set B. The crucial hypothesis is that we assume that marbles have two different probabilities of being selected. Specifically, m marbles have a probability to be sampled proportional to weight w_2 (heavy), whereas the remaining $T-m$ marbles have a probability to be sampled proportional to w_1 (light), and we define the weight ratio as $w = w_2/w_1 > 1$. The weight models the heterogeneity in set B. We'll focus on Eq.(2.1), and show that the expected value of $\mathrm{cov}(i,j)$ is, in general, different from zero, if $w > 1$.

In this model, each node i in set A samples a total of K_i marbles, of which k_i^w are heavy and the remaining $K_i - k_i^w$ are light. In a biased urn problem without replacement, a single variable w is sufficient to describe the system, with the stochastic variable $k_i^w \in [\max(0, K_i - T + m), \min(K_i, m)]$ following the Wallenius non-central hypergeometric distribution [190].

If all marbles are distinguishable, for example labeled, we now ask ourselves what would be the intersection n_{ij} between the marbles sampled by two different nodes, i and j, in A. The expected number of sampled objects $\mathbf{E}[n_{ij}|k_i^w, k_j^w]$ in common between i and j will be the sum of the expected number of heavy marbles in common, n_{ij}^w, and the expected number of light ones in common, n_{ij}^1,

$$\mathbf{E}[n_{ij}|k_i^w, k_j^w] = \mathbf{E}[n_{ij}^w|k_i^w, k_j^w] + \mathbf{E}[n_{ij}^1|k_i^w, k_j^w]. \tag{2.4}$$

The underlying probability distribution, since each weight-group is now homogeneous, is the Hypergeometric distribution. Specifically, the probability that both nodes sampled exactly n_{ij}^w heavy marbles in common, out of the m available ones, is given by $P(n_{ij}^w; k_i^w, k_j^w, m)$. Similarly, the corresponding probability for the n_{ij}^1 light marbles in common is $P(n_{ij}^1; K_i - k_i^w, K_j - k_j^w, T - m)$. Since the sampling processes are independent, variables n_{ij}^w and n_{ij}^1 are independent as well, so that the joint probability distribution is just the product of the previous two. The expected numbers of common heavy and light marbles can be easily calculated,

$$\mathbf{E}[n_{ij}^w|k_i^w, k_j^w] = \frac{k_i^w\, k_j^w}{m} \quad \text{and} \quad \mathbf{E}[n_{ij}^1|k_i^w, k_j^w] = \frac{(K_i - k_i^w)(K_j - k_j^w)}{T - m}, \tag{2.5}$$

thus the expected number of marbles in common between i and j turns out to be:

$$\mathbf{E}[n_{ij}] = \sum_{k_i^w, k_j^w} \left(\mathbf{E}[n_{ij}^w|k_i^w, k_j^w] + \mathbf{E}[n_{ij}^1|k_i^w, k_j^w]\right) W(k_i^w)\, W(k_j^w) = \frac{\mu_i\, \mu_j}{m} + \frac{(K_i - \mu_i)(K_j - \mu_j)}{T - m}, \tag{2.6}$$

where μ_i (μ_j) is the expected value of k_i^w (k_j^w) calculated with the Wallenius distribution PMF $W(k_i^w)$ ($W(k_j^w)$).

Unfortunately, no exact formula for the mean of the Wallenius distribution is known [190], however, the approximate solution of the following equation is reasonably accurate [131]:

$$\frac{\mu_i}{m} + \left(1 - \frac{K_i - \mu_i}{T - m}\right)^w = 1. \tag{2.7}$$

Finally, by calculating the Taylor series up to second order of $\mathbf{E}[n_{ij}]$ in Eq.(2.6) near $w = 1$ and due to the linearity of operator expectation $\mathbf{E}$, the expected value of the covariance can be approximated by:

$$\begin{aligned}\mathbf{E}[\mathrm{cov}(i,j)] &= \frac{\mathbf{E}[n_{ij}]}{T} - \frac{K_i K_j}{T^2} \simeq \\ &\simeq \frac{m(T-m)}{T^2}[(1-\frac{K_i}{T})\ln(1-\frac{K_i}{T})][(1-\frac{K_j}{T})\ln(1-\frac{K_j}{T})](w-1)^2\end{aligned} \tag{2.8}$$

For a graphical representation of the dependency of $\mathbf{E}[\mathrm{cov}(i,j)]$ on K_i, K_j see Fig.2.2.

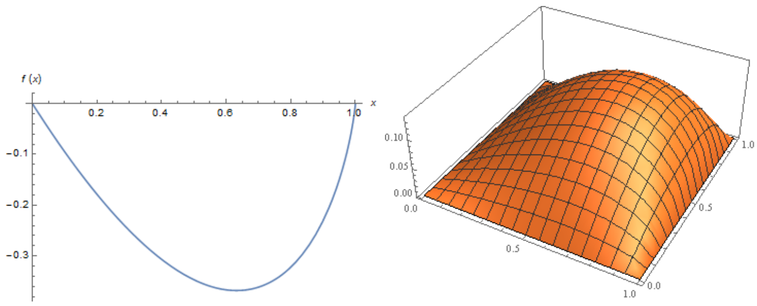


Figure 2.2: Left panel: plot of $f(x) = (1-x)\ln(1-x)$ for $x \in [0,1]$, the function is strictly negative and displays a minimum in $x_m = 1 - 1/e \simeq 0.632$. Right panel: 3D plot of $f(x,y) = (1-x)\ln(1-x) \cdot (1-y)\ln(1-y)$ for $x, y \in [0,1]$, the function is strictly positive and shows a maximum in $\{x_M, y_M\} = \{1-1/e, 1-1/e\}$.

The expected value of the correlation coefficient in Eq.(2.2) can be calculated from Eq.(2.8) dividing by the standard deviations, which depend only on fixed parameters:

$$\mathbf{E}[\rho_{ij}] \simeq \frac{m(T-m)}{T\sqrt{K_i\left(1-\frac{K_i}{T}\right)K_j\left(1-\frac{K_j}{T}\right)}}\left(1-\frac{K_i}{T}\right)\ln\left(1-\frac{K_i}{T}\right)\left(1-\frac{K_j}{T}\right)\ln\left(1-\frac{K_j}{T}\right)(w-1)^2. \tag{2.9}$$

From Eq.(2.8) and Eq.(2.9) it's easy to see how the expected value of both the covariance and the correlation coefficient depends on i's and j's degrees, K_i and K_j, as well as on w, which is the ratio of w_2 to w_1 (here representing the heterogeneity of the other set, B, in the bipartite system). Thus, we've shown there exists a bias due to the interplay between both sets' heterogeneity in a bipartite system. In the next section, we introduce estimators of covariance and correlation coefficient, whose expected value is zero in any randomly rewired network, that is, they are bias free.

2.2.3 Multivariate weighted covariance estimator

In the most general case, we're dealing with $n < T$ groups, each containing $\mathbf{m} = \{m_1, m_2, ..., m_n\}$ marbles of weight $\mathbf{w} = \{w_1, w_2, ..., w_n\}$. Each node i samples k_i^q marbles out of group q, for a total of marbles equal to its own degree K_i. Our aim here is to show that the bias in the expected value of the covariance can be completely removed by opportunely weighing the original binary vectors. Thus, re-normalizing the vectors leads to the definition of a new covariance estimator, $\hat{\text{cov}}(i,j)^{\mathbf{w}}$, which possesses the desirable property that its expected value is zero.

Specifically, focusing on node i, a component q of vector $\mathbf{v}_{\mathbf{i}}^{\mathbf{w}}$ is now set equal to $1/f(w_q, K_i)$ if i randomly sampled a marble out of group q and 0 otherwise. We can then reorder each user's weighted vector $\mathbf{v}_{\mathbf{i}}^{\mathbf{w}}$ as follows:

$$\mathbf{v}_{\mathbf{i}}^{\mathbf{w}} = \left\{ \frac{\delta_1}{f(w_1, K_i)}, ..., \frac{\delta_{m_1}}{f(w_1, K_i)}, \frac{\delta_{m_1+1}}{f(w_2, K_i)}, ..., \frac{\delta_{m_1+m_2}}{f(w_2, K_i)}, ..., \frac{\delta_{T-m_n+1}}{f(w_n, K_i)}, ..., \frac{\delta_T}{f(w_n, K_i)} \right\},$$

where each δ_s is either 1 or 0, and the following constraints hold,

$$\sum_{s=1}^{m_1} \delta_s = k_i^1, \cdots, \sum_{s=T-m_n+1}^{T} \delta_s = k_i^n; \quad \sum_{s=1}^{T} \delta_s = \sum_{q=1}^{n} k_i^q = K_i; \quad \sum_{q=1}^{n} m_q = T.$$

Having thus re-normilized the original vectors by the weight functions $f(w_q, K_i)$, we can now define the weighted covariance estimator as:

$$\hat{\text{cov}}(i,j)^{\mathbf{w}} = \frac{1}{T} \sum_{q=1}^{n} \frac{\hat{n}_{ij}^q}{f(w_q, K_i) f(w_q, K_j)} - \frac{1}{T^2} \left(\sum_{q=1}^{n} \frac{k_i^q}{f(w_q, K_i)} \right) \left(\sum_{q=1}^{n} \frac{k_j^q}{f(w_q, K_j)} \right), \tag{2.10}$$

where $\hat{n}_{ij}^q$ is the number of marbles of weight w_q in common between i and j.

Working under the multivariate version of the biased urn model introduced in Section 2.2.2, we're now in the position to calculate the expected value of the weighted covariance. Under the Hypergeometric distribution hypothesis, see Eq.(2.6) we have that,

$$\mathbf{E}[n_{ij}^q | k_i^1, ... k_i^n, k_j^1, ... k_j^n] = \frac{k_i^q \, k_j^q}{m_q}, \tag{2.11}$$

so that the expected value of the weighted covariance in Eq.(2.10) can be written as:

$$\mathbf{E}[\text{cov}(i,j)^{\mathbf{w}}] = \frac{1}{T} \sum_{q=1}^{n} \left[\frac{\mathbf{E}[k_i^q]}{f(w_q, K_i)} \left(\frac{\mathbf{E}[k_j^q]}{m_q \, f(w_q, K_j)} - \frac{1}{T} \sum_{p=1}^{n} \frac{\mathbf{E}[k_j^p]}{f(w_p, K_j)} \right) \right] \tag{2.12}$$

From Eq.(2.12), we can define the group of weight functions $\{f(w_1, K_j), ..., f(w_n, K_j)\}$ as those which zero the expected value of the weighted covariance, that is, the solutions of the following system of equations:

$$
\begin{aligned}
&\frac{\mathbf{E}[k_j^1]}{m_1\, f(w_1, K_j)} - \frac{1}{T}\sum_{p=1}^{n} \frac{\mathbf{E}[k_j^p]}{f(w_p, K_j)} = 0\\
&\frac{\mathbf{E}[k_j^2]}{m_2\, f(w_2, K_j)} - \frac{1}{T}\sum_{p=1}^{n} \frac{\mathbf{E}[k_j^p]}{f(w_p, K_j)} = 0\\
&\qquad\vdots\\
&\frac{\mathbf{E}[k_j^n]}{m_n\, f(w_n, K_j)} - \frac{1}{T}\sum_{p=1}^{n} \frac{\mathbf{E}[k_j^p]}{f(w_p, K_j)} = 0.
\end{aligned}
\tag{2.13}
$$

System (2.13) is indeterminate and can be solved after assigning an arbitrary value to one of the weight functions, for example $f(w_1, K_j)$. Then all the other weight functions can be written relative to $f(w_1, K_j)$:

$$
\frac{f(w_q, K_j)}{f(w_1, K_j)} = \frac{m_1}{m_q}\,\frac{\mathbf{E}[k_j^q]}{\mathbf{E}[k_j^1]}, \qquad \text{with } q \in [2, n]. \tag{2.14}
$$

Thus, by defining the weight functions $\{f(w_1, k_j), ..., f(w_n, k_j)\}$ with Eq.(2.14), it's guaranteed that the expected value of the weighted covariance estimator in Eq.(2.10) is zero.

In the multivariate case, the Wallenius distribution PDF for the vector of variables $\mathbf{k_j} = \{k_j^1, k_j^2, ..., k_j^n\}$, with weight vector $\mathbf{w} = \{w_1, w_2, ..., w_n\}$ and number of marbles per weight group $\mathbf{m} = \{m_1, m_2, ..., m_n\}$, takes the form:

$$
W(\mathbf{k_j}; \mathbf{m}, \mathbf{w}) = \prod_{q=1}^{n} \binom{m_q}{k_j^q} \int_0^1 \prod_{q=1}^{n} (1 - t^{w_q/D})^{k_j^q}\, dt, \tag{2.15}
$$

where $D = \mathbf{w} \cdot (\mathbf{m} - \mathbf{k_j}) = \sum_{q=1}^{n} w_q(m_q - k_j^q)$. The group means $\mu_q = \mathbf{E}[k_j^q]$ with $q \in [1, n]$ satisfy the system of equations [40]:

$$
\left(1 - \frac{\mu_1}{m_1}\right)^{1/w_1} = \left(1 - \frac{\mu_2}{m_2}\right)^{1/w_2} = ... = \left(1 - \frac{\mu_n}{m_n}\right)^{1/w_n}, \tag{2.16}
$$

with the constraint $\sum_{q=1}^{n} \mu_q = K_j$. From this constraint and Eq.(2.14), we can write each group mean μ_q in terms of the weight functions,

$$
\frac{\mu_q}{m_q} = \frac{K_j\, f(w_q, K_j)}{\sum_{p=1}^{n} m_p\, f(w_p, K_j)}, \tag{2.17}
$$

and inserting Eq.(2.17) in Eq.(2.16), we find a set of equations for the weight functions:

$$
\left(1 - \frac{k_j\, f(w_1, k_j)}{\sum_{p=1}^{n} m_p\, f(w_p, k_j)}\right)^{1/w_1} = ... = \left(1 - \frac{k_j\, f(w_n, k_j)}{\sum_{p=1}^{n} m_p\, f(w_p, k_j)}\right)^{1/w_n}. \tag{2.18}
$$

System (2.18) provides a way to directly calculate the weight functions, without having to compute the group means first.

2.2.4 Multivariate weighted correlation estimator

In this section, we write down the weighted estimator for the correlation coefficient and quantitatively show the improvement it offers over the unweighted one.

From Eq.(2.12) it's straightforward to define the weighted correlation coefficient estimator as the Pearson correlation coefficient of the weighted vectors:

$$\hat{\rho}_{ij}^{\mathbf{w}} = \frac{\hat{\mathrm{cov}}(i,j)^{\mathbf{w}}}{\hat{\sigma}_i^{\mathbf{w}}\ \hat{\sigma}_j^{\mathbf{w}}} = \frac{\sum_{q=1}^{n} \frac{n_{ij}^q}{f(w_q,K_i)f(w_q,K_j)} - \frac{1}{T}\left(\sum_{q=1}^{n} \frac{k_i^q}{f(w_q,K_i)}\right)\left(\sum_{q=1}^{n} \frac{k_j^q}{f(w_q,K_j)}\right)}{\sqrt{\left[\sum_{q=1}^{n} \frac{k_i^q}{f(w_q,K_i)^2} - \frac{1}{T}\left(\sum_{q=1}^{n} \frac{k_i^q}{f(w_q,K_i)}\right)^2\right]\left[\sum_{q=1}^{n} \frac{k_j^q}{f(w_q,K_j)^2} - \frac{1}{T}\left(\sum_{q=1}^{n} \frac{k_j^q}{f(w_q,K_j)}\right)^2\right]}}. \tag{2.19}$$

Unfortunately, from Eq.(2.19) one realizes immediately that having $\mathbf{E}[\mathrm{cov}(i,j)^{\mathbf{w}}] = 0$ is not a sufficient condition for $\mathbf{E}[\rho_{ij}^{\mathbf{w}}] = 0$, since variables $\{\mathbf{k_i}, \mathbf{k_j}\}$ now appear in the denominator as well. However, we can approximate $\mathbf{E}[\rho_{ij}^{\mathbf{w}}]$ by its Taylor series near $\mathbf{w} = \mathbf{1}$ and show that its value is less than the Taylor series of $\mathbf{E}[\rho_{ij}]$.

2.2.5 Comparison of correlation coefficients near w=1

We now proceed to show the improvement of the weighted estimator over the unweighted one, by comparing the Taylor series of their expected values. Out of simplicity, we show our results in the bivariate case, with $n = 2$ groups and $w = w_2/w_1$. The Taylor series of $\mathbf{E}[\rho_{ij}]$ near $w = 1$ was calculated in Section 2.2.2, Eq.(2.9).

We now calculate the Taylor series of $\mathbf{E}[\rho_{ij}^w]$, starting from the expected value of ρ_{ij}^w given k_i^w, k_j^w, which can be calculated from Eq.(2.19) when $n = 2$:

$$\mathbf{E}[\rho_{ij}^w | k_i^w, k_j^w] = \frac{\left[(T-m)\,k_i^w - m\,f(w,K_i)(K_i - k_i^w)\right]}{m\,T\,\sigma_i^w f(w,K_i)} \frac{\left[(T-m)\,k_j^w - m\,f(w,K_j)(K_j - k_j^w)\right]}{(T-m)\,T\,\sigma_j^w f(w,K_j)}. \tag{2.20}$$

From Eq.(2.20), remembering that the Wallenius distribution in $w = 1$ becomes the Hypergeometric distribution, we can calculate the zero order term in the Taylor series, which turns out to be null. To calculate the first and second order terms, we define the function:

$$F(k_i^w, k_j^w, w) = \mathbf{E}[\rho_{ij}^w | k_i^w, k_j^w] \cdot W(k_i^w) \cdot W(k_j^w),$$

which, summed over all possible values of $\{k_i^w, k_j^w\}$ gives $\mathbf{E}[\rho_{ij}^w]$. Thus, we can calculate the derivatives as follows,

$$\left.\frac{d\mathbf{E}[\rho_{ij}^w]}{dw}\right|_{w=1} = \sum_{k_i^w, k_j^w} \left[\frac{d}{dw}\mathbf{E}[\rho_{ij}^w | k_i^w, k_j^w] W(k_i^w) W(k_j^w)\right]_{w=1} = \sum_{k_i^w, k_j^w} \left.\frac{dF(k_i^w, k_j^w, w)}{dw}\right|_{w=1}, \tag{2.21}$$

by exploiting the advantage of first evaluating the derivatives of $F(x_i, x_j, w)$ near $w = 1$, and then summing over the variables. The first non-null term is the second order one, so that the expected value of the weighted correlation coefficient near $w = 1$ is:

$$\mathbf{E}[\rho^w_{ij}] \simeq \frac{m(T-m)}{T\sqrt{K_i(1-\frac{K_i}{T})K_j(1-\frac{K_j}{T})}}(1-\frac{K_i}{T})[h_{(T)} - h_{(T-K_i)} + (1-\frac{1}{K_i})\ln(1-\frac{K_i}{T})]\cdot$$
$$\cdot(1-\frac{K_j}{T})[h_{(T)} - h_{(T-K_j)} + (1-\frac{1}{K_j})\ln(1-\frac{K_j}{T})](w-1)^2, \tag{2.22}$$

where $h_{(n)} = \sum_{k=1}^{n} 1/k$ is the n-th harmonic number, that is, the sum of the reciprocals of the first n natural numbers.

A graphic comparison between the unweighted estimator in Eq.(2.9) and the weighted estimator in Eq.(2.22) is shown in Fig 2.3, where the improvement of the latter is clear.

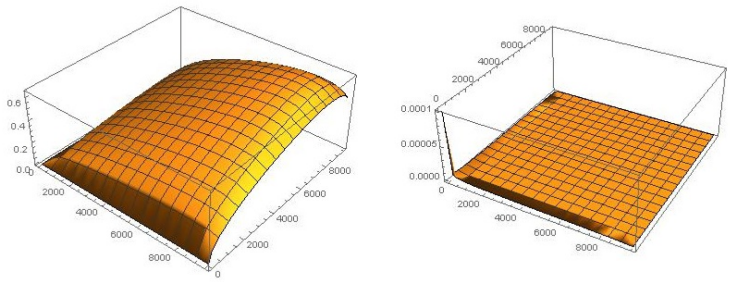


Figure 2.3: Plot of the expected value of the unweighted correlation coefficient (left) against the weighted one (right) as a function of k_i and k_j. Parameters are: $T = 10^4 = 2m$, where m is the number of marbles in either group, according to the bivariate biased urn model, $w = \frac{w_2}{w_1} = 2$, while k_i and k_j can vary between 1 and 95% of the number of marbles in the urn (T), that is, we let k_i and k_j to span a range large enough to describe sparse, as well as dense networks. Both correlation estimates assume the same value of 0.0001 when $k_i = k_j = 1$. Notice that the vertical scales are different in the left and right plots.

Finally, to quantify the improvement offered by the weighted estimator over the unweighted one, we use the asymptotic expansion of the harmonic number,

$$h_{(T)} - h_{(T-K_i)} \simeq -\ln\left(1-\frac{K_i}{T}\right) - \frac{1}{2T}\left(\frac{K_i/T}{1-K_i/T}\right), \tag{2.23}$$

valid when $T \to \infty$ and $T >> K_i$.

Within the former asymptotic limit, we have that the ratio of the expected value of the weighted correlation coefficient to the unweighted one, near $w = 1$, is

$$\frac{\mathbf{E}[\rho^w_{ij}]}{\mathbf{E}[\rho_{ij}]} = \left[\frac{h_{(T)} - h_{(T-K_i)}}{\ln\left(1-\frac{K_i}{T}\right)} + 1 - \frac{1}{K_i}\right]\left[\frac{h_{(T)} - h_{(T-K_j)}}{\ln\left(1-\frac{K_j}{T}\right)} + 1 - \frac{1}{K_j}\right] \simeq$$
$$\simeq \left(\frac{1}{K_i} - \frac{1}{2T}\right)\left(\frac{1}{K_j} - \frac{1}{2T}\right) \simeq \frac{1}{K_iK_j}. \tag{2.24}$$

Thus, when $T >> K_i, K_j$, which occurs, for instance, when the bipartite network is sparse, we find that the expected value of the weighted correlation estimator is $1/K_iK_j$ times the expected value of the unweighted one.

2.2.6 Wallenius' distribution: weight-groups and odds-ratio estimation

In the previous section, unbiased weighted estimators for the covariance and correlation coefficient have been introduced, which can be calculated by modifying the original 0/1 incidence matrix on the basis of the degree distributions of both sets nodes in the bipartite network. That is done, in practice, by dividing the 1's of the matrix by the weight function $f(w_q, k_j)$ if user j has drawn a marble belonging to weight-group q.
Now, since $f(w_q, k_j)$ depends on both the expected number of marbles (according to a Wallenius' experiment) drawn by a user with degree k_j and the weight w_q, a problem of estimation arises. In fact, once we collect the data, the composition of the "urn" (marble set) must be characterized, that is, the number and dimension of groups $\mathbf{m}$ and the weights must be estimated.

The only information we have about the marbles is given by their degree, that is the number of users they are linked to. So, on the basis of that, we need to put together marbles which are as similar as possible. The most intuitive and easy choice would be to assume that the odds-ratios $\mathbf{w}$ are exactly equal to the degree of set B in the bipartite system. For example, in a bipartite system of parliament members and private initiatives (see next section for details), the weight of an initiative could be set equal to the number of members who signed it. Such a rough estimate has the benefit of automatically defining the weight-groups vector $\mathbf{m}$, by grouping together all the initiatives which have the same weight, with the simple idea of just dividing the original vectors $\mathbf{v_i}$ ($\mathbf{v_j}$) by the weight $\mathbf{w}$ defined by set B's heterogeneity, as inspired by Newman [150], which shall henceforth be referred to as Newman's estimator. Basically, Newman's estimator may work well when one is dealing with datasets with low heterogeneity, so that the noise can be modeled as a multinomial distribution, but it becomes dramatically biased as heterogeneity on both sides of the system grows, as is typically the case in many complex systems. In truth, the estimation of the odds-ratios in a Wallenius distribution with different sampling processes, that is, a different number of total marbles sampled by each user, is not straightforward and has not been investigated in the literature.

A very simple and effective method in this case is given by the K-Means algorithm, which, starting with some initial centers values, iteratively assigns each marble to the closest mean, until no marble is moved any more [95]. The problem about the K-Means algorithm is its deterministic nature, indeed the number of clusters to find must be given a priori by the researcher. However, it turns out that the classification performed by K-Means corresponds with the one performed by the maximum likelihood approach assuming that data come from a Gaussian Mixture Model (GMM), with clusters distributed normally with same variances. Via an Expectation Maximization (EM) algorithm it is possible to maximize the likelihood of the mixture model and compute the usual BIC statistics, which allows one to find the optimal number of weight-groups [75]. Once the number of weight-groups and their dimension are available, it's quite straightforward to estimate the odds-ratios parameter vector $\mathbf{w}$ of the Wallenius distribution, according to Eq.(2.16), as:

$$w_q^i = \frac{\ln\left(1 - k_q^i/m_q\right)}{\ln\left(1 - k_n^i/m_n\right)}. \tag{2.25}$$

The estimation of groups can be performed by using the function *WGroupsEst*, while the function *WeightsEst* is used to estimate the odds-ratios given the groups (both functions are available in the R package WestC, which is available upon request to the authors). From Eq.(2.25) it's possible to reconstruct each weight by averaging over all the users and keeping in mind that, in a multivariate Wallenius distribution, the odds-ratios are distributed according to a log-normal:

$$\langle w_q \rangle = \exp\left(\left\langle \ln\left(w_q^i\right)\right\rangle_i\right) \tag{2.26}$$

The odds-ratios estimates obtained from Eq.(2.26) get more and more accurate as the number of users and marbles grows. Obviously, when going from Eq.(2.25) to Eq.(2.26), one needs first to remove all the values of w_q^i that are either 0, 1 or infinite.

2.3 Empirical Analysis

In this section, we employ the weighted covariance and correlation estimators we developed, against the unweighted ones, with the aim of showing how the new estimators outperform the others in 1) revealing no community structure in randomly rewired networks and 2) highlighting community structure in two real networks. As a matter of fact, in order to calculate the weighted covariance and correlation, we simply derive the weight functions as shown in section 2.2.3 and use them to weigh users' vectors, over which we then compute the covariance and correlation coefficients. The first step will be identifying the weight-groups and estimating their corresponding odds-ratios.

2.3.1 Data

The datasets taken into consideration are two, one pertains to the social sciences and the other one to the biological sciences. The social database [159] consists of 1,808 private initiatives submitted between 2011 and 2014 by 201 members of the Finnish parliament, along with information on who signed each initiative. Data cover an entire parliament of the duration of four years. The resulting bipartite system displays Members of Parliament (MP) on one side and initiatives they signed on the other. Info on MP include their party and district of election. Parties in Finland are:Kristillisdemokraatit - Christian Democrats (KD), Keskusta - Centre Party (KESK), Kokoomus - National Coalition Party (KOK), Perussuomalaiset - Finns Party (PS), Ruotsalainen kansanpuolue - Swedish People's Party (RKP), Sosialidemokraattinen puolue - Social Democratic Party (SDP), Vasemmistoliitto - Left Alliance (VAS) and Vihreä liitto - Green League (VIHR). Electoral districts are 15.

The biological data comes from the Clusters of Ortholous Group (COG) database[3], which stands for Clusters of Orthologous Groups of proteins, from the sequenced genomes of prokaryotes and unicellular eukaryotes. The database consists of 4,873 COGs present in 66 genomes of unicellular organisms, belonging to 3 broad macro-groups: Archaea, Bacteria or Eukaryota. The corresponding bipartite system consists of organisms on one side and COGs present in their genome on the other. Organisms belong to 12 different phyla: Actinobacteria (Act), Archaea of type Crenarchaeota (ArC) and Euryarchaeota (ArE), Cyanobacteria (Cya), Eukariota (Euk), Gram-negative Proteobacteria of type α (Gr-a), β (Gr-b), ϵ (Gr-e), γ (Gr-g), Gram-positive bacteria (Gr+), Hyperthermophilic bacteria (HyT) and other bacteria (Oth). This database has been widely studied, see for example [178] and [179].

Table 2.1 shows that both datasets present a high degree of heterogeneity in both sides of the bipartite system, which is at the origin of the bias observed with usual sample correlation and covariance estimators. However, such a high degree of heterogeneity is frequently found in bipartite systems.

[3] Available at http://www.ncbi.nlm.nih.gov/COG

Data

	Finnish parliament	*COG*
T	1,808	4,873
$w_m - w_M$	2-150	3-66
N	201	66
$K_m - K_M$	2-793	362-2,243
n_L	28,568	83,675

Table 2.1: T is the number of initiatives/COGs; $w_m - w_M$ is their heterogeneity, that is, the range (min-max) of degree distributions; N is the number of MP/organisms; $K_m - K_M$ is the range (min-max) of their degree distributions; n_L is the number of links in the bipartite network.

2.4 Empirical evidence about the performance of the weighted estimators

2.4.1 Real and randomly rewired bipartite networks: a comparison of estimators

If we want to assess how the heterogeneity of nodes affects the correlation matrix computed according to Eq.(2.2), one of the approaches used in the literature [46] is the rewiring of the bipartite network, since it keeps constant the degree of each node, and generates a network where the expected correlation between two nodes, based on their connectivity patterns, is zero. The rewiring algorithm samples randomly a pair of MP/organisms according to a probability distribution equal to their degree distribution, then samples randomly two initiatives/COGs out of those already linked to the first sampled pair, again according to the degree distribution of initiatives/COGs. Then, if neither in the pair is already linked to the other's sampled initiative/COG, the two links are swapped, otherwise the swap is rejected. Such an algorithm performs a random rewiring of the entire bipartite system, preserving both sides degree distributions. To efficiently rewire large bipartite networks a Monte Carlo procedure known as the switching-algorithm (SA) [109] can be used. This algorithm can be performed by using the function *Rewiring* of our R package.

We can now compare the weighted estimators against the unweighted ones, over both datasets. The first result, as shown in Fig. 2.4, is that the weighted covariance estimator completely destroys the structure still present in the unweighted covariance matrix of the rewired network. This feature translates also to the weighted/unweighted correlation coefficients in Fig. 2.5, although the expected value of the weighted correlation estimator is only approximately zero. In Fig. 2.5, we show how the weighed correlation outperforms the unweighted correlation in randomly rewired networks. Indeed, according to Fig.2.5, the weighted correlation does not indicate the presence of any structure

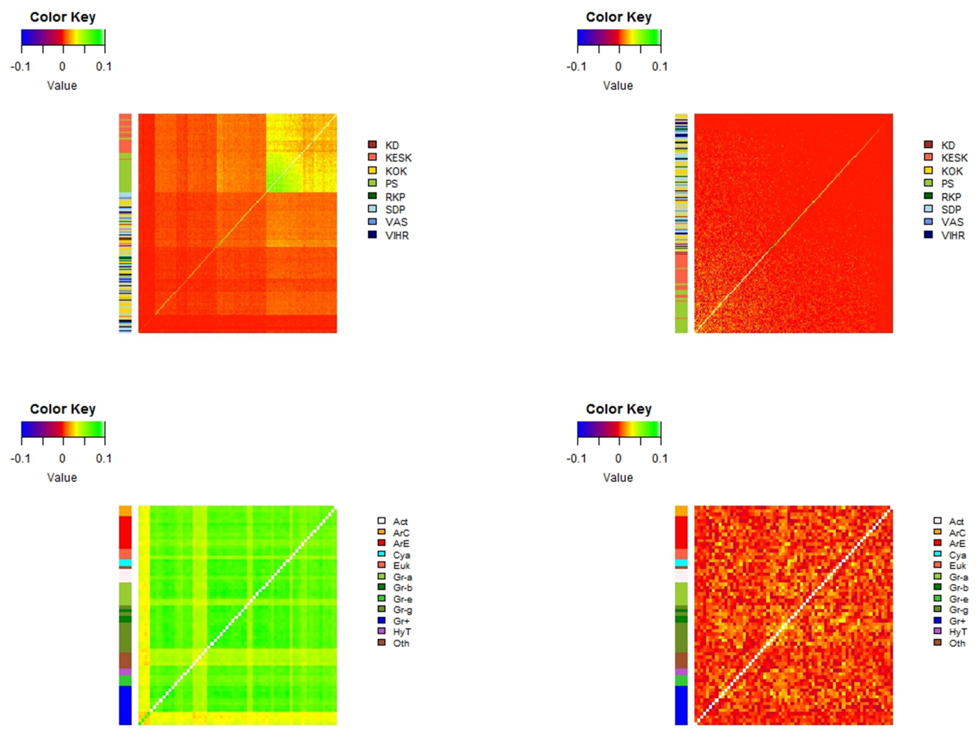


Figure 2.4: Covariance matrices of MP (top-row) and organisms (bottom-row) after random rewiring of the original bipartite network, calculated without weighing the vectors (left) and weighing them (right). MP/organisms are ordered by increasing degree with respect to columns and by decreasing degree with respect to rows. The Color Key scale is identical in all figures.

in the system, whereas the unweighted one does. Furthermore, Fig.2.6 shows that the weighed correlation better highlights the cluster-structure present in the real system. Indeed, the weighted correlation matrix better identifies the clusters in the original COGs bipartite system (bottom row), by encompassing a broader scale of values, displayed within the matrix in violet (negative correlations), zero (red), orange (low), yellow (average) and green (high) against the unweighted matrix which only features the positive correlations, making it harder to distinguish sub-clusters. Indeed the right weighted matrix shows sub-clustering corresponding to organisms' phyla. For example, it neatly discriminates Archaea (red and orange in the left color-bar), Eukariota (Salmon) and Bacteria (all the rest), by also grouping together Gram-negative bacteria (shades of green), Gram-positive bacteria (blue), Hyperthermophilic bacteria (violet), Actinobacteria (pink) and Cyanobacteria (cyan).

Concerning the Finnish parliament dataset (term 2011-2014), results reported in top-row panels of Fig. 2.6 show how the weighing destroys the cluster of party KESK, implying that this cluster is more due to the heterogeneity and consequent bias in the unweighted correlation estimator than to

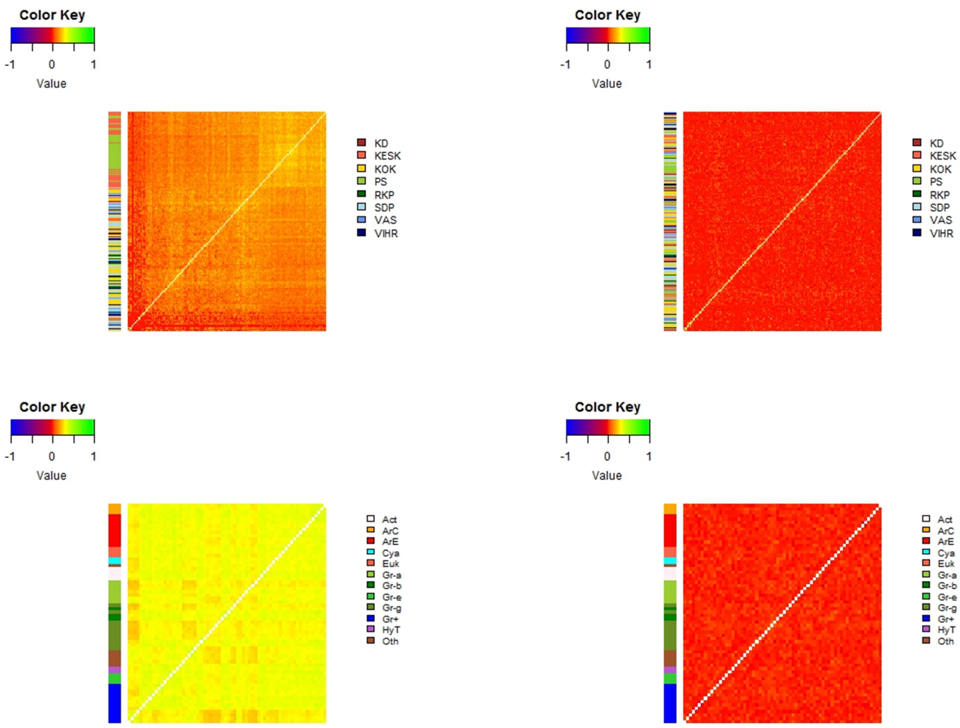


Figure 2.5: Correlation matrices of MP (top-row) and organisms (bottom-row) after random rewiring of the original bipartite network, calculated without weighing the vectors (left) and weighing them (right). MP/organisms are ordered by increasing degree with respect to columns and by decreasing degree with respect to rows. The Color Key scale is identical in all figures.

a real collaboration between MP, while, at the same time, weighing preserves the cluster of party PS. This finding is in agreement with the general trend observed in [159], where the evolution of this network over 4 Finnish parliament terms is studied. In fact, during previous terms, MP collaborated by district and by party both, with party being more characterizing in the opposition and district sub-clustering within the government. If we look at the unweighted matrix, it appears that not only the two opposition parties strongly cluster and display a negative correlation with each other, but also the government splits in two right-wing left-wing sub-clusters. Such a change from the previous terms was attributed to the sudden rise in numbers of the populist party PS. From the weighted matrix instead we can see that the situation is more in line with previous terms, with district subclustering reappearing.

2.4.2 Weight-groups and Odds-ratios Estimation

In this subsection our proposed estimation method will be applied to a simulation study as well as to the real datasets discussed before to show the improvement it brings over the unweighted and Newman covariance/correlation

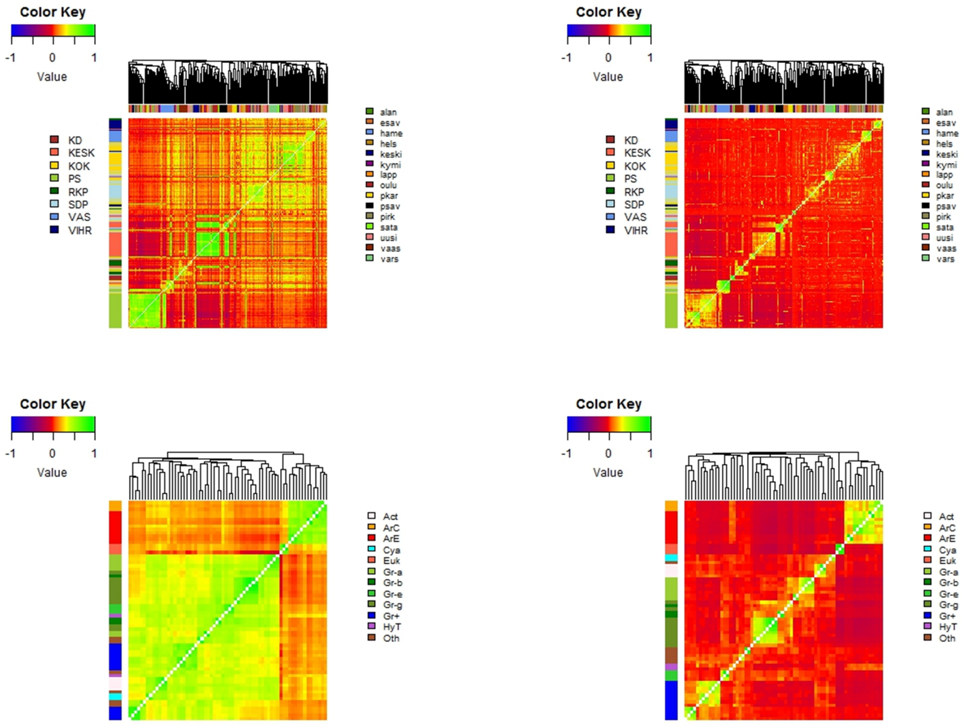


Figure 2.6: Unweighted (left) against weighted (right) correlation matrices of MP (top) and organisms (bottom), ordered by hierarchical clustering with average linkage performed on each matrix [10]. The left-side bar is colored according to party (left legend) or phylum (right legend), the top bar is colored according to districts (right legend). Diagonals have been colored white. The Color Key scale is identical in all figures.

estimates. The setting of the simulation is as follows: we define set A heterogeneity, by fixing $\mathbf{v_i}$'s degree for every i, we consider five groups of marbles of equal size, and set the odds-ratios as $\mathbf{w} = \{15, 10, 5, 2.7, 1\}$, since all the weights can be normalized in terms of any of the other weights, in this case normalizing with respect to the lightest weight-group. We ran an exploratory simulation with $\mathbf{m} = \{500, 500, 500, 500, 500\}$, encompassing the whole spectrum of values of K_i, from 10 to 1990 in steps of 30 for a total of 83 users. With these initial parameters, the simulation runs a random sampling from a biased urn with odds-ratios $\mathbf{w}$, one user at a time. Then, all of the marbles sampled by each user are labeled randomly from 1 to the total of 2,500 marbles, so that the corresponding user's profile binary vector can be constructed. Finally, the incidence matrix is built from all the profile vectors, after taking care of having removed any marble labels which were never sampled by any user (which usually doesn't happen if the number of users is not too low and their heterogeneity is not too poor).

Having thus constructed our synthetic database, we can easily calculate Newman's covariance and correlation estimators by simply dividing every row

of the matrix by its corresponding weight, which is just the number of users who sampled it, and then computing the unweighted estimators on the resulting matrix.
For what concerns our newly proposed weighted estimators, in order to calculate the weight functions $f(w_h, K_i)$ one needs to estimate both the weight-groups $\mathbf{m}$ and the odds-ratios $\mathbf{w}$ from the synthetic dataset. In Fig. 2.7 we report the results of the exploratory simulation, by showing the plot with the estimated partition of marbles, the BIC curve with points starting from two clusters (so that BIC_{min}=19,717.7; therefore 5 is the optimal number of groups to choose), the plot of both covariance and correlation estimators calculated with Newman's weight and with our weighted estimators as a function of users' degree: K_iK_j/T^2, $\forall i, j > i$.

From the simulations we ran, it's quite clear that the weighted estimators perform better than Newman's ones in terms of accuracy (Fig. 2.7). In fact, the latter ones are still affected by a bias growing as user's degree increases. In Fig. 2.8, we compare the estimators in terms of their precision. The results indicate that precision of all the three estimators is comparable in spite of the degree. In conclusion, the weighted estimator turns out to be more accurate than the other estimators, especially when high values of degree are considered, and all the estimators show a similar precision. The performed analysis suggests that, while there are many other ways in which one can attempt to identify the weight-groups in empirical datasets when they are unknown a priori, our approach, which is quite simple, works well enough to provide estimates of the parameters that allow the introduced weighted estimators of covariance and correlation to outperform the other considered estimators.

In Fig. 2.9 and 2.10 we show the above described method to identify groups and relative odds-ratios for the rewired matrices of the Finnish parliament and COGs databases. The parameters we obtained from the algorithm are summarized in Table 2.2.

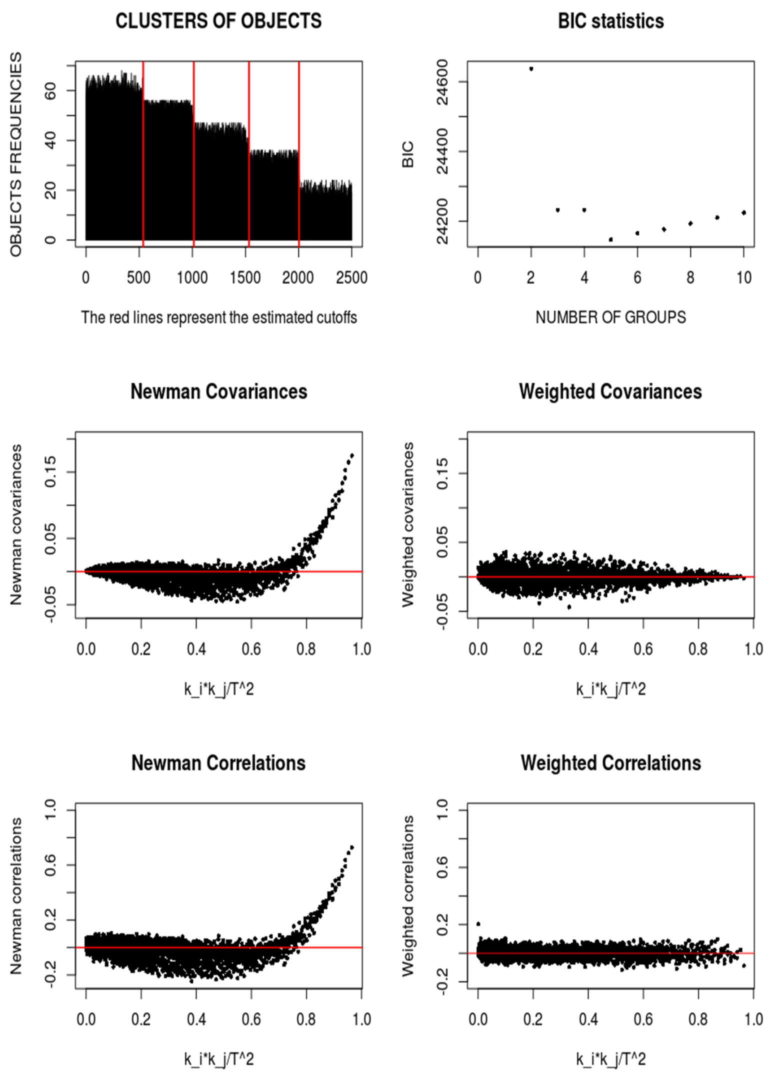


Figure 2.7: Exploratory simulation, top row shows the estimation process of the number and dimension of groups, mid row shows the plot of Newman's covariance (left) and weighted covariance (right) as a function of $K_i K_j / T^2$ and the bottom row shows the same plot of Newman's correlation and weighted one.

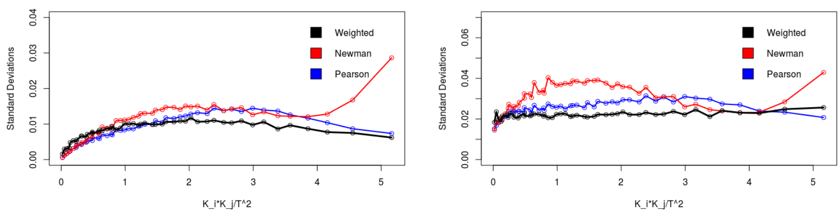


Figure 2.8: Standard deviations of covariances (left) and correlations (right) for the Pearson, Newman and weighted estimators. Standard deviations are calculated over non overlapping moving windows of the support $(k_i\, k_j / T)$, each one including 500 points.

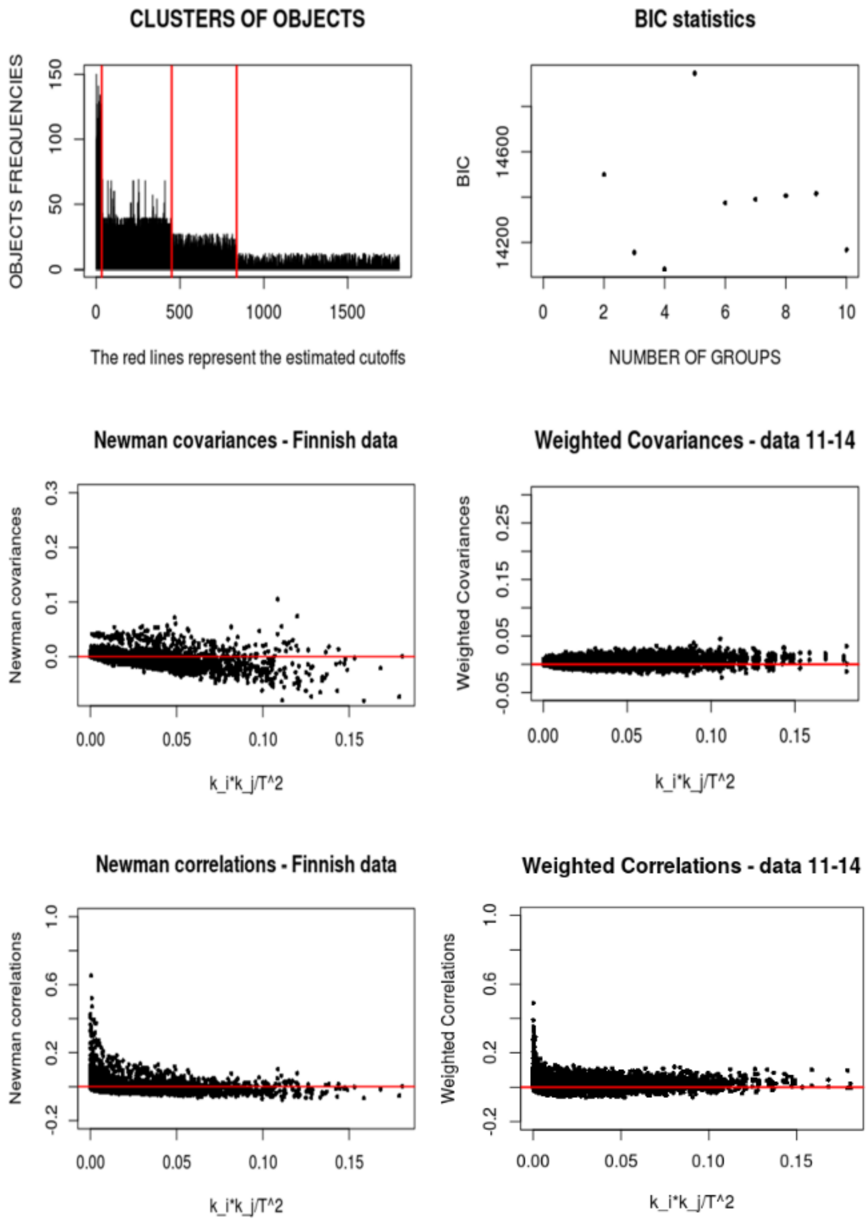


Figure 2.9: Finnish parliament rewired data, top row shows the groups estimation process, mid row shows the plot of Newman's covariance (left) and weighted covariance (right) as a function of $K_i K_j / T^2$ and the bottom row shows the same plot of Newman's correlation and weighted one.

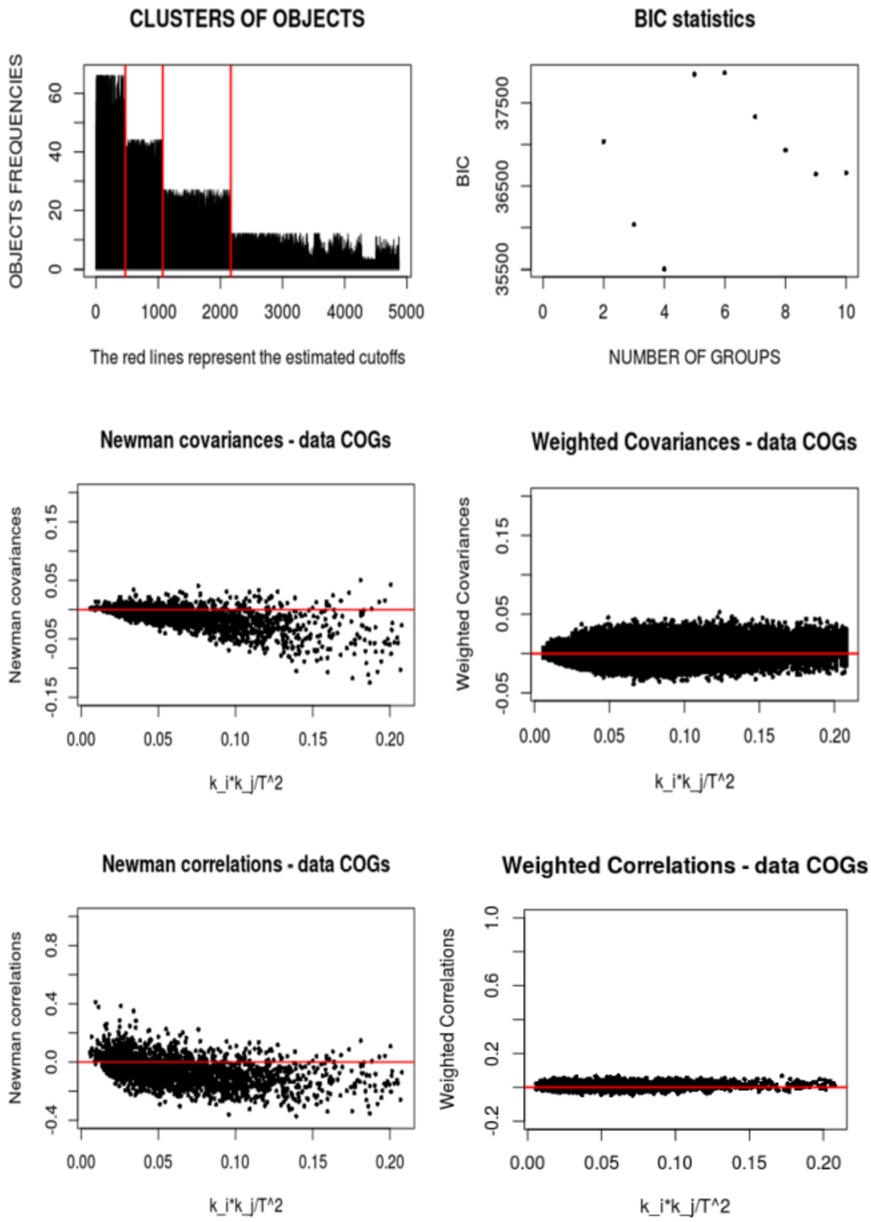


Figure 2.10: COGs rewired data, top row shows the groups estimation process, mid row shows the plot of Newman's covariance (left) and weighted covariance (right) as a function of K_iK_j/T^2 and the bottom row shows the same plot of Newman's correlation and weighted one.

Parameters from the algorithm

Exploratory simulation					
$N.groups$	5				
BIC	19,717.7				
$\hat{\mathbf{m}}$	537	476	520	470	497
$\hat{\mathbf{w}}$	12.4	8.4	4.6	2.5	1
Finnish Parliament 11-14 data					
$N.groups$	4				
BIC	14,082.9				
$\hat{\mathbf{m}}$	33	417	388	970	
$\hat{\mathbf{w}}$	38.12	5.98	2.21	1	
COGs data					
$N.groups$	4				
BIC	35,502.8				
$\hat{\mathbf{m}}$	470	603	1094	2706	
$\hat{\mathbf{w}}$	28.98	10.95	4.16	1	

Table 2.2: Parameters obtained by running the algorithms implemented by the R package WestC. The algorithm first estimates the number of groups via GMM likelihood approach and then calculates the best partition according to the k-means algorithm, from which the weight-groups vector $\mathbf{m}$ is obtained (this can be performed by the function *WGroupsEst*), while the corresponding odds-ratios vector $\mathbf{w}$ is calculated according to Eq.2.26 (function *WeightsEst*). The estimates are sorted according to a decreasing weight, with the lighter fixed to 1.

2.4.3 Unbiased weighted estimators in a community detection framework

We have also compared the proposed estimators as applied to a more complicated, yet controlled, synthetic system. Specifically, we have considered the actual marginals observed in the Finnish parliament dataset, i.e., the degree (number of signers) of initiatives and the degree (number of signed initiatives) of parliament members, in such a way to be assured that a double heterogeneity is included in the model. We have then randomly sorted out parliament members in three non overlapping groups, G_1 and G_2 including 60 MP each, and G_3 with the remaining 81 MP. Each one of the 1808 initiatives has been randomly labeled according to four categories, in order to mimic, in the simulation, the presence of first signers, i.e., proposers, and the group(s) they belong to. Specifically, 482 initiatives have been assumed to be proposed by a member of group G_1 and labeled P_1, 514 initiatives proposed by a member of G_2 and labeled P_2, 542 proposed by a member of G_3 and labeled P_3, and, finally, 270 initiatives proposed by one member of G_2 and one member of G_3 and labeled P_4. Then the simulation consisted in randomly selecting, independently for each initiative, the list of signers in the following way. For each initiative m with label P_i and degree k, k MP have been randomly selected, without restitution, from the list of the 201 MP with probability proportional to the degree of MP times a weighting factor only depending on the label P_i of the initiative, that is, the group(s) the proposer belongs to. Specifically, if $i = 1, 2$, or 3 then the degree of members of the group(s) G_i (i=1,..,3) has been multiplied by a factor w_i, whereas the degree of the other MP remained the same, and, if $i = 4$, then the degree of members of both G_2 and G_3 has been multiplied by a factor w_4. Weights used in the simulation are $w_1 = 5$, $w_2 = 2$, $w_3 = 2$, and $w_4 = 3$. Weights w_1, w_2, and w_3 are used to increase the probability that MP belonging to the same group co-sign initiatives proposed by a member of their group, while weight w_4 plays a double role: on the one hand it increases the probability of intra-group co-signing for groups G_2 and G_3, on the other hand it introduces a mixing factor between these groups, since it also increases the probability that a member of G_2 and a member of G_3 co-sign the same initiative. According to the way in which simulation has been performed, empirical values of the degree of initiatives are exactly preserved in the synthetic realization, whereas the empirical degree of each MP is preserved only on average, that is, the expected value of the degree of each MP in the simulation corresponds to the one empirically observed. At least to our knowledge, the expected value of connectivity covariance or correlation between any two MP is unknown for this model.

Once a simulated network has been obtained, we prove here that the information carried by the introduced weighted estimator turns out to be useful when performing community detection, for instance, by applying deterministic algorithms, such as the k-means, but also methods based on generative model estimation, such as the Stochastic Block Model (SBM) [105].
With respect to a large majority of community detection techniques, SBM has the advantage of explicitly stating the underlying assumptions of the model, which improves the interpretability of results. Since the introduction of the SBM [105], a lot of improvements have been subsequently made to basic SBM scheme, in order to make it more versatile by increasing the number of model parameters. Prominent examples are the degree-corrected SBM [117], which takes into account the heterogeneity of vertex degrees within the same communities, the biSBM for analyzing bipartite networks [123], and the hierarchical SBM (hSBM) [156] to overcome the so-called "resolution limit" problem of community size, that is, the fact that well-defined small clusters were not detectable when dealing with very large networks. In general, for the SBM model specification, the number of groups can be given independently, otherwise users are required to resort to heuristics, or more complicated inference approaches based on the computation of the model evidence, which are not only numerically expensive, but can only be done under onerous approximations.

There is a subtle difference between SBM and the estimation of similarity patterns between nodes of a network. On the one hand, the main objective of SBMs estimation is addressing community detection problems. Its estimation is performed through the inference of parameters of a given specification of the model, obtaining values of parameters as the ones that best explain the observed network (Maximum likelihood). On the other hand, the method proposed in this paper is not based on the estimation of parameters of a generative model, but rather, on the opportune modification of the original incidence matrix. This can be easily done by estimating the strategic weight functions $f(w, k)$ that allow the purification of the covariance/correlation matrix from the presence of the spurious correlations due to the heterogeneity of both sets of a bipartite network. From an operative point of view this approach is similar to the Newman's one in that both act directly on the binary vectors of the original incidence matrix. The weighted covariance/correlation estimators turn out to be a good instrument to highlight similarity patterns between the objects of a bipartite network, similarity patterns that eventually are useful in a community detection framework.

Therefore, we first performed the Louvain's clustering algorithm [27], which is based on the maximization of the weighted modularity function, to estimate the optimal number of communities in the projection of the synthetic bipartite network discussed above. In particular, we applied it to three different weighted projected networks, in order to make a direct comparison between the clustering algorithm performances depending on the kind of weights considered in the projected network. Specifically, links of the projection of our synthetic network were weighted according to Pearson's, Newman's, and our weighted correlation coefficients. Since weights have to be positive, the sequence $w' = (w - w_{min})/(w_{max} - w_{min})$ was considered to allow weights to vary within the interval [0, 1]. While the optimal number of groups detected using the network with weights according to Pearson is two, and the optimal one using the network with weights according to Newman's approach is four—thus underestimating and overestimating the number of groups, respectively—the network weighted according to our weights leads the algorithm to correctly uncover the three groups of objects. With respect to other clustering algorithms we used, the k-means algorithm with 3 groups proved to have the best class predictive power. Therefore, here we report the results obtained by using the k-means algorithm with three groups to compare the three weighting methods when used as classifiers. The confusion matrix associated with each estimator has been calculated, as well as the corresponding multivariate Matthews Correlation Coefficient (MCC) [83], which has been used as an overall measure of performance of the classifiers. The confusion matrices obtained for each correlation estimator are:

$$C(\text{biased urn}) = \begin{pmatrix} 55 & 5 & 0 \\ 0 & 54 & 6 \\ 0 & 45 & 36 \end{pmatrix} ; C(\text{Newman}) = \begin{pmatrix} 55 & 4 & 1 \\ 2 & 29 & 29 \\ 2 & 20 & 59 \end{pmatrix} ; C(\text{Pearson}) = \begin{pmatrix} 56 & 4 & 0 \\ 0 & 31 & 29 \\ 2 & 29 & 52 \end{pmatrix},$$

where, each row corresponds to the original classification of MP in the synthetic network and each column to the classification elicited from the simulated network. The matrices show that all of the estimators easily allow to separate MP belonging to group G_1 from the others, while distinguishing between groups G_2 and G_3 is more difficult due to the mixing weight w_4 used in the simulation. The three class Matthews correlation coefficients associated with the confusion matrices above are $MCC(biasedurn) = 0.63$, $MCC(Newman) = 0.56$, $MCC(Pearson) = 0.53$.

We also wanted to investigate the possibility that our weighting method might prove useful in the SBM framework. Therefore the degree-corrected hierarchical SBM (DC-hSBM) was applied to our synthetic network, in the following two settings:

1. the unweighted bipartite network, represented by the original 0/1 incidence matrix;

2. the weighted bipartite network, where links are weighted according to the components of vector $\mathbf{v_i^w}$ (functions of $f(w_j, K_s)$), which depend on both the degree of subject s and the weight-group of marble j.

By maximizing the models' posterior distribution, it is possible to estimate the optimal number of groups of objects, given the graph and the other parameters of the model.

In case (i), the upper three hierarchical levels of the estimated DC-hSBM highlighted respectively 5, 2 and 1 clusters, meaning that, according to DC-hSBM, the number of estimated groups of MP closest to the one used to generate the synthetic network was two. On the contrary, when case (ii) is considered, the hierarchical levels of the model unveiled respectively 16, 3 and 1 clusters, suggesting how the introduction of our weights helps the model to reveal the true underlying structural properties of the analysed bipartite network, that is, 3 groups of MP. To further improve the classification provided by DC-hSBM as applied to case (ii), which corresponds to a value of MCC equal to 0.47, we used the optimal number of groups revealed by DC-hSBM, i.e. 3 groups, as a prior information for the estimation of the degree-corrected bipartite SBM [123], leading to a very high level of accuracy in the prediction of membership of MP. Indeed, the confusion matrix of the classification for the DC-biSBM is:

$$C\big[\text{biSBM(3 groups)}\big] = \begin{pmatrix} 60 & 0 & 0 \\ 0 & 53 & 7 \\ 1 & 7 & 73 \end{pmatrix},$$

The Matthews correlation coefficient associated with this confusion matrix is 0.91, that is far higher than the ones obtained using the k-means clustering algorithm. Although we are aware that this is just a preliminary analysis, it suggests that the biased urn model might be usefully integrated with SBM. However, an in depth analysis of that is out of the scope of the present paper and is left for future work.

2.4.4 Robustness analysis

Since the proposed weighted estimator depends on the heterogeneity of both sets of elements in a bipartite network, if we sample a subset of elements from the group of interest (MP/organisms), then the degree of elements on the other set (initiatives/COGs) decreases as well and, as a result, the weighted correlations may change for the sampled elements in the set of interest. In other words, the correlation coefficient between two elements would potentially depend on the composition of the subset, and therefore a robustness analysis is in order, to show how the weighted estimator holds up when subsetting data. We ran 1,000 independent random samplings of 90%, 80% and 70% MP/ organisms from the randomly rewired network, and calculated the Frobenius distance between (i) pairs of weighted correlation matrices (by considering only elements included in both samplings), (ii) weighted correlation matrices and the identity matrix (which corresponds to the noiseless null-model) and (iii) unweighted correlation matrices and the identity matrix [106]. In order to compare matrices of different dimensions, we renormalized each distance by $\sqrt{n(n-1)}$, where n is the size of the pair of matrices over which the distance is calculated.

According to Fig. 2.11, the variability of the distribution of distances increases as the percentage of sampled elements decreases, while their expected value remains the same.

The distribution of the Frobenius distances between the weighted correlation matrices and the identity matrix is the first one from the left in each panel, while the the distribution of the Frobenius distances between the unweighted correlation matrices and the identity matrix is at right side of each panel. Furthermore, the distribution of distances between weighted correlation matrices is always in between the other two distributions. These results indicate a larger accuracy of the weighted estimator.

2.5 Conclusions

Elements' heterogeneity is a common feature of many real-world bipartite systems, and we have provided evidence of biasing in the binary covariance and correlation estimators when applied to bipartite systems with a high degree of heterogeneity on both sides. Such a bias becomes apparent when looking at the correlation and covariance matrices of a randomly rewired network, which is supposed to be completely randomized, whereas both the unweighted

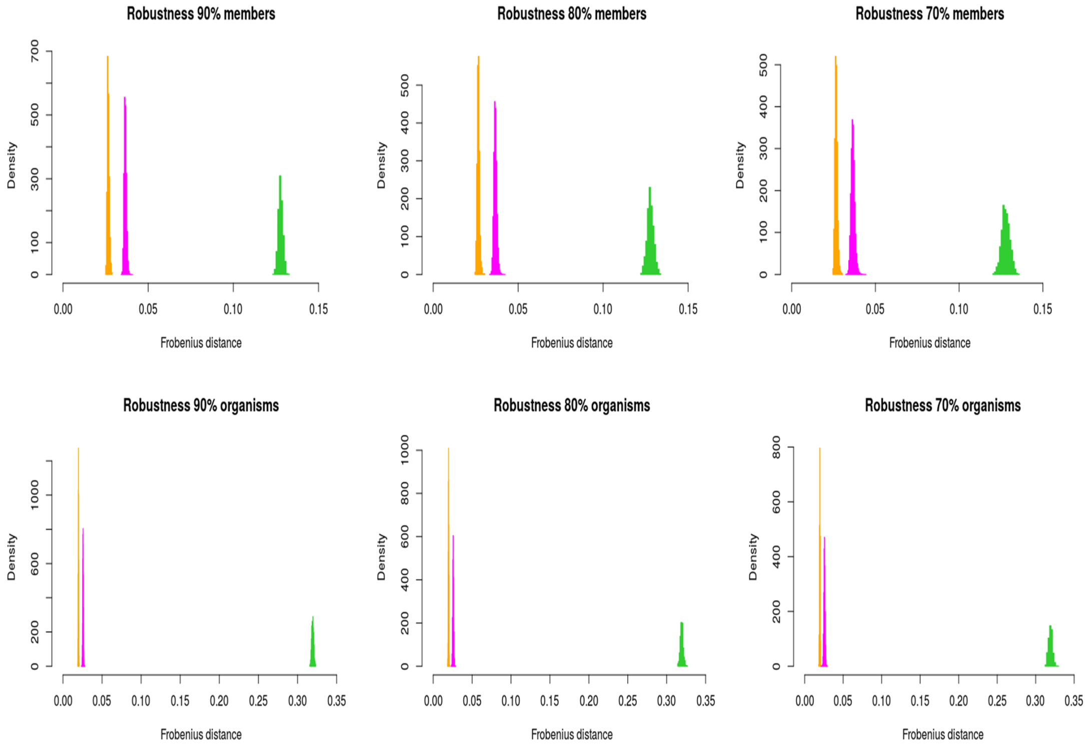


Figure 2.11: Robustness analysis performed on the weighted correlation coefficient between MP (top) and between organisms (bottom) in the rewired network. We display in violet the distribution of Frobenius distances between weighted correlation matrices, in yellow the distribution of weighted-Identity distances, in green the distribution of unweighted-Identity distances.

correlation and covariance matrices turn out to be structured instead.

To explain the former structure and devise an unbiased estimator, we developed a simple theoretical model of the rewiring process, as a sampling without replacement from a biased urn. Such a model is an approximation of the randomly rewired network, in the sense that the degrees of the set we are projecting on is exactly preserved in the model, like in the randomly rewired network, while the degrees of the other set of nodes is only preserved on average, while it is exactly preserved in the randomly rewired network. According to the biased urn model, two users randomly and independently pick a number of marbles equal to their degree, the underlying distribution being, therefore, the Wallenius non-central hypergeometric distribution. One can then calculate the expected value of random co-occurrence within each weight-category, that is the number of marbles with the same label randomly sampled by two users, by using the standard hypergeometric distribution. The model predicts a second order correction to the expected value of the unweighted sample covariance, which depends on both users degree and quadratically on the weight, when $w \simeq 1$.

The starting point to construct the unbiased estimator lies on the idea of including weighs in the binary vectors, in order to remove the bias. Weights are chosen in such a way as to satisfy the requirement of zeroing the expected value of the covariance in the purely random case. By doing so, we automatically end up with a new estimator of covariance whose expectation value is zero under random rewiring, thus being unbiased. By using the same weighting functions used to estimate the covariance, the expected value of the correlation keeps showing a second order bias in w. However, such a bias is much smaller than the one in the unweighted estimator: it is $1/(K_i K_j)$ times the unweighted one, where K_i and K_j are the degrees of the considered users. Furthermore, from a more practical point of view, we've shown that such an improvement in the correlation estimator de facto zeroes the expected value of the correlation coefficient under rewiring as well, at least for a broad range of users' degrees, in both real-world examples analysed in the paper.

Finally, the introduced covariance and correlation estimators perform better than the unweighted ones at grasping the clustered structure of the real bipartite networks considered in the paper. Specifically, they better capture aggregation by phyla in the COGs dataset and better discriminate between real and noise-induced clusters of members of the Parliament in the Finnish dataset of initiatives.

We have also assessed how similarity patterns described by the proposed weighted correlation coefficients can be very helpful in a community detection framework. We proved it in the specific case where the observed bipartite network presented a hierarchical cluster structure and double heterogeneity.

Of course, we rely on the fact that the improvement brought by our methodology can have a positive impact in other real situations as well - for example - referring to the machine learning algorithms for online recommendation which currently uses the simple unweighted correlation coefficients to find patterns of similarity in the data.

In conclusion, our paper serves both as a warning to other researchers when using binary correlation and covariance to investigate bipartite systems with a high heterogeneity on both sides, and as a solution to the problem, in that we propose weighted estimators, which get rid of the bias problem.

The R package named *WestC* has been implemented, with functions that, among others, give the user the possibility to calculate bias free correlations and covariances in bipartite systems, and which is available upon request to the authors.

Chapter 3

Emergent phenomena in bipartite complex networks: detection of fraudsters' communities and motifs in the Italian insurance sector

Abstract

Fraud is a social phenomenon and fraudsters often act in collaboration with players having different roles. Supervised methods, although they add value to the analysis, show two main drawbacks: first, their calibration is based on a set of known frauds that are very difficult to obtain, and that are a very small sample with respect to the total claims. Second, they miss a peculiar feature of frauds in motor insurance, i.e., the existence of "criminal infrastructures".

We develop an investigation system based on the application of bipartite networks to highlight the relationships between subjects and accidents or vehicles and accidents. Starting from the dense complex network, we construct statistically validated networks to prune connections that do not show statistical anomaly if compared to the random case. We formalize the filtering rules through probability models and test specific methods to assess the existence of communities for very large networks and propose new alert metrics of suspicious structures. We apply the methodology to a real database—the Antifraud Integrated Archive (AIA)—and compare results to out-of-sample fraud scams assessed by the judicial authorities.

3.1 Introduction

Information and communication technologies allow storing big mole of data in very efficient, and cost effective, data warehouses. This is also possible by consolidating and integrating data with different levels of heterogeneity and variety of sources, including social media, email, archives and documents. In the car insurance industry, accident claims are an example of heterogeneous and multidimensional data as they include—not being exhaustive—coded identity of all the subjects directly involved in an accident, such as, drivers, passengers, car owners, witnesses, and pedestrians; professionals, such as, doctors, lawyers, car repairs, as well as details about injuries, fatalities, requested amount, property damage, place and time of the accident, and all about the vehicles involved.

Such a variety and volume of data can be properly exploited through large-scale techniques, integrating ad-hoc mathematical models and fast algorithms in a context where powerful computers can process enormous amounts of data in tiny time frames. A specific field that can take advantage of such techniques is the detection of organized insurance frauds. The aim is to enhance the predictive power of analytical tools by bringing to the surface the hidden interconnections between subjects and events. Indeed, such interactions are usually buried under noisy or spurious relationships and only by means of targeted strategies and appropriate technologies we will be able to dig out the signal content.

The extension of the fraud phenomenon in insurance varies between countries and depends on how the product classifies: life, health, motor and benefit. Experts[1] admit that "across Europe, 10% of all claim euros paid out are considered fraudulent with 21% to 36% of claims potentially possessing elements of fraud." In their annual report—*UK Insurance & Long Term Savings Key Facts*—the Association of British Insurer dedicates a section to the fraud phenomenon and they allege that "fraudulent motor claims were the most common, with over 68,000 cases in 2016" and they are valued up to £780m, which is 60% of the total volume of detected cases of attempted claims fraud in 2016 [180]. The phenomenon is very wide and it goes from one side of the spectrum where opportunists invent or exaggerate a claim, to the other extreme where highly organized criminal gangs set up sophisticated motor fraud scams. To this purpose, in 2012 ABI launched the Insurance Fraud Register (IFR) to convey all data on known fraudsters in a single database, and also equipped it with a comprehensive package of analytics used to provide insurance intelligence.

[1] http://www.interfima.org/publications/insurance-fraud-expert-insights-may-2015-part/.

Along the same line, in 2012 the Italian Parliament passed a bill[2] to entrust the IVASS[3]—the Institute for the Supervision of Insurance—with the task to "fight against fraud in the motor liability insurance sector by analyzing and evaluating the information obtained from the claims data bank", by also giving IVASS the responsibility to manage the AIA an industry-wide database where insurance companies are compelled to upload a detailed description of all the claims for motor policies. Unlike IFR, AIA is a database collecting information about the many actors involved in an car accident: from the drivers to the subjects injured (if any) also including lawyers, medical examiners, insurance repairers, witnesses, amount claimed, vehicles and many other aspects. In this respect, AIA can be considered a *"data lake"* [196]. It is a comprehensive and exhaustive register of the claims issued from 2012, where, however, no explicit information about fraudsters is given, and any conclusion must be drawn relying on statistical analysis and specific analytical tools. Since 2011, IVASS developed a set of alerts to signal its stakeholders unusual levels of some indicators (e.g., number of accidents of a driver, number of involved injuries, claimed amount). Usually, such indicators are binary, measuring the presence or absence of a specific claim characteristic, and an alert is triggered when they trespass a given thresholds based on recurrences and cross-checks criteria.

The scientific literature offers a rich set of statistical tools to identify insurance fraud patterns. They can be partitioned in two wide classes whose main distinctive feature is if they make use of training sets from the fraud and the non-fraud groups (supervised methods), or they rely on "unlabelled" data where account of frauds, together with their covariates, are not available (unsupervised methods). Both approaches have pros and cons, and there is no "fit-for-all" method. (See, [54, 189] for a review and [22, 36, 37, 26] for model specifications and implementations.)

As observed, fraud is a social phenomenon and fraudsters often act in collaboration with players having different roles. Supervised methods, although they add value to the analysis, show two main drawbacks: first, their calibration is based on a set of known frauds that are very difficult to obtain, and that are a very small sample with respect to the total claims. Second, they miss a peculiar feature of frauds in motor insurance, i.e., the existence of "criminal infrastructures" that also encompass the professional profiles operating in this field. Network models have been proved to be a successful methodology to identify social phenomena. In particular, networks methods are suitable to disentangle complex patterns and obtain hidden signals from large and noisy

[2] Decree-Law No 179/2012, article 21, converted to Law 221/2012.
[3] Istituto per la Vigilanza sulle Assicurazioni, `http://www.ivass.it`.

set of data ([148] e [60]).

In the vehicle insurance context, many software companies offer products implementing social network analysis to extract fraudsters patterns from data lakes. Nevertheless, scientific literature is lacking of a formal and rigorous discussion on the subject matter. To the best of our knowledge, the sole article interlacing graph theory and insurance fraud is by [176], who describe a decision support system, to unveil odd network structures in motor insurance claims. Their approach draws from two basic characteristics of the fraudulent behaviour: (i) the "collaborative nature" of fraudsters, involving many different actors, and (ii) the continuous innovation in fraud mechanisms that necessitates a flexible approach, so that "unlabelled relationships" can emerge as soon as they are committed.A major drawback of [176]'s system is the limited size of data samples it can handle. Indeed, [176] build networks upon police records. That is very restrictive since most of the claims do not go through police investigation activities. When only data lakes are available—as in our case—the structures of the suspicious have to first be validated by means of a "filtering" stage, in order that only statistically significant relationships are kept.

The main contribution of our paper is threefold. First, we start by building bipartite networks to highlight the relationships between subjects and accidents or vehicles and accidents. This is a general approach that allows to include the whole spectrum of actors around a claim: from the drivers to the legal professionals. The dense networks obtained has to be filtered out to prune those connections that score a low likelihood level with respect to random chance. In this respect, only structures with very strong ties will appear, thus signalling potential group of fraudsters. Clearly, we are aware that a statistical anomaly cannot be considered a guilty sentence. But, such an information is vital for investigating units as it strongly reduces the—virtually—uncountable number of structures, and, therefore, the cost and the time to liquidate honest claimants.

Second, we formalize the filtering rules through probability models and we will also test specific methods to assess the existence of communities for very large networks and propose new alert metrics of suspicious structures.

Third, we apply the above methodology to a real database—the AIA—and compare results to out-of-sample fraud scams assessed by the judicial authorities. We carry over longitudinal analyses from 2011 to the present to assess possible persistence phenomena of suspicious relationships, and cross-section analyses to collect insights about the spatial structures of frauds throughout the entire Italian territory.

3.1.1 Main challenges: heterogeneity, non-stationarity, localization effects and community detection

The whole methodology is tailored to deal with a very large mole of data. Indeed, AIA is a fully-fledged *data lake* containing detailed information about all of the accidents occurred in Italy since 2011, with overall 15M accidents, 20M subjects, and 13M vehicles. The database AIA is a truly, fully-fledged, *data lake* gathering dozens of tables and millions of records of disparate types (see subsection 3.2.1 for a more precise description). The complexity of AIA requires specific analytical tools to extract the fraudulent patterns and poses challenges that need to be addressed through an advanced multi-level system. We list below the main challenges we identified in preliminary discussions with IVASS's fraud analysts, and that we faced in analyzing AIA during the project development:

Challenge I *Curse of dimensionality.* The complexity of AIA arises from the combination of two dimensions: to one extent, the variegate forms of its data that carries the information related to each claim; to the other extent, the massive size of records that could undermine—or make impossible—to apply methods that proved to be effective for small–medium size samples. *Community detection* is one such example (see subparagraph 3.3.3).

Challenge II *Identification and frequency of frauds.* Labelling as fraudulent a claim is not an easy matter. The investigation units of the insurance companies usually adopt regression models based on a set of indicators sensitive to the detection of fraud and whose output is the probability that a given instance contains elements of fraud. Not all the claims deemed as *"suspected"* are then prosecuted. In general, the decision to open an in-depth investigation depends on the cost of the claim settlement. Once triggers activate an inquiry, negotiations also start. The possible result is that an agreement is reached and the case is closed, or that the claimant withdraws his complaint, or that the case is taken to the Court. The only information available to IVASS (but not included in AIA) are the *claim withdrawals*. Their number, however, is very small compared to the whole AIA and they cannot really assumed to be frauds. Even smaller is the number of frauds assessed by the Court. The acquisition of such information is not systematic because legal authorities have no obligation to inform IVASS.

Challenge III *Heterogeneity.* The database AIA is populated with information about all the actors involved in the *"accident/claim chain"*: from the claimant to the insurance adjuster; from the witness to the lawyer; from the injured to the physician. In principle, no subject or professional can be excluded *a priori* from the scam investigation[4]. The main consequence is that subjects with very few connections (a witness, or an injured) will *"live with"* others highly connected (lawyers or car repairers). The challenge here is that any statistical model used to test for anomalies has to account for such a heterogeneity to avoid that actors with few connections will be deemed as not statistically significant.

Challenge IV *Time and space localization.* The data contained in AIA includes claims in the time span between 2011 and 2016, and it covers all the accidents occurred within the Italian territory. Any probabilistic model or data mining approach working with the whole database will run into a serious issue: a small *"perturbation"* (the statistical anomaly) in the calm of the *"sea of noise"* (the null hypothesis) will be readily highlighted, even though it is just a *"ripple"* and not a *"tsunami"*. Out of metaphor, two lawyers exercising their activity in the same city could interact in a significant number of accidents, if compared to the whole accidents in Italy. On the contrary, if we restrict to the number of accidents occurred in the nearby of the city, such a relation might lose its anomalous character. Similar examples can be found for the temporal extent. Note that, focusing the investigation on ex-ante spatial or temporal sub-samples of AIA is not a viable solution, since network of fraudsters, although they have a restricted temporal or spatial perimeter, cannot be confined to administrative boundaries, or limited to artificial temporal segments (years, semesters, etc.). Returning to the lawyers example, without any spatial restriction, we run the risk that lots of relationships, like that described, are signalled as anomalies, whereas to a lower scale (region, city, etc.) would be considered as normal ones.

Challenge V *Homophily.* "Similarity breeds connections" [135], this is in synthesis an outline of the concept of homophily. In crimes related to frauds, homophily plays a relevant role as frauds require a rather high degree of cooperation, coordination, and, therefore, trust among the fraudsters. If not friends, they should be at least acquaintances, which suggests that, unless an external factor destroys the relationship, the same fraudsters are likely to be involved in several frauds together over time.

[4] In reality, subjects with a specific role in the same insurance company are excluded in advance. For example, the lawyer and the car repairers of the same company is very unlikely that they participate to a fraud together.

3.2 Data

3.2.1 The IVASS Integrated Antifraud Archive

The Antifraud Integrated Archive (AIA) is the result of the integration of several databases, managed by both public and private bodies. In fact, the main source is given by the claims database. In addition, AIA gets information from six external databases: vehicle register; driver license register; insurance coverage database, black box files; insurance expert list; public vehicle register. Among other information, insurance companies are compelled to upload in real time detailed descriptions of all the claims for motor policies reported to insurance undertakings. It collects and organizes information about the many actors involved in a car accident: from the drivers to the subjects injured (if any) also including lawyers, medical examiners, insurance repairers, witnesses, amount claimed, vehicles and any other person or company directly or indirectly involved in the accidents. In this respect, AIA can be considered a "data lake" [196]. It is a comprehensive register of the claims issued since 2011, where, however, no explicit information about fraudsters and frauds is provided. Therefore, suspected frauds and fraudsters must be detected on the basis of a statistical analysis of data and the application of specifically devised software. Since 2011, IVASS developed a set of alerts to signal its stakeholders and prosecutors accidents with anomalous levels of some indicators. Usually, such indicators are simply binary variables, denoting the presence/absence of a specific characteristic of the claim. The weighted combination of two or more binary variables are used as an alert, which is triggered when they trespass given thresholds. To give an idea of its size, AIA recorded 16,050,689 accidents and 21,574,410 people at the end of January 2018, and it is quickly increasing. Indeed, the corresponding amounts are 18,592,317 (increase of 15.8%), and 23,943,787 (increase of 10.9%), respectively, at the end of February 2019. AIA represents a complex set of interrelations between subjects and between vehicles, which turn out to be connected whenever they are involved in one or more car accidents together. A way to filter all these random interrelations out of the network, is our main objective.

Such a filtering procedure must properly take into account the heterogeneity of subjects. Indeed, the graphs reported in figure 3.1 indicate that, while accidents show a limited heterogeneity with respect to the number of subjects involved (130 at most), the heterogeneity of subjects is extreme, with a few subjects (companies, of course) involved in more than 100,000 accidents over a period of six years. Therefore, at difference with systems that display a double

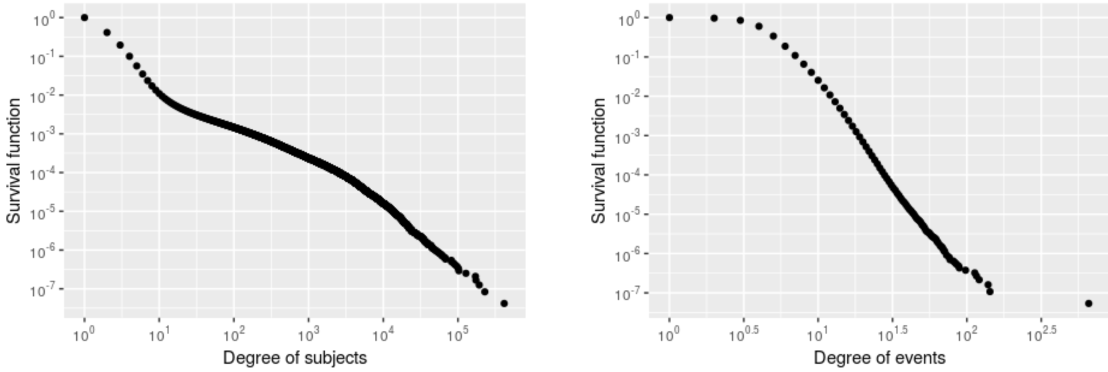


Figure 3.1: Survival function of the degree distributions of subjects (left) and accidents (right) in a log-log scale.

source of heterogeneity ([184, 160]), only the heterogeneity of subjects really matters here, and must be taken into account when filtering the network, in order to detect anomalous patterns. Unlike the bipartite network *subjects - accidents*, the network *vehicles - accidents* shows a lower source of heterogeneity on both sets.
An appropriate white list for the network was constructed, adding subject IDs (for example referring to the army, the police, the government as a legal entity) to the subject IDs that formed the initial AIA white list. This step is necessary since a lot of professionals had a very high degree in the network, being connected to many accidents just for their normal professional activity and not because of fraudulence.

3.3 Methods

3.3.1 ISAAC: an investigation system for Antifraud activity in the motor insurance sector

ISAAC (**I**nvestigation **S**ystem for **A**ntifraud **AC**tivity) is a system to investigate the existence of networks of fraudsters in the motor claims sector. Investigation System for Antifraud ACtivity (ISAAC) faces issues raised by IVASS's fraud experts and who are responsible for the maintenance and management of AIA, a database collecting any car accident claim that occurs in the Italian soil. One of the IVASS's mission is to return to its stakeholders (the insurance companies) analysis of the fraud phenomenon and alerts about potential criminal networks. In principle, IVASS benefits of a privileged position since AIA encompasses the whole insurance claims in the motor market, and it is not limited to the perspective of a single company.

3.3.2 The subject-accident bipartite network

The implementation of ISAAC starts with the construction of a preliminary SVN of subjects. This is done by projecting the subjects-accidents bipartite network with respect to the set of subjects and then perform a statistical test for each link of the resulting projected network of subjects. As described in paragraph 1.2.2 of chapter 1, for each pair of subjects we test the hypothesis of randomness of co-occurrences (accidents they have in common), considering the hypergeometric null distribution of eq. 1.6 and adjusting the statistical significance level according to the Bonferroni correction for allowing multiple comparisons (described in paragraph 1.2.3). Obviously, due to the huge dimension of the SVN, it is practically impossible to view it all. Rather, smaller parts of it can be viewed. As an example, Figure 3.2 shows a connected component belonging to the SVN of subjects.

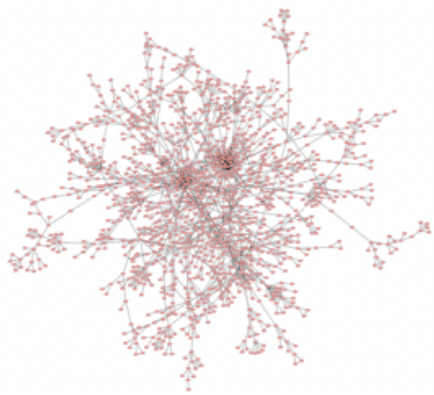


Figure 3.2: A connected component of the SVN of subjects.

Notice that attention must be paid to the effects that time and geo-localization of accidents may have on the rate of false positive links, i.e. links formed by subjects who did not behave in a fraudulent manner and that are classified as potential fraudsters. This aspect is apparent, for instance, when two professionals work in the same restricted area. They could show a lot of co-occurrences due not to fraudulent activity, but just because they operate in the same area, therefore having a high probability of being involved in the same accidents together in a certain time window. To overcome this problem, we introduce a Robustness score (R-score) R_{ij}, computed for each validated link. Given the pair of subjects i and j,

$$R_{ij} = \log_{10} T - \log_{10} m^*_{ij} \tag{3.1}$$

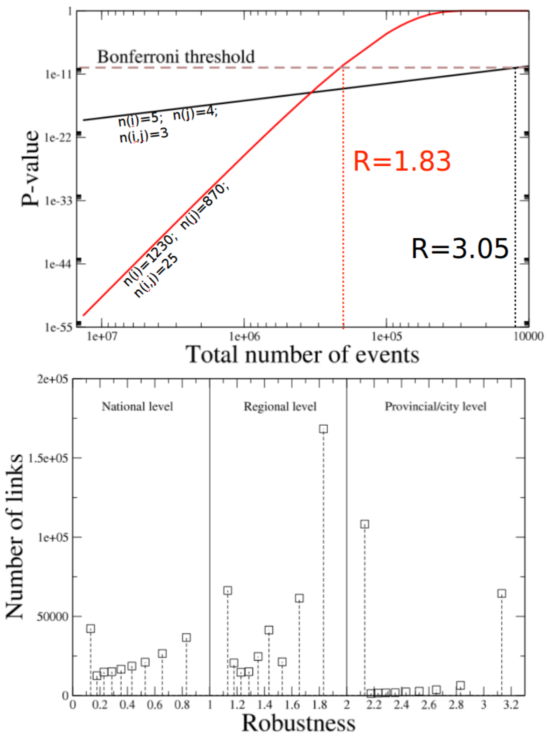


Figure 3.3: Example of computation of the R-score (left) and its distribution (right).

where T is the total number of accidents in the system regardless of the place of occurrence, and m^*_{ij} is the minimum value of T such that link between subjects i and j is statistically validated. Fig.3.3 shows the rationale behind the computation of the R-score.

The lower m^*_{ij}, i.e. the higher R_{ij}, the more robust the link between subjects i and j will be. Once the R-score has been assigned to every link in the SVN, decision about whether they must be discarded or not comes after a community detection procedure.

3.3.3 Community detection

Community detection is a fundamental step in the analysis of the AIA database, and in particular of the SVN, in order to highlight organized groups of suspected fraudsters. Community detection in large networks, such as the present one, is challenging due to the intrinsic nature of the problem. Qualitatively speaking, a community in a network is a list of nodes (subjects in our case) more closely connected among them than to the others. Despite the simplicity of such a qualitative definition, community detection is challenging from several points of view. First of all, it is necessary to introduce a suitable utility function, which should incorporate the properties of the network, e.g., directionality of links, weights, quality of nodes, etc.. The most popular and adaptive utility function for community detection is modularity, which, in its basic form, has been introduced by [78]. The modularity of a partition is an additive function of the modularities associated with each specific community of nodes, and the modularity of a community is calculated as the difference between the number of links actually observed among community members and its expected value under the hypothesis of random connectivity [151]. Therefore, in principle, modularity should be calculated for all the possible partitions (in any number of communities) of the vertices of a network, and the optimal partition is the one that corresponds to the maximum value of the modularity. Community detection is an NP-complete problem, and many heuristic methods have been devised to provide sub-optimal solutions in polynomial time ([151, 74]). Alternative methods to modularity optimization have been proposed in the literature, most of them relying upon the idea of a process running on the network, e.g., a random walk. If one considers a random walk in which a particle can travel on the network from one node to another only crossing existing links and randomly selecting the link to cross, it is intuitive that it should spend more time cruising in a community, which is unknown yet, than traveling across communities. A popular method of community detection that is based on this idea is the Infomap, which has been proposed by [165]. It is worth saying that community detection methods based on modularity optimization and methods based on processes running on the network can bring to rather different partitions of vertices. How to choose the most appropriate method in real networks? It depends on the nature of the network and on the information available, if any, about the expected size and structure of communities. Our polar star in the present analysis is highlighting groups of potential fraudsters. This objective sets weak boundaries on the size of communities. Indeed it is unreasonable to envision the existence of organized groups of fraudsters made of thousands of individuals. On the other hand, it is useless to focus on very little communities, made of two or three subjects involved in a little number of accidents, since the cost of performing an actual investigation of related events could be much higher than the value of the fraud itself. Therefore our main focus should be on communities made of tens to hundreds of individuals. A hundred might also appear a large number, however empirical evidence indicates that groups of fraudsters of such a dimension actually exist, in connection with organized crime. In our case, modularity optimization seems to be the most appropriate approach, as our network is essentially a network based on co-occurrence, and no information naturally flows on it. However, IVASS uses a SAS procedure to perform community detection relying on modularity optimization, which in turn involves a tuning parameter.

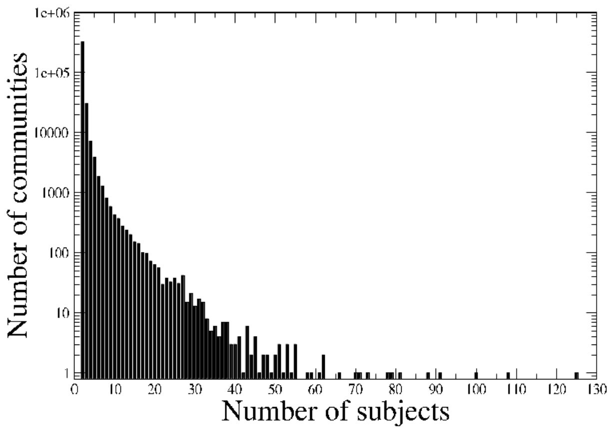


Figure 3.4: Distribution of community dimension.

We set the value of the tuning parameter as the one that leads to a partition of nodes as close as possible to that obtained by the infomap algorithm, and we have used a combination of different heuristics, such as extreme optimization ([58]), taboo search, etc., and introduced weak constraints on community size, as discussed above, as well as time and geographical corrections, when appropriate.

Community characterization

Characterization of communities is an important task for modeling the homophily that is showed by subjects through their behaviour. The same approach used for the validation of links (see Eq.1.10) is now used for associating each community with one or more over-expressed attributes, which can be referred to one or more geographic areas (region or province), years of occurrence of car accidents, and subjects' roles (in Fig.3.5 we report some examples).
Denoting by N the number of subjects within the network, N_c by the number of subjects within community c, N_p by the number of pedestrians in the network, and $N_{p,c}$ the number of pedestrians who belong to community C, the probability linked to $N_{p,c}$ is equal to Eq. 1.6, where $x = N_{p,c}$, $N_c = n_i$, and $N_p = n_j$. To say that an attribute, e.g. *pedestrian*, is over-expressed for a certain community c, we apply the hypergeometric test of Eq. 1.10.

If the observed value of $N_{p,c}$ is statistically greater than what we would observe in a situation of completely uniform distribution of attributes in the system, then we'll say that attribute *pedestrian* is over-expressed, and therefore, characterizes community c, that is, if $P(N_{p,c}^{obs} \geq N_{p,c}^{0.05}) < 0.05$, then we'll say that attribute *pedestrian* is over-expressed in community c. In the particular situations where communities have few nodes or where the attribute we study is rare in the system, the hypergeometric test leads to unreliable results

Comm. ID	Size (events)	Years over-expressed	Regions over-expressed	Provinces over-expressed
1	152,906	2015, 2016	SARDEGNA, LOMBARDIA, LAZIO	VA, TV, TP, TO, SS, RM, RN, RG, PO, PT, PE, PV, PD, MI, LO, LC, LT, CO, CL, CA, BG, MB, OG, VI, VR, AG
2	117,396	2011, 2012	CAMPANIA*, NA	NULL, SA, AV, NA, CE
3	123,216	-	TOSCANA*, NA	NULL, SI, PO, PT, PI, AR, LU, FI
4	115,573	-	PIEMONTE*, VALLE_D'AOSTA	VC, TO, AT, AO, CN, BI
5	88,799	-	BASILICATA, PUGLIA*, NA	NULL, BA, TA, PZ, MT, FG, BR, BT
6	92,177	-	FRIULI_VENEZIA_GIULIA, VENETO*	VE, UD, TV, RO, PN, PD, FE, VI, VR, BL
7	83,589	-	SICILIA*	TP, PA, AG
8	132,361	-	LAZIO*	RM, RI, LT, VT
9	73,537	-	SICILIA*, NA	NULL, SR, RG, ME, EN, CT, CL
10	71,974	-	EMILIA_ROMAGNA*	RN, RA, OR, MO, FC, FE, BO
11	100,036	2015, 2016	LAZIO*	RM, RI, LT, FR, VT
12	69,680	2011	FRIULI_VENEZIA_GIULIA, VENETO	VE, UD, TV, PN, PD, NO, GO, VI, BL
13	65,887	-	LIGURIA, NA	NULL, SV, SP, IM, GE, AL
14	64,568	-	LAZIO, NA	NULL, RM, LT, VT
15	68,079	2015	CAMPANIA*	SA, AV, NA, CE
17	57,989	-	EMILIA_ROMAGNA*, NA	NULL, RE, PR, MO, MN, FE, BO
23	65,884	2016	LOMBARDIA	VA, PV, MI, LO, LC, CR, CO, BG, MB
25	51,217	-	LOMBARDIA, NA	PC, MN, LO, CR, BS, BG, VR

Figure 3.5: Example of communities with over-expressed years, provinces and regions.

due to its discrete nature. Therefore, we say that an attribute characterizes a community when at least 90% of nodes in the community has that attribute. *Example*: community c has 3 subjects, all witnesses. The test for the value of $N_{p,c}$ may not be statistically significant but, since the attribute *witness* is the role of 100% of subjects in the community, we will say that the attribute *witness* characterizes that community.

R-score at the community level

Once a first detection of communities is completed, we associate each of these communities with a value of R-score:

$$R_k = \log_{10} T - \log_{10} n_k^* \tag{3.2}$$

where T is the total number of accidents in the system and n_k^* the number of accidents occurred in the place (or places) and in the year (or years) that characterize community k. We compare the R-score (R_{ij}) of a link between a generic pair of nodes i and j with the R-score (R_k) computed for the community they belong to. This comparison provides a way to remove links that are not very robust compared to other links belonging to the same community in the SVN. Indeed, remembering that m_{ij}^* is the minimum value of T such that link between subjects i and j is statistically validated,

$$R_k - R_{ij} = \log_{10}\frac{T}{n_k^*} - \log_{10}\frac{T}{m_{ij}^*} = \log_{10}\frac{m_{ij}^*}{n_k^*} \quad \Rightarrow \quad 10^{R_k - R_{ij}} = \frac{m_{ij}^*}{n_k^*} \tag{3.3}$$

On one hand, if $m_{ij}^* < n_k^*$, then $R_k - R_{ij} < 0$ meaning that the link between i and j is very robust and should be kept within community k. On the other hand, if $m_{ij}^* > n_k^*$, then it means that the link between i and j is not validated when considering a number of accidents that exceed the number of accidents characterizing community k, therefore being less robust than expected within the same community. Specifically, we remove the link between nodes i and j if

$$R_k - R_{ij} > t^* \quad \forall i \neq j : \{i, j\} \in \text{community } k$$

The threshold t^* is fixed to 0.1, that is, when m_{ij}^* is about 26% greater than n_k^*. The choice of t^* is made in order for us to be not too restrictive when deleting links from the SVN. Also, there is no unique way to choose this threshold. Eventually, this procedure will bring the benefit of reducing potential false positive links from the SVN, leading to the final SVN. After this step is completed, the community detection algorithm used before is again performed to find the new community structure in the SVN, together with the characterization of its communities.

Bipartite SVN and enlarged SVN

The SVN allows one to spot anomalous relationships between subjects but it does not give explicit information about the accidents these subjects were involved in. In fact, accidents may represent our unit of interest in order to further investigation activity. Starting from the SVN of subjects one can define the bipartite SVN, linking subjects to the accidents that contributed to the statistical validation of their relationships. If we also include all the subjects that were directly involved in the accidents of the bipartite SVN, then we refer to the enlarged SVN, which leads to an increase of 2 people per person on average.

3.3.4 The vehicles-accidents network

The approach used for the construction of the SVN of subjects, aimed at the detection of anomalous relationships between subjects, can be extended to the study of the bipartite network vehicles-accidents in order to detect anomalous relationships between vehicles.

Unlike the SVN of subjects, the SVN of vehicles is much less structured as in general a vehicle is linked to a limited number of subjects (see Tab. 3.1). Therefore, community detection and the correction for time-space localization are not needed in this case and the focus is given to small highly connected components.

Table 3.1: Dimension of SVN of subjects and SVN of vehicles.

	Nodes	**Links**	**Connected Components (CC)**	**Dimension of the biggest CC**
SVN of subjects	2,016,505	1,919,897	638,878	651,267
SVN of vehicles	112,771	61,311	54,563	12

The information carried by the SVN of the vehicles-accidents network is useful to be integrated with that of the SVN of subjects-accidents network. Its inclusion in the detection fraud activity will allow to study a complete set of complementary knowledge of the linkages between subjects, vehicles and accidents.

3.3.5 Network structure and properties

Relying on the data stored in AIA at the end of February 2019, the number of communities detected within the SVN reaches 488,362. About the 60.2% of these communities is made up by only four nodes (two subjects and two accidents), while about 9,767 communities (the highest 2% of all the communities) has a number of nodes between 26 and 13,778.
In Tab. 3.2 we report the number of communities belonging to each combination of the macro-groups formed according to the characterization of roles of subjects and time/space localization.[5].
A description of the network community indicators and a descriptive analysis

	P	NP	P-NP	$\overline{P}$-$\overline{NP}$	None	Overall
# of communities	15,403	112,103	310	300,564	59,982	488,362
# accidents (average)	58.5	2.3	45.3	3	3	4.6
# subjects (average)	6.2	2.1	10.1	2.2	2.3	2.3
# links (average)	123.2	4.7	97.3	6.2	6.2	9.6

Table 3.2: Number of communities and average of nodes, subjects and links, according to community characterization: professional roles only (P); non-professionals only (NP); both professionals and non-professionals (P-NP); only time and/or space attributes ($\overline{P}$-$\overline{NP}$); no characterization.

of their conditional distributions according to macro-categories classification are reported in Tables 3.8 and 3.9 respectively. Communities characterized

[5] communities characterized only by time and/or space attributes show a limited variability in the network indicators, as shown in Tab. 3.2 under column $\overline{P}$-$\overline{NP}$

by attributes related to professionals are basically the biggest ones, highly connected and with more robust links, also showing a higher variability in the community network indicators. On the other hand, communities that are characterized only by attributes related to non-professionals, tend to be smaller and sparser. Communities belonging to the combinations (P-NP and $\overline{\text{P}}$-$\overline{\text{NP}}$) are halfway between the previous extreme cases.

3.3.6 SVN for classification purposes

The objective of this work is to enhance the IVASS antifraud activity with a very powerful and effective tool, but also simple for usage and interpretation at the same time. Once the SVN is constructed, the objective is predicting the degree of statistical anomaly of co-occurrence of future accidents. With the definition of an integrated indicator of statistical anomaly, we will be able to give a simple and immediate way to tell insurance undertakings which accidents, subjects and so communities of subjects they should pay closer attention to. First, since we start from a set of correlated variables describing the aspects of size, connectivity and robustness of a community at the network level, as well as indicators at the individual level of accidents, we perform a PCA to capture all the core information in the system and make the predictive model as parsimonious as possible by reducing redundant information from the data. The number of principal components is chosen based on the Random Matrix Theory (RMT) (see Fig. 3.6), showing that three eigenvalues (and so, principal components) are actually useful to grasp a statistically significant proportion of the variance in the system. Second, we use a classification model to discriminate reported accidents and random ones. Many machine learning algorithms could be used to deal with binary classification problems, such as logistic model, Support Vector Machine, binary classification trees etc. We use the logistic model to estimate the predictive power of the principal components. This choice is preferred to other approaches because of its simplicity and easiness in the interpretation of results. This phase of the analysis exploited the information of 9,199 accidents, 4,566 of these being accidents reported by insurance undertakings to IVASS and 4,633 accidents being a random sample of accidents picked from AIA, sampled based on an opportune stratification of AIA according to geographical and time localization, which reflects that of reported accidents.

When estimating the model the fourth principal component is statistically significant (but not the fifth) and therefore relevant in discriminating between random and reported events. This step allows us to associate the estimated coefficient with each principal component, and finally, use these coefficients to build our final indicator.

$$II = \hat{\alpha}_1 PC_1 + \hat{\alpha}_2 PC_2 + \hat{\alpha}_3 PC_3 + \hat{\alpha}_4 PC_4 \tag{3.4}$$

where $\hat{\boldsymbol{\alpha}} = (0.113, 0.213, 0.368, -0.833)'$.
Accidents reported by the insurance undertakings tend to have higher values of the principal components, and so of the integrated indicator (see Fig. 3.7 (right)).
Moreover, for practical reasons, the integrated indicator was used to define four classes of statistical anomaly, specifically *null, low, medium, and high.* The thresholds are chosen based on the percentiles of the distribution of the integrated indicator, and in particular, the 33^{th} percentile, that is approximately the mode of the distribution, and the 66^{th} percentile, that is approximately the value for which the Matthews Correlation Coefficient is maximized. Another aspect that is considered when classifying accidents is whether they belong to the SVN or not (see Tab. 3.3).

	$a \notin$ SVN	$a \in$ SVN
$X(a) \leq t_{33^{rd}}$	null	low
$t_{33^{rd}} < X(a) < t_{66^{th}}$	low	medium
$X(a) \geq t_{66^{th}}$	medium	high

Table 3.3: Classes of statistical anomaly according to the value of the integrated indicator and to whether the accident a belongs to the SVN or not.

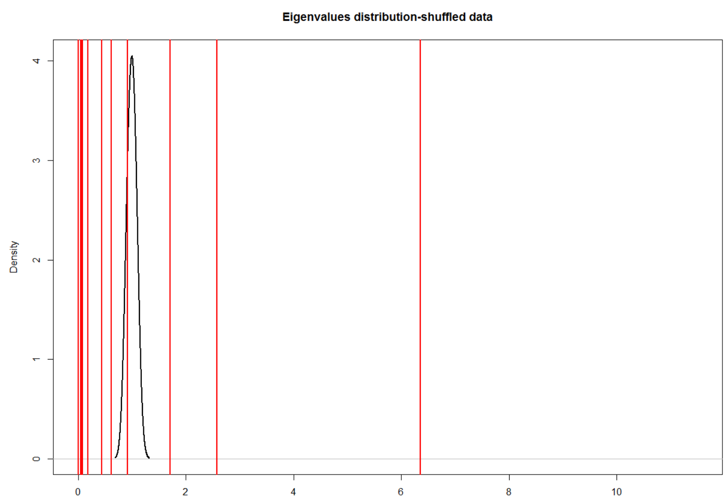


Figure 3.6: The set of eigenvalues under the random case of no correlation structure in the data is represented by the black distribution (centered in 1). Red vertical lines are the eigenvalues of the correlation matrix of the observed standardized data.

3.3.7 Statistical anomaly of communities

Since any accident can be associated with a level of statistical anomaly, consequently any community of the SVN can also be associated with a level of statistical anomaly, based on the anomaly of its accidents: for instance, one way of associating a community with a "high" statistical anomaly could be based on whether the community contains a given number of accidents with a high statistical anomaly, depending on the dimension of the community.

We focus the attention on communities that include at least 4 accidents, since start detecting very small communities is not convenient in terms of costs and benefits comparison. Moreover, we say that a community is statistically highly anomalous when at least the 66.7% of its accidents shows a high score of the integrated indicator. Also, we take into account for the presence of accidents that belong to two or more communities. In fact, these accidents show a higher proportion of accidents with a high score of the integrated indicator, 70% (175,304 out of 250,370) against the 54% characterizing the accidents belonging to only one community (1,092,222 out of 2,014,525). Therefore, the 6.1% (29,965 out of 488,362) of communities are associated with a "high" level of statistical anomaly.

3.3.8 Effectiveness of the method: case studies and out-of-sample validation

The usual approach to solve this kind of classification problems involves the quantities shown in Table 3.4. By varying the value of the threshold x_0,

	Random	Reported	
$X \leq x_0$	TN True Negatives	FN False Negatives	TN+FN Negatives
$X > x_0$	FP False Positives	TP True Positives	TP+FP Positives
	TN + FP Real Negatives	TP+FN Real Positives	

Table 3.4: Quantities involved in a classification problem; x_0 is the generic threshold for the composite indicator X.

the aim is minimizing the number of false positives, as the more they are, the more the costs for insurance undertakings in terms of time and money will be, but also false negatives. Measures that give the idea of correct classification in terms of true positives and true negatives are, respectively, *sensitivity* as TP/(TP+FN) and *specificity* as TN/(TN+FP). Instead, a measure that takes into account true and false positives and negatives and is generally regarded as a balanced measure is the Matthew's correlation coefficient (MCC) introduced

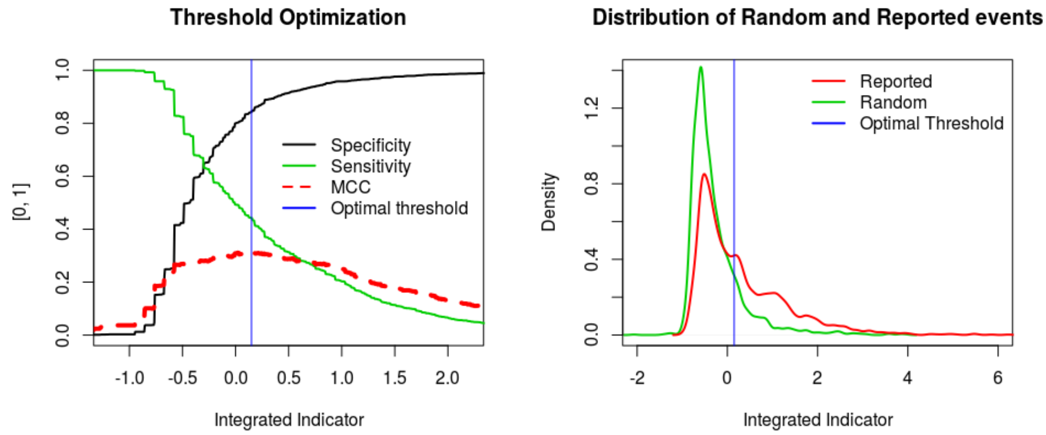


Figure 3.7: Estimation of the optimal threshold based on MCC maximization (left) and the two kernel distributions of random and reported sub-samples (right).

by [133].

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{3.5}$$

We perform an out-of-sample validation process. Specifically, the initial dataset was partitioned in two parts such that the 80% (7,359 units) forms the training set and the remaining 20% (1,840 units) the test set. Also, the same proportion of reported and random units was maintained while forming the training set. This procedure was iterated 500 times so that the first two moments of the sampling distributions of the main performance measures could be studied (see Tab. 3.5).

Using this approach, the integrated indicator is reasonably sensitive, classifying as fraudulent the 67.7% of true frauds, and specific, classifying as non-fraudulent the 57.1% of the accidents belonging to the random group of accidents drawn from AIA. It is worthy to note that while frauds are associated with a hard label, controls are associated with a soft label, since AIA consists of about the 20% of frauds. Also, the ability of the model to detect frauds among true frauds is higher than that among the controls, eventually reaching on average an accuracy of 62.3%

3.3.9 K-fold cross-validation performance of the model

In the previous subsection we described the integrated indicator used by the IVASS for associating an accident of interest with a score of statistical anomaly. In this subsection we run logistic regressions through a 5-fold cross-validation technique, using, this time, the set of original variables. We show

Performance measure P	$\mathbb{E}[P]$ $(se[P])$
MCC	0.253 (0.023)
Specificity	0.571 (0.090)
Sensitivity	0.677 (0.087)
Relative Risk	1.601 (0.059)
Accuracy	0.623 (0.012)

Table 3.5: Out-of-sample empirical expected values and standard errors (in parentheses) of main classification performance measures.

how the introduction of SVN improves the classification performance of the classifier when compared to the case where, in fact, only the score AIA was used. In particular, we derive Receiver Operating Characteristic (ROC) curves for three cases: (1) we consider only the score AIA as explanatory variable; (2) we consider only network variables and a dummy variable indicating whether or not the accident belongs to the SVN, and (3) we consider both points (1) and (2) together. Moreover, for each of the three cases, we trained the model under both balanced and unbalanced data settings. Results are shown in Table 3.6, and Figure 3.8 shows the ROC curves. It is worth to note that

	Reported vs Random			
AUC	I - **400 vs 400**	II - **400 vs 4,000**	III - **400 vs 40,000**	IV - **400 vs 400,000**
(1) AIA score	0.62	0.62	0.61	0.62
(2) Network	0.86	0.86	0.83	0.83
(3)=(1)+(2)	0.87	0.87	0.85	0.85

Table 3.6: AUC for the balanced and unbalanced cases (ratios reported/random accidents: 1:1, 1:10, 1:100, 1:1000). Results are shown for three cases: (1) score AIA only; (2) network indicators only; (3) score AIA and Network indicators.

the performance of the model increases thanks to the application of the SVN, and the AUC increases when considering both the score AIA and community network indicators as features of the model. Also, the same results hold in the case of unbalanced data.

	Random	Reported	Random $\in$ SVN	Reported $\in$ SVN	Random	Reported $\in$ SVN
$X \leq x*$	307 (76.7%)	79 (19.8%)	358 (89.5%)	87 (21.8%)	371 (92.7%)	13 (3.3%)
$X > x*$	93 (23.3%)	321 (80.2%)	42 (10.5%)	313 (78.2%)	29 (7.3%)	387 (96.7%)

Table 3.7: Confusion matrices under the case I-(3) reported by Table 3.6 – comparison between reported accidents and random accidents both belonging to the SVN, and between reported accidents belonging to the SVN and random accidents from AIA; $x*$ is the probability threshold that maximizes MCC. In parentheses the percentages with respect to column totals are reported.

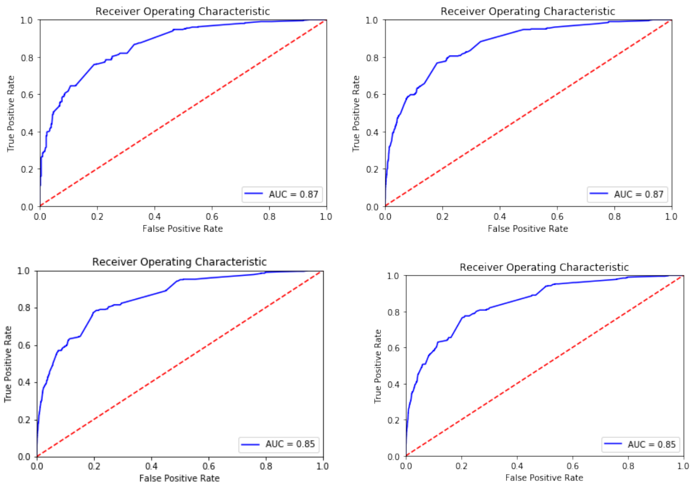


Figure 3.8: From top-left to bottom-right: ROC curves for cases I, II, III, IV, with the specification of (3) according to Tab. 3.6

3.3.10 Effectiveness of the method: three case studies of detected communities of fraudsters

A crucial aspect in evaluating the effectiveness of our method concerns the ability to spot empirical cases of fraudulent organizations that are referred to IVASS from external sources, assessing the presence of fraudulent people and accidents in the SVN. This paragraph remarks the positive impact that our investigation system brings to the fraud detection activity performed by IVASS. Specifically, we report here three empirical case studies of fraudulent organizations, that are structurally different in terms of link formation, nature of nodes, and scale dimension.

The first case study considers the information about three fiscal codes belonging to three out of the five components of a family. For this case, the father, that divorced his wife, was the one claiming to the insurance company that the wife and one of their children were organizing frauds. We first checked for their presence in the SVN, and after that, we observed how many car accidents they were involved in. Consequently, we added all subjects that were involved in the accidents of the SVN, obtaining the enlarged SVN. Fig. 3.10 shows the fraudulent sub-network with accidents involving at least one of the family members, which highlights the connections between the mother, the father, their three sons (one of them being 3 years old), two mother's relatives and two professionals, specifically a physician and a technical expert.

It's important to notice that the method is able to detect fraudulent organizations acting on very different scale dimensions—small in the latter instance—and it also manages to integrate information that is not known a priori: two out of three children and two relatives of the mother were not initially claimed by the father to IVASS, while they are spotted in the SVN. Moreover, six out of seven (85.7%) accidents have been associated by the integrated indicator with a high level of statistical anomaly (marked in red in the graph), and one accident with a medium level of statistical anomaly (marked in orange in the graph).

The second case study consists of a network on a larger scale if compared to the previous one of family members. It comes from nineteen fiscal codes reported to IVASS by the prosecutor office of an Italian city, and it describes the fraudulent activity of people belonging to organized criminality (Fig. 3.11). Also in this case, the integrated indicator manages to associate the majority of accidents with a high level of statistical anomaly (60% and most of them being in the deepest and most connected part of the network), a 20% of accidents is associated with a medium level, and therefore the remaining 20% with a low level of statistical anomaly. Note that no accident is associated with a null level of anomaly as long as it belongs to the SVN.

Finally, the third case study consists of a network on an even larger scale if compared to the previous networks. It's a network of people and accidents involving 313 car plates in the context of a legal identity theft reported to IVASS by the prosecutor office. The number of car accidents and subjects linked to the 313 plates are 874 and 3,004 respectively in AIA. When we look at the bipartite SVN, 1,313 of those subjects are involved in 88,672 car accidents, forming a total of 979 communities. One of the subjects (marked with a bigger black node in Fig. 3.12) is linked to the VAT number of the robbed company, covering a central position/role in the network. The integrated indicator classifies as highly potential frauds the 42.2% of the accidents, while the 19.4% and 38.2% are classified as having, respectively, a medium and a low level of statistical anomaly. Therefore, starting with external information about a set of claimed subjects/accidents/car plates, and despite the relatively low proportion of subjects and accidents being in the SVN (8.4% and 13.3% of respectively subjects and accidents that are in the SVN), the method proved to be able to detect frauds and to integrate them with other useful information.

3.3.11 Life-cycle of communities

We also studied the dynamics of communities of fraudsters. The principled idea is that any community has to have a starting point, a phase of proliferation, and, when they are discovered, a progressive decline. We analysed the dynamics of the communities of fraudsters considered in subsection 3.3.10. Fig. 3.9 shows the time series of the average of the integrated indicator of

Formula 3.4 over the years for the three communities of fraudsters. The network of family members (black solid line) lasts four years, starting in 2012 and ending in 2015. It is rather cohesive and every accident has a high level of statistical anomaly leading to a high average value each year of its existence. The organized criminality network (red solid line) starts in 2011 and its statistical anomaly begins to decrease starting from 2014. That's because in that year some of the criminals are detected by the legal authorities. Finally, the legal identity theft network (blue solid line) starts in 2014, and again, after about three years of activity and proliferation, its anomaly start decreasing from 2017, when some of the people are detected and stopped.

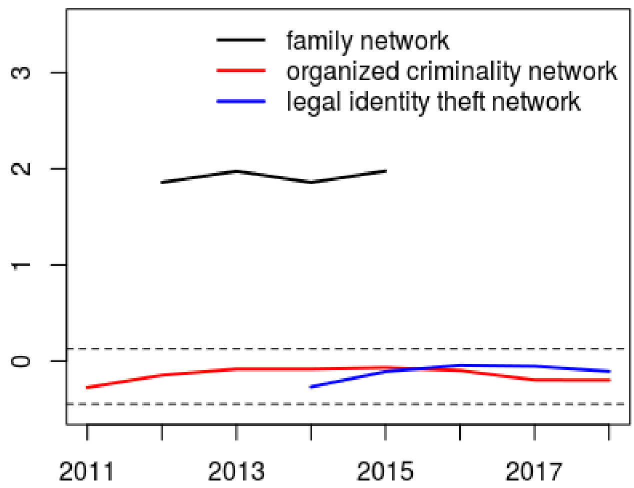


Figure 3.9: Yearly average values of the integrated indicator for the three case studies. Dashed lines represents the thresholds separating respectively low-medium, and medium-high classes of statistical anomaly.

3.3.12 Fraud detection activity from the user-perspective

A dedicated Graphical User Interface (GUI) has been implemented at the IVASS in order for an analyst to be able to benefit from ISAAC. Specifically, the enabled user interfacing with the GUI may input the name and surname or the fiscal code of a subject, or the ID of a car accident, or even a car plate number to search a vehicle. After entering the requested data, the system will output the level of statistical anomaly of accidents/subjects/vehicles according to the value of the integrated indicator, and some descriptive statistics in the mask can be viewed if the user wants to, as additional information, such as the

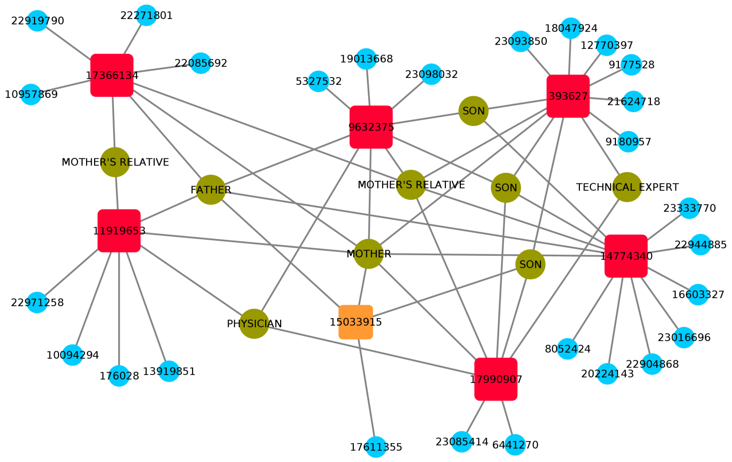


Figure 3.10: Enlarged SVN with accidents involving the reported fraudsters (colored in black). Rectangular nodes are accidents while circular nodes are subjects. Accidents are in red if they have been assigned a "high" level of anomaly according to the integrated indicator; accidents are in orange if they have been assigned a "medium" level of anomaly according to the integrated indicator.

number of people involved in an accident, or the number of accidents linked to a subject, number of links, clustering coefficient, H-K score, etc. Moreover, if a subject/accident/vehicle is in the SVN, then the system will plot the community or communities that contain it, allowing the user to choose between a projected and a bipartite (enlarged or not) network. Also, the user will be able, if interested, to view a particular shell of a network rather than all the network.

3.4 Discussion and conclusions

In this work we developed a novel statistical tool for the detection of frauds and fraudsters' communities in Italy. In particular, we used a statistically validated network approach to analyse AIA, the comprehensive and exhaustive Antifraud Integrated Archive managed by the IVASS. The method proved

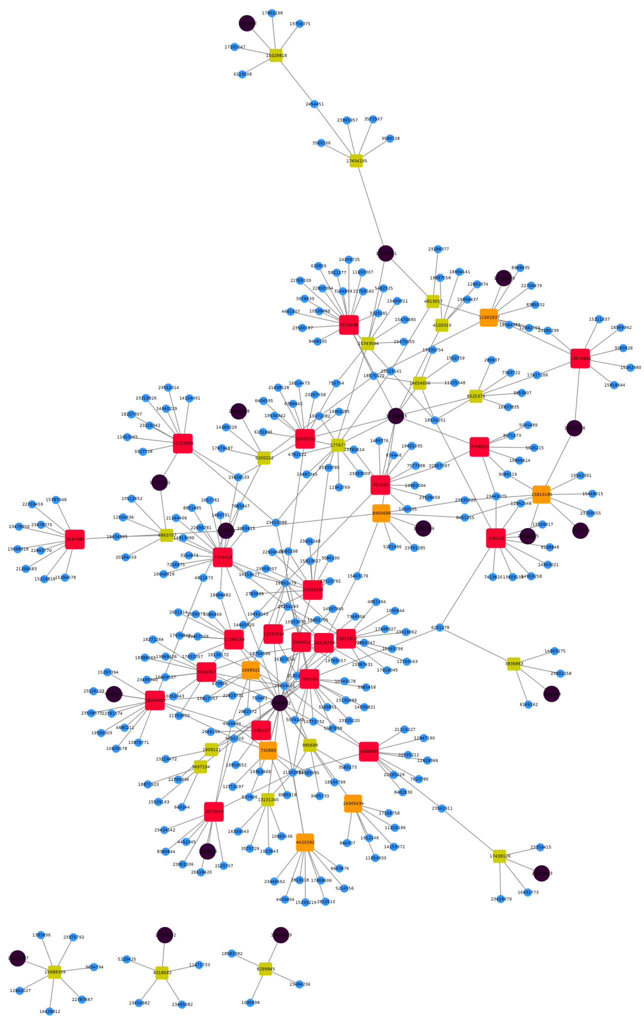


Figure 3.11: Enlarged SVN with accidents involving the reported fraudsters (colored in black). Rectangular nodes are accidents while circular nodes are subjects. Accidents are: in red if they have been assigned a "high" level of anomaly; in orange if they have been assigned a "medium" level of anomaly; in yellow if the have been assigned a "low" level of anomaly according to the integrated indicator.

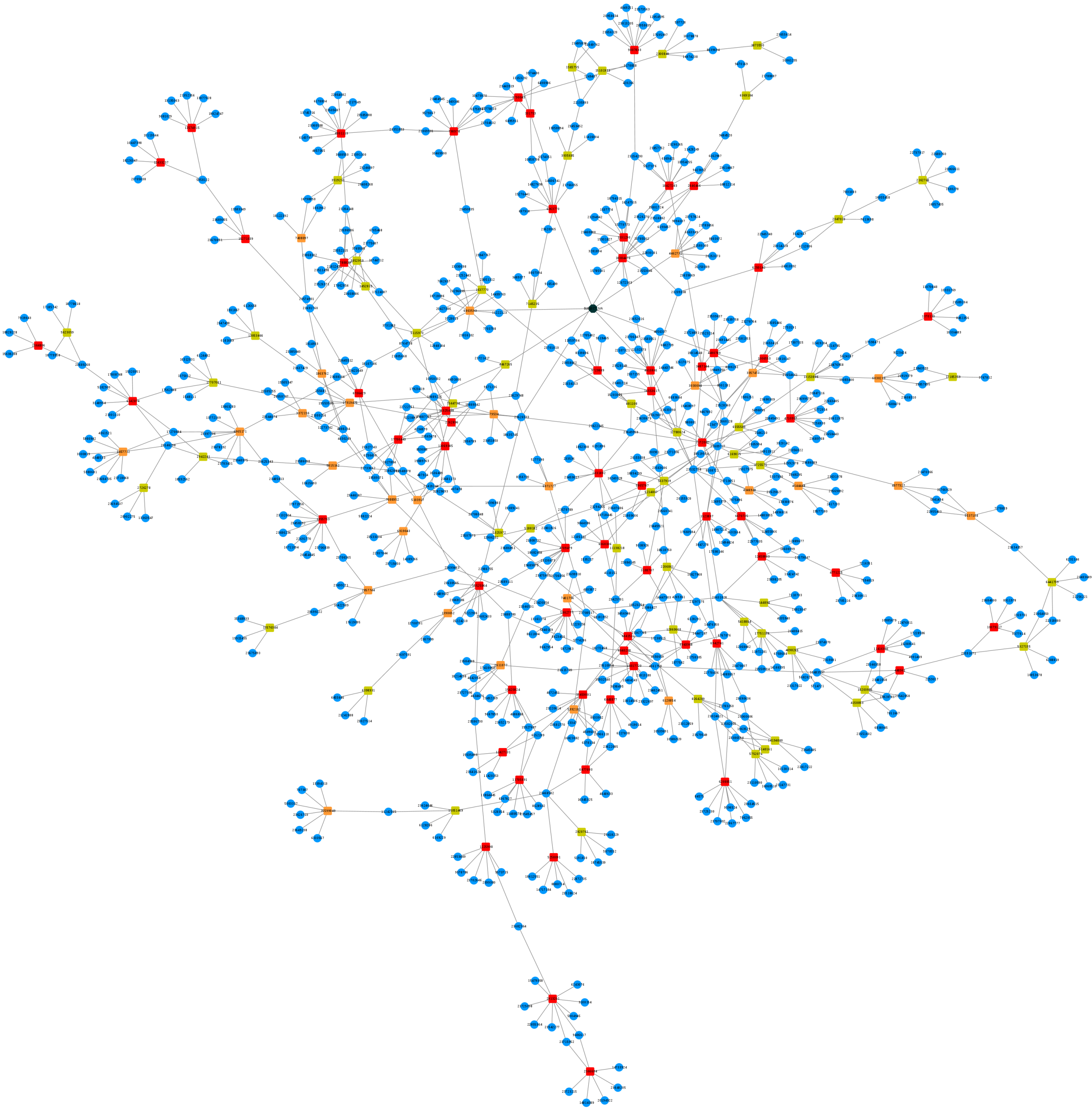


Figure 3.12: Enlarged SVN with accidents involving the reported car plates. Rectangular nodes are accidents while circular nodes are subjects. Accidents are: in red if they have been assigned a “high” level of anomaly; in orange if they have been assigned a “medium” level of anomaly; in yellow if the have been assigned a “low” level of anomaly according to the integrated indicator.

to be very effective in uncovering the anomalous patterns between subjects in the bipartite complex system *subjects-accidents* and between vehicles in the bipartite complex system *vehicles-accidents*. We construct an integrated indicator that synthesizes the information at node and system/network level to define a level of statistical anomaly of car accidents, and so subjects and vehicles linked to them. Moreover, we showed that the introduction of the SVN improves the ability of the model to detect frauds with respect to the case where only the score AIA is considered. Based on the evidence that emerges from the new tool, IVASS will inform all the competent authorities, police, prosecutor offices, eventually restraining fraudulent activities and improving the efficiency of the car insurance market in Italy.

3.5 Future research: triplets tests and recommendation methods for fraud detection

Triadic closure is a social mechanism that lies on the more fundamental concept of homophily, is also relevant for frauds [A. Rapoport, Bulletin of Mathematical Biophysics 15(4), 523-533 (1953)]. Indeed, triadic closure represents a simple mechanism through which fraudsters may learn to collaborate with each other. Let's suppose that fraudster A cooperates, separately, with fraudster B, and fraudster C, and nonetheless, B and C don't even know each other. Triadic closure suggests that the presence of A as a common associate provides the *opportunity* (that B and C come to know each other), the *trust* (due to the common trust in A) and the *incentive* (A may want to perpetrate a fraud with both B and C together) to the possibility that B and C become associates (in frauds). Therefore, as a future research advancement the presence of a series of frauds in which the same subjects appear and the presence of triplets and triangles of cooperation should both be taken into account to spot potential frauds among car accidents.

Moreover, fraud detection activity can be perceived as a recommendation system task. In principle, it is possible to suggest or associate any accident with a list of other accidents based on their similarities. There are many algorithms that allow to construct recommendation lists, which are based on similarities between accidents or between people/vehicles involved in the accidents, or again, a hybrid version involving the two cases [198].

Table 3.8: Community network indicators

Dimensions	Indicators	Description
Size	Subjects	Number of subjects in the community
Size	Accidents	Number of accidents in the community
Size	Vertices	Number of vertices in the community
Connectivity, Centrality	Links	Number of links in the community
Connectivity, Centrality	Connectivity ratio	Average number of links per vertex.
Size	RSS	Average number of accidents per subject.
Size, Connectivity, Centrality	HK-core (α, β)	the generalization for bipartite networks of the K-core introduced by [167]. It carries information about the connectivity and dimension of communities. Depending on the importance we want to attribute to accidents and subjects, it involves 2 parameters, α and β: $HK(\alpha,\beta) = \max_{H,K}\sqrt{H^{\alpha}K^{\beta}}$. where H and K are positive integers and refer to the number of subjects and accidents respectively. HK-core is obtained after a "pruning" of the bipartite SVN, removing at each step all the accidents linked to less than H subjects and all the subjects linked to less than K accidents. We take the maximum of the weighted geometric average of H and K to describe how deep and connected a community is, with weights α and β. While β is fixed, we choose α such that the indicator depends on the network structure rather than on the macro-category of communities.
Robustness	R-mean	Average of R-score of the community.
Robustness	R-max	Maximum value of R-score of the community.

Table 3.9: Average, median, standard deviation (σ) and Fisher's skewness (γ) of the community network indicators according to the four different combinations of the macro-categories of community characterization, including also communities without any characterization and the overall case, that is, regardless of macro-category classification.
P=professionals; NP= non professionals; P-NP=professionals and non-professionals; $\overline{\text{P}}$-$\overline{\text{NP}}$=neither professionals nor non-professionals (communities characterized by time and/or space attributes); None = communities with no characterization.

	P(3.15%)	**NP(22.95%)**	**P-NP(0.06%)**	**$\overline{\text{P}}$-$\overline{\text{NP}}$(61.54%)**	**None(12.28%)**	**Overall**
Community indicators	**Mean - Median ($\sigma\backslash\gamma$)**	**Mean - Median ($\sigma\backslash\gamma$)**	**Mean - Median ($\sigma\backslash\gamma$)**	**Mean - Median ($\sigma\backslash\gamma$)**	**Mean - Median ($\sigma\backslash\gamma$)**	**Mean - Median ($\sigma\backslash\gamma$)**
N. accidents	58.55 - 22.00 (251.67\26.48)	2.34 - 2.00 (1.06\10.51)	45.35 - 2.00 (99.68\4.77)	3.01 - 2.00 (3.02\10.62)	3.02 - 2.00 (1.54\2.63)	4.64 - 2.00 (45.89\141.20)
N. subjects	6.26 - 4.00 (6.71\4.33)	2.07 - 2.00 (0.36\9.11)	10.06 - 2.00 (14.96\4.12)	2.21 - 2.00 (0.88\11.05)	2.32 - 2.00 (0.69\2.77)	2.32 - 2.00 (1.63\17.56)
RSS	6.80 - 4.00 (9.87\9.41)	1.12 - 1.00 (0.32\4.64)	2.87 - 1.00 (4.44\5.38)	1.28 - 1.00 (0.48\2.81)	1.26 - 1.00 (0.39\1.81)	1.42 - 1.00 (2.05\39.24)
Vertices V	64.81 - 27.00 (255.22\26.09)	4.41 - 4.00 (1.32\10.52)	55.41 - 4.00 (112.39\4.60)	5.22 - 4.00 (3.84\10.89)	5.35 - 4.00 (2.12\2.65)	6.96 - 4.00 (46.72\137.77)
N. links	123.30 - 44.00 (567.36\27.50)	4.77 - 4.00 (2.46\11.49)	97.31 - 4.00 (210.57\4.72)	6.18 - 4.00 (6.51\10.58)	6.24 - 4.00 (3.38\2.58)	9.61 - 4.00 (103.12\147.69)
Conn.	2.05 - 2.00 (0.09\5.34)	2.03 - 2.00 (0.15\14.96)	2.09 - 2.00 (0.20\4.62)	2.03 - 2.00 (0.15\6.34)	2.05 - 2.00 (0.16\4.49)	2.03 - 2.00 (0.15\8.41)
H	2.00 - 2.00 (0.10\11.96)	2.02 - 2.00 (0.17\14.81)	2.04 - 2.00 (0.22\6.34)	2.02 - 2.00 (0.17\7.11)	2.03 - 2.00 (0.20\6.08)	2.02 - 2.00 (0.17\8.79)
K	19.73 - 9.00 (40.93\9.48)	2.25 - 2.00 (0.66\5.62)	11.00 - 2.00 (42.31\11.80)	2.56 - 2.00 (1.02\4.32)	2.48 - 2.00 (0.81\2.38)	3.02 - 2.00 (7.99\46.77)
HK-core($\alpha=\beta=1$)	5.20 - 4.24 (3.53\2.94)	2.11 - 2.00 (0.26\3.35)	3.52 - 2.00 (3.12\5.48)	2.24 - 2.00 (0.37\2.29)	2.22 - 2.00 (0.33\1.62)	2.31 - 2.00 (0.89\11.95)
HK-core($\alpha=2.48;\beta=1$)	1.83 - 1.79 (0.22\0.89)	1.90 - 1.83 (0.17\3.08)	1.72 - 1.54 (0.24\1.80)	1.88 - 1.75 (0.19\1.87)	1.91 - 1.78 (0.19\1.45)	1.71 - 1.63 (0.18\5.16)
R-mean	1.30 - 1.29 (0.84\0.32)	0.35 - 0.00 (0.79\2.17)	0.54 - 0.001 (0.71\1.26)	0.70 - 0.001 (1.02\1.15)	0.61 - 0.00 (0.97\1.42)	0.63 - 0.001 (0.98\1.31)
R-max	2.10 - 2.24 (1.15\-0.33)	0.38 - 0.00 (0.83\2.12)	1.32 - 0.001 (1.52\0.55)	0.77 - 0.001 (1.09\1.05)	0.71 - 0.00 (1.06\1.16)	0.71 - 0.001 (1.08\1.18)

Chapter 4

Assessing the impact of the REF on scientific excellence in the UK

Abstract

The Research Excellence Framework (REF) is the main UK government policy on public research in the last 30 years. The primary aim of this policy is to promote and reward research excellence through competition for scarce research resources. Surprisingly, and despite the severe criticisms, little has been done to systematically evaluate its effects. In this paper we evaluate the impact of the REF 2014.

We exploit a large database that contains all publications in Economics, Business, Management and Finance available in Scopus since 2001. We use a synthetic control method to compare the performance of each of the 85 universities from the UK with a counter-factual similar unit in terms of past research constructed using 121 US universities. Among other interesting insights, we find an overall increase of the number of published papers, but the effect reverses when we focus on per-capita productivity. The proportion of papers published in a 3, 4* or 4** journal had a significant increase in 2012 but the proportion of articles published by Economics Department decreased. The twenty-four universities belonging to the Russell group reported almost only benefits, and when negative effects took place, they were the units that suffered the least.*

4.1 Introduction

4.1.1 The politics of the REF (ex RAE) in the UK

The main government policy on public research in the last 30 years has been the university RAE, formally known as Research Selectivity Exercise, then as Research Assessment Exercises, and now as REF. The RAEs produce comparable ratings of research performance of all the departments of all the universities and public research institutions in the UK. Based on the results of this assessment, undertaken every three to seven years (1986, 1989, 1992, 1996, 2001, 2008, 2014), core government funding for the subsequent years is allocated. But, besides universities' funding, the RAE results also influence the UK departments.

The primary aim of this policy is to promote and reward research excellence through competition for scarce resources. The RAEs facilitate the concentration of research funding in better-performing institutions [101, 24, 100]. But, even after several modifications, the RAEs are still receiving severe criticisms, both in terms of the benefits obtained as well as on the costs incurred [132]. Some commentators question whether they are really fostering high quality research (e.g. the University and College Union). Others claim that, as they are currently designed, the RAEs favour the "old", large universities and those represented on the decision panels [57, 162, 24, 43] and also show that panels were biased in favour of the Russell-Group Universities [177]. Critics also complain that the RAEs have substantial costs of preparation and submission and even more costly side-effects or indirect costs [94]. Some claim, for example, that the RAEs have distorted universities' hiring decisions, especially in the years around RAE submission deadlines [99, 121].

Surprisingly, and despite the severe criticisms, little has been done to systematically evaluate the effects of such an ambitious policy. Probably because of lack of data, most existing analyses are descriptive, bibliometric or apply sociological perspectives [114, 57, 24, 79, 141, 171, 177]. More recent papers use the output submitted to the REF to create a ranking of economics journals [104] or to predict the results of the next REF using departmental h-index [147].

Among the few quantitative studies, [191] analyse thirty years of UK aggregate publication data, identify three structural changes at the national level, and relate one of them to one RAE. At the international level, [76] provide evidence that country-level incentives rewarding research performance in the OECD lead to more submissions and publications in the academic journal Science. [103] presents a review of fourteen performance-based research funding systems (PRFSs) policies in different countries (including the RAE), stating that while the aim of these policies is to increase excellence of a nation's research, it may compromise other important values such as equity or diversity.

This paper investigates if the REF of 2014 increased research in economics, business and management in terms of quantity and quality, both in total and on a per-capita basis. To analyse the impact of the 2014 REF on academic performance, we make use non-UK departments' exposure to the 2014 REF using US economics departments and business schools. To do so, we apply the synthetic control method (SCM) which allows the creation, for each university in the UK, of a comparable research unit combining a set of US universities.

Our results indicate that the REF increased significantly the overall number of publications in the UK in the years 2012 to 2015. In terms of quality—number of publications in journals graded as 3* and 4* in the Academic Journal Guide (AJG)[1]—it also increased significantly from 2013 onwards. We analyse the effect of the REF 2014 for the Russell Group Universities, per author, proportion of publications in Economics and Econometrics journals and publications in Economics only. These extensions show that the REF had a more positive impact in terms of quantity and quality for the Russell Group Universities and that there was a negative and significant effect on the proportion of publications in Economics and Econometrics journals graded as 3*, 4*, 4** for 2014 and 2015. Moreover, results also show a negative and signifcant effect of the REF on the number of publications in journals per author.

4.2 Data

This research is possible thanks to the Scopus Database from which we were authorised to download all articles published by all the academics in the UK and in the US, for the last 15 years (2001-2015), in the fields of Economics and Econometrics and Business and Management.

Our sample includes all published articles by authors affiliated to universities in the UK that submitted their research to the Economics and Econometrics REF Panel (Panel 18) and to the Business and Management Panel (Panel 19) in 2014. This amounts to 103 UK universities. In order to create a control group not exposed to the REF 2014, we select the publications of the top 25% Departments of Economics and the top 25% Business Economics (in terms of *RePEc* number of publications) in the US, which amounts to 135 US universities. Further, we only include publications of universities that published an average of at least 10 papers in the pre-treatment period 2001-2007. As a result, our final dataset includes articles of 121 US universities and 85 UK universities.

The definition of our output variables is reported in Table 4.1. The first measure refers to the total number of publications, the second to their quality. We use the classification of scientific journals by the Academic Journal Guide

[1] http://www.CharteredABS.org/academic-Journal-Guide-2018.

(AJG) for 2018 as a proxy of the quality of published papers. The possible values of this classification can be 1* (worst), 2*, 3*, 4* and 4** (most influential journals). We assume - and believe it is reasonable - that the classification of journals remains almost invariant over time.

Table 4.1: Description of the research output measures considered in the analysis.

Number of publications in journals	Count the number of unique publications by institution and year in only scientific journals, and so, after deducting all the publications in books and/or conferences.
Number of publications in a 3*, 4*, and 4** journal	Papers published in journals with an Academic Journal Guide (AJG) 2018 grade of 3*, 4*, and 4**, by institution and year.

In Figure 4.1, we present the total number of research papers published and the proportion of papers that are 3* and 4* from 2001 to 2015 for both the UK and the US.

Figure 4.1: Descriptive analysis, comparing UK and US universities.

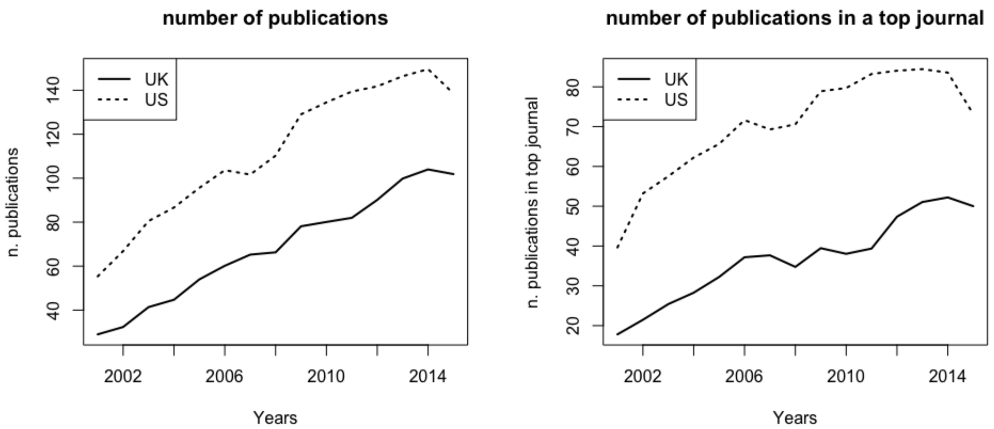


Figure 4.1 reveals that, on average, the net number of publications increases over time for both UK and US units. The proportion graded as 3*, 4* and 4**, show a slightly more volatile trajectory.

Tables A1 and A2 of Appendix A present the list of universities included for the US and UK, and the summary statistics of the outputs along with the average number of co-authors per article, number of affiliated authors and number of papers per author, by university and country, both for the pre and the post treatment periods. We sort the university in decreasing order according to the average number of publications in column (1).

4.3 Methods

4.3.1 The Synthetic Control Method (SCM)

To estimate the impact of the REF 2014 on the Economics, Econometrics and Business research output, we use the Synthetic Control Method (SCM). This method was introduced by [2] to evaluate the effect of an intervention on a unit (region) in terms of a certain output of interest by comparing it to that of an artificial unit created as a convex combination of multiple untreated units. [2] proposes that a convex combination of some untreated units (controls) allows to reproduce the characteristics of the treated one better than when using just a single control unit. The artificial comparator group is chosen taking into account a series of covariates which have good predictive power over the pre-intervention period. The artificial or counterfactual unit provides information on what the treated unit would have experienced in absence of the intervention. Thus, the comparison takes into account the difference, which we denote by $\hat{\alpha}_t$, between the actual values of the outcome, Y, for the treated unit and the artificial one, ${Y_t}^*$, i.e. $\hat{\alpha}_t = Y_t - {Y_t}^*$. Moreover, unlike the difference-in-differences model, which has been used many times in the literature for comparative case studies, the SCM allows for the presence of unobserved confounders whose effects can vary over time, and , also, it does not rely on the *parallel trend assumption* [3]. Indeed, [4] states that, intuitively, only units that are alike in both observed and unobserved determinants of the outcome variable as well as in the effect of those determinants on the outcome variable should produce similar trajectories of the outcome variable over extended periods of time. One limitation of the SCM is that traditional statistical inference is inappropriate when there are small number of treated and control units and the fact that units are not sampled probabilistically [29].

Because the REF 2014 is an intervention that affects all UK universities submitting to the Economics, Econometrics and Business REF panel, we apply a variation of the original SCM designed to the case of multiple treated units as opposed to one [7].

Our control group is made of US universities, not exposed to the REF 2014 by definition. The treatment period is from January 1 of 2008 to the of December of 2014, which is the deadline of the submission to the REF panels. The modified SCM allows us to create as many artificial units combining US universities as UK universities there are, i.e. the SCM creates a control artificial university for each UK university.

To create the artificial control group for each UK university we use information on each outcome variable(s) in Table 4.1, one at a time. The pre-treatment period covariates used to run the matching algorithm are the means over all pre-treatment period (2001-2007) of: the number of publishing authors; the total number of publications; the total number of publications in a 3*, 4* or 4** journal; the total number of publications in a 4* journal; and the outcome in interest. Also, we use the last value of the outcome in the pre-treatment period (2007).

Therefore, the SCM follows an iterative two-step optimization process:

(i) in the *inner optimization* step, we estimate the weights that minimize the distance between treated and untreated units' covariates over the pre-treatment period

$$\mathbf{w} = \arg_{\mathbf{w}} \min ||\mathbf{X}_1 - \mathbf{X}_0\mathbf{w}||_{\mathbf{V}} = \arg_{\mathbf{w}} \min \sqrt{(\mathbf{X}_1 - \mathbf{X}_0\mathbf{w})'\mathbf{V}(\mathbf{X}_1 - \mathbf{X}_0\mathbf{w})} \quad (4.1)$$

where $\mathbf{X}_1$ is the matrix containing the values of the covariates over the pre-treatment period for the treated units; $\mathbf{X}_0$ the same but for the untreated units; $\mathbf{w}$ is the vector of optimal weights to create a convex combination of untreated units; and $\mathbf{V}$ is a positive-definite and diagonal matrix, which is initialized at the beginning of the iterative algotithm and allows to assign some weights to the variables used in the optimization process;

(ii) in the *outer optimization* step we use the current optimal value of $\mathbf{w}$ to estimate $\mathbf{V}$. Specifically, matrix $\mathbf{V}$ is chosen to be the one minimizing the Mean Square Predictive Error (MSPE) for the outcome over the pre-treatment period. Thus, denoting the pre-treatment period by $(1, 2, \ldots, T_0)$, where T_0 is the time prior to intervention, and by Y_{it} the value of the outcome for the treated unit i at time t

$$\begin{aligned} Y^*_{it} &= \textstyle\sum_{j\in \text{untr.}} w_{ij} Y_{jt} && (4.2) \\ \text{MSPE}_i &= \tfrac{1}{T_0} \textstyle\sum_{t=1}^{T_0} (Y_{it} - Y^*_{it})^2. && (4.3) \end{aligned}$$

Steps **(i)** and **(ii)** are repeated iteratively until convergence.
To implement the SCM to estimate $\mathbf{w}_i$, $\forall i = 1, 2, \ldots, N_T$, where N_T is the number of treated units, we use the R packages *Synth* and *improveSynth*. The estimated coefficients, $\mathbf{w}_i$, are reported in Table A5.

4.3.2 Robustness check: placebo based p-values

Once all the effects have been estimated, $\alpha_{it} = Y_{it} - Y^*_{it}$ $\forall i = 1, 2, \ldots, N_T$, where N_T is the number of treated units and $t = 2008, 2009, \ldots, 2015$, we check if these differences between the actual and counterfactual values are due to chance or, actually, to a statistically significant effect of the REF2014. We conduct exact inference on these parameters, running the so-called placebo tests [3].

Performing placebo tests allows us to construct the null distributions of the placebo effects against which we compare or actual estimates. To do so, we use our untreated units as if they were the treated ones and apply SCM to them. So, eventually, we obtain 121 placebo patterns of gaps over time. If the REF did not have any effect on UK universities, we would expect the placebo effects to be similar to the ones computed for the treated units.

Then, we conduct a two-sided hypothesis test on the placebo effects. The p-values for a generic treated unit i at time t can be calculated as

$$p_{it} = \frac{\#\{|\alpha_{it}^{PL}| \geq |\hat{\alpha}_{it}|\}}{N_{PL}} \quad \forall i = 1, 2, \ldots, N_T,\ t = T_0 + 1, \ldots, T \tag{4.4}$$

where N_{PL} is the number of generated placebo effects.
Between all placebo patterns, we remove from the computation of p-values the ones that have a pre-treatment MSPE greater or equal than twice that of the treated unit [3].

4.3.3 Average Treatment Effect on the treated

To calculate the overall effect that REF had on the whole treated group, at the system level, we obtain the so-called Average Treatment effect on the Treated (ATT).
As suggested by [7], a fit-weighted ATT can be computed as:

$$\hat{\text{ATT}} = \frac{\sum_{i \in Treat} \left(\frac{\sum_{t=T_0+1}^{T} \hat{\alpha}_{it}}{\hat{\sigma}_i} \right)}{\sum_{i \in Treat} \frac{1}{\hat{\sigma}_i}} \tag{4.5}$$

where $\hat{\sigma}_i = \sqrt{\frac{\sum_{t=1}^{T_0} \hat{\alpha}_{it}^2}{T_0}}$, that is, the RMSPE over the pre-treatment period, and $\hat{\alpha}_{it}$ is the estimated effect for the treated unit $i = 1, \ldots, N_T$ at time $t \in [T_0 + 1, \ldots, T]$ where, again, N_T is the number of treated units and $[T_0 + 1, \ldots, T]$ the post-treatment period.

Equation 4.5 describes a weighted average of the effects using the inverse of the RMSPE over the pre-treatment period as weights. This implies that universities with a better matching have a higher impact on the estimate of ATT which provides an unbiased estimate of ATT. To compute the p-value, again, a null distribution of placebo ATT effects is needed. [7] suggest forming 5,000 placebo treatment groups of size N_T from the N_C controls.

4.3.4 Quality of the matching

Although there is currently no consensus on what constitutes a ‘good fit’ or how to judge similarity between treated and control units [29], most of the works making use of SCM consider the RMSPE of the estimates within the two groups of units in the pre-treatment period to assess the quality of the matching. Therefore, to assess the goodness of the matching, we consider the proportion of placebos that have a pre-treatment RMSPE at least as large as the average RMSPEs of the treated units in the pre-treatment period. If placebo RMSPEs are basically smaller than those of the treated, then it means that the control group is not able to properly replicate the patterns of the treated units. Moreover, we assume that control units are somehow similar, in the sense that we should not expect their RMSPEs to be too high. Therefore, if the control group can reasonably reproduce the treated units, we expect the two RMSPE distributions to be very close one another. On the other hand, if that value is significant (small proportion of placebos with pre-treatment RMSPE at least as large as the average RMSPEs of the treated units), then RMSPEs of the treated are higher and there is concern about the quality of the matching.

4.4 Results

Below, we present the results for our two outcomes of interest: the total number of publications and the total number of papers published in top journals (3*, 4* and 4**). We show the ATTs.

We introduce our results in a variety of ways so that we compare the impact of the REF2014 on the number of publications and publications in top journals for different types of universities and for different fields.

We compare the results for the Russell group universities to the non-Russell group ones. We also distinguish the universities that submitted to the Economics and Econometrics panel of the REF 2014 and compare them to the ones that did not. We also examine the impact in terms of number of publications in journals (all ranks and top ranked) in Economics and Econometrics and in Finance and Management journals (see Table A7 of Appendix Appendix A).

The goodness of fit of our estimates is discussed at the end of the section. The weights, w_i, that matching algorithm gives US universities to create the artificial control group for each UK university is included in Table A5 in the Appendix.

4.4.1 ATT for the number of publications

Table 4.2 shows the estimated ATT on the number of publications in scientific journals associated to the REF 2014 by post-treatment year, by university and overall.

The overall results are in the second-last and last columns, which contain the universities' ATTs across the post-treatment period (2008 to 2014) and the ATT including year 2015 (2008-2015), respectively. Overall, the ATT aggregated for all universities is positive and about of 150.74 publications. In the Appendix A (from page 134) we report the graphs of the estimated effects for each of the UK universities.

Universities in the Russell group (top panel) experience positive or negative effects in specific years but the aggregated effects (in the last two columns) are not significantly different than zero for all universities. For instance, this is the case of Cardiff or Newcastle Universities. However, overall results for the Russell group show that they experienced a positive effect on publications due to the REF 2014 as the average up to 2014 is of a significant increase of 11.42 and up to 2015 of 15.16.

Within this group, the most exceptionally striking results are for the University of Cambridge - as it has positive ATTs almost all years and an overall average above 85 publications. Oxford University has more variability but has 96.66, 83.89 and 96.04 the last three years and ATTs of 42.22 more publications up tp 2014 and 48.95 up to 2015. In the case of Nottingham, nevertheless, the effect is negative for almost all years and for the average over the post-treatment period.

The Non-Russell group has a non significant overall average treatment effect. For this group, the effects are of smaller magnitude than the Russell group: Bournemouth University experienced a significant overall effect (35.95 and 39.37 up to 2014 and 2015, respectively), and so did City, University of London (32.25 and 39.86), University of Essex (22.29 and 28.36). Instead, Glasgow Caledonian University and University of Aberdeen suffered a significant reduction in the number of publications (over 20 in both cases).

Comparing Russell and non-Russell group, results show a significant difference in ATTs between the two groups of about 10.28 and 12.44 publications per year in favour of the Russell group (up to 2014 and 2015, respectively). To assess if the effect of the REF has been significantly different for universities belonging to the Russell group versus not we ran placebo sampling tests. To do so, we create a null (placebo) distribution against which we test the point estimate of the difference, 10.28 and 12.44 reported in Table 4.2 .

To generate the null distribution for the difference, we construct two groups of universities by selecting 24 (number in the Russell group) and 61 (number of the non-Russell group) units randomly from the original full set of universities. We calculate the ATT for each group, overall and by year. We repeat this process 100,000 times and obtain the null (placebo) distribution set of the difference. The point estimate - in the second-last row- is 10.28 and 12.44 - in the last row- and significant at a level of 1%. The same approach is used to test the difference year by year.

Moreover, the same approach was used to compare universities that left the Economics and Econometrics panel and the ones that remained. Results show a significant and positive difference in ATTs between the two groups of about 11.1 and 11.9 publications per year in favour of the universities that remained in the Economics and Econometrics panel (up to 2014 and 2015, respectively). As before, we run placebo tests to associate these figures with p-values.

4.4.2 ATT for the number of papers in a 3^*, 4^*, and 4^{**} journals

In Table 4.3 we present the estimated ATTs associated to the REF 2014 by post-treatment year, by university and overall on the number of publications in scientific journals which quality is ranked 3^*, 4^*, and 4^{**}. Table 4.3 shows that the overall ATT is 24.26 up to 2014 and 49.38 up to 2015, which are statistically significant. This effect is lower than our previous finding in number of publications. With respect to yearly ATTs, it is negatively significant for the year 2008, and positively significant for years 2011, 2013, 2014, and 2015.

Regarding the Russell group universities (top panel), there are only four universities that experience a positive aggregated effect up to 2014 or 2015, i.e. University of Oxford (48.44 and 51.86 up to 2014 and 2015, respectively), University of Warwick (23.74 up to 2015), Imperial College London (23.02 and 21.13 up to 2014 and 2015, respectively) and University of Cambridge (22.62 and 23.10 up to 2014 and 2015, respectively). However, overall results for the Russell group show that they experienced a positive effect on 3^*, 4^*, and 4^{**} journal's publications due to the REF 2014 only for the aggregated figure until 2015, 8.61.

The Non-Russell group has a non significant overall average treatment effect, which goes in line with the previous on the total number of publications analysis. For this group, only Lancaster University (18.56 and 25.48 up to 2014 and 2015, respectively) and University of Kent (21.55 and 26.96 up to 2014 and 2015, respectively) experienced a positive and significant overall effect. For the rest of universities of this group, even if the aggregated ATT

Table 4.2: ATT for the REF 2014 by post-treatment year on the number of publications in scientific journals.

Russell group	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**ATT$_{-2015}$**	**ATT**
Cardiff University	3.31	37.10***	24.68	-30.54*	-0.19	20.01	3.56	13.04	8.27	8.87
Imperial College London	-9.62	-3.18	-16.96*	-3.32	1.02	2.63	15.47	5.03	-1.99	-1.11
King's College London	21.49**	3.83	3.30	-17.36	8.45	31.54**	37.88**	51.96***	12.73	17.63
LSE	9.28	-9.46	14.15	45.35**	30.15	58.20**	52.91*	69.21**	28.65*	33.72*
Newcastle University	-19.00**	11.01*	-13.38*	-10.36	4.26	37.81****	25.86*	61.72****	5.17	12.24
Queen Mary University of London	12.25	3.61	10.41	-1.34	17.43	34.36*	41.11*	33.42	16.83	18.90
Queen's University Belfast	10.35	2.70	-0.70	-7.57	6.06	-9.16	-0.99	6.34	0.09	0.87
University College London	13.06	-13.18	-34.99*	-25.67	13.39	52.71**	86.44***	86.74****	13.10	22.31
University of Birmingham	16.13*	18.56*	10.39	25.87*	35.76**	22.61*	10.45	31.07**	19.96*	21.35*
University of Bristol	1.50	9.86	18.06	8.47	-4.70	18.12	46.57*	28.25	13.98	15.77
University of Cambridge	45.26**	73.51****	84.38***	94.09***	95.63****	109.08***	92.72***	114.10****	84.95****	88.59****
University of Durham	-13.40	-36.20***	-32.60**	-26.99	20.70	14.87	25.92	57.33**	-6.81	1.20
University of Edinburgh	-3.53	19.91	-0.94	1.26	-4.79	25.98	35.91*	47.64**	10.54	15.18
University of Exeter	-31.96**	-4.15	-19.18	1.14	18.64	29.11	22.11	29.53	2.24	5.65
University of Glasgow	15.95***	-6.70*	-9.60*	1.47*	22.49**	43.31****	26.68**	34.42***	13.37	16.00
University of Leeds	-6.14	-24.47	-23.96	-28.06	-5.03	24.00	39.89	64.04*	-3.39	5.03
University of Liverpool	4.58	-0.33	14.26	16.30	12.18	49.96**	53.62**	66.36**	21.50	27.11*
University of Manchester	53.75**	29.96*	-37.79*	-52.62**	4.86	-10.35	2.87	-7.48	-1.32	-2.09
University of Nottingham	-9.51	-21.44	-83.84**	-84.64***	-33.15	-64.39**	-67.90**	-78.55**	-52.12**	-55.42**
University of Oxford	18.58	45.56**	10.02	-6.59	47.45*	96.66***	83.89**	96.04****	42.22**	48.95**
University of Sheffield	18.94	-4.43	-3.24	-8.13	-18.60	-24.11	8.87	54.15*	-4.38	2.93
University of Southampton	-7.08	3.86	8.97	20.93	61.94***	77.09**	89.63****	92.50****	36.47**	43.48**
University of Warwick	15.89	16.19	-53.1****	22.48	26.33	28.79	44.32*	26.15	14.41	15.88
University of York	-7.74	-12.63	9.85	-6.38	-20.25	12.45	21.73	8.70	-0.42	0.71
Total Russell group	6.34	5.81	-5.07	-3.01	14.17*	28.38*	33.31**	41.32***	11.42*	15.16**
Non-Russell group	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**ATT$_{-2015}$**	**ATT**
Aberystwyth University	-6.32	-4.56	-3.17	-12.04	5.86	-3.49	9.62	-4.87	-2.01	-2.37
Aston University	3.62	-3.75	2.15	12.08	-14.06	25.15	-3.45	14.94	3.10	4.58
Bangor University	-0.92	13.00	8.60	16.93	25.88*	23.29	35.47*	21.61	17.98	17.46
Birkbeck College	-5.23	2.40	-18.59	-10.86	-23.69	-3.51	-19.69	-9.22	-11.31	-11.05
Bournemouth University	1.31	23.74*	22.04	27.28	36.79*	66.52**	74.02**	63.35**	35.95**	39.37**
Brunel University London	-2.28	22.73*	-45.85**	-11.24	-7.21	-4.55	-39.58*	-2.71	-12.57	-11.33
City University London	24.89	22.99	4.16	21.05	23.64	53.42**	75.63**	93.08****	32.25*	39.86**
Coventry University	-4.88	-14.04	0.30	-6.03	-12.80	-18.26	12.74	7.72	-6.14	-4.40
Cranfield University	0.58	-9.19	-24.88	-2.34	-21.57	-9.65	-35.44*	-4.83	-14.64	-13.41
De Montfort University	1.82	-10.00	10.13	-8.03	4.37	20.36	9.59	-13.19	4.03	1.88
Edinburgh Napier University	-12.37	-13.90	-4.46	-9.01	-18.26	-35.70	-7.33	-7.75	-14.43	-13.59
Glasgow Caledonian University	-22.41	-33.59**	-17.66	-15.20	-41.89**	-21.30	-22.04	-4.45	-24.87*	-22.32
Heriot-Watt University	3.54	3.23	-11.68	1.40	-18.48	5.26	4.30	36.32	-1.77	2.98
Keele University	-19.70	-1.90	-0.37	-11.31	-11.32	-26.66	-7.07	-8.80	-11.19	-10.89
Kingston University	-11.64	3.89	3.86	10.13	20.90*	-1.12	23.17	36.47*	7.02	10.70
Lancaster University	2.77	10.00	-21.61	8.52	17.45	44.28*	42.50	72.50**	14.84	22.05
Leeds Beckett University	-5.75	5.45	15.22	4.02	16.71	13.82	21.51	21.77	10.13	11.59
London Business School	-29.63*	-11.04	-7.63	-29.66	-17.30	-25.85	-25.25	-51.35*	-20.91	-24.71
London Metropolitan University	-0.01	5.14	11.39	20.83	3.59	-1.44	-12.00	-18.23	3.93	1.15
London South Bank University	-4.47	-4.19	-4.40	-10.45	-25.13	-29.65	-24.58	-12.49	-14.69	-14.42
Manchester Metropolitan University	-0.91	13.74	-16.76	-17.26	-23.26	0.81	-31.58	-6.57	-10.74	-10.22
Middlesex University	-15.79	-7.73	3.33	8.66	8.14	24.47	6.18	57.04**	3.89	10.54
Nottingham Trent University	6.10	20.18*	16.44	-2.19	20.98*	8.40	23.00	29.96*	13.27	15.36
Open University	2.61	23.72*	18.57	18.38	8.92	22.04	29.30	26.55	17.65	18.76
Oxford Brookes University	16.74	-0.23	-4.77	-1.01	8.40	21.98	10.09	18.75	7.31	8.74
Robert Gordon University	5.89	-1.27	-4.14	-4.63	-2.64	3.33	6.21	3.64	0.39	0.79
Royal Holloway, University of London	-6.35	13.28	-4.74	1.63	11.12	36.74**	21.88*	25.47**	10.50	12.38
Sheffield Hallam University	-17.27*	1.62	-10.04	-1.04	9.69	-1.38	3.99	13.20	-2.06	-0.15
Staffordshire University	-12.86	-16.01	-8.08	-15.01	-6.86	0.07	0.00	-7.58	-8.39	-8.29
Swansea University	3.97	-13.18	-4.49	-12.41	-29.91*	-31.72*	-26.54	-30.44	-16.32	-18.08
University of Aberdeen	-22.34*	-14.42	-12.64	-21.13	-36.26**	-10.56	-38.21**	-55.35***	-22.20*	-26.36*
University of Bath	17.84	18.61	4.78	7.92	-2.05	29.89*	24.64	37.26*	14.51	17.36
University of Bedfordshire	-11.62	-7.21	-6.00	-0.21	-2.30	8.40	3.21	5.98	-2.24	-1.22
University of Bradford	-9.61	-10.96	-26.95	-32.68	-28.75	-43.46*	-40.35	-39.52	-27.54	-29.03
University of Brighton	-7.57	3.62	-2.46	18.57	-0.32	13.81	18.55	30.05	6.31	9.28
University of Central Lancashire	-6.36	-1.13	1.21	0.28	17.25	11.94	21.37	7.23	6.36	6.47
University of Dundee	-14.35	-18.06	-29.55*	-23.43	-11.63	-19.64	-31.30	-19.95	-21.14	-20.98
University of East Anglia	6.17	-6.02	21.25	5.88	26.89*	30.09*	47.94**	40.55*	18.88	21.59*
University of East London	-12.63	-7.11	0.82	4.93	9.67	13.23	-0.01	2.15	1.27	1.38
University of Essex	3.99*	19.36***	16.14**	31.87***	25.50**	14.09*	45.16****	70.84****	22.29***	28.36****
University of Greenwich	-7.50	-8.11	-4.28	-13.23	5.92	19.81	21.50	8.69	2.01	2.85
University of Hertfordshire	0.49	3.65	2.26	7.79	20.30	13.47	19.78	3.68	9.67	8.92
University of Hull	12.37	-1.80	4.10	-3.09	21.63	10.04	25.04	52.36*	9.75	15.08
University of Kent	26.09*	17.15	24.14	27.14	22.65	56.89**	37.83*	70.97**	30.26**	35.36**
University of Leicester	15.44	32.39**	6.02	8.77	20.43	14.82	-4.17	29.39	13.38	15.38
University of Northumbria at Newcastle	-3.27	8.11	7.84	2.37	24.31	29.15	45.40*	62.18**	16.27	22.01
University of Plymouth	-4.58	-1.20	3.34	-4.11	17.92	21.22	52.61**	19.36	12.17	13.07
University of Portsmouth	-4.67	-14.90	-0.95	-18.00	-9.62	-3.60	-8.42	26.33	-8.59	-4.23
University of Reading	2.69	12.90	-21.00	-18.42	-31.09	-3.15	5.87	26.94	-7.45	-3.15
University of Salford	-1.89	5.36	-2.30	-2.35	-3.87	-6.21	-20.29	-25.35	-4.50	-7.11
University of South Wales	5.59	3.15	-11.43	-1.04	-0.30	-2.22	-1.37	-8.48	-1.09	-2.01
University of St Andrews	2.19	35.06****	23.21*	11.00	7.67	14.89	25.33*	45.08***	17.04	20.55
University of Stirling	18.71*	9.24	-6.99	1.10	-6.36	17.60	14.82	16.34	6.87	8.05
University of Strathclyde	15.99*	-16.61*	-2.94	15.82	11.62	15.67	15.69	-5.12	6.26	7.89
University of Sunderland	-19.67	-18.00	-12.00	-26.00	-15.17	-13.33	-7.00	-11.34	-15.88	-15.31
University of Surrey	0.89	-19.49	-12.89	-20.04	-11.48	-42.96*	-3.76	15.67	-15.67	-11.75
University of Sussex	9.64	-18.32*	-22.92	-7.70	-4.59	12.90	31.95*	52.35***	0.13	6.67
University of the West of England, Bristol	6.89	-4.26	6.82	15.78	10.29	18.38	17.61	26.58	10.21	12.26
University of Ulster	-0.71	-33.55***	-6.35	-22.03	-34.43**	5.82	-41.76**	-25.03*	-18.99	-19.75
University of Westminster	-7.99	10.21	15.06	-4.54	25.74	-2.88	19.12	10.61	7.81	8.16
University of Wolverhampton	-3.84	5.72	-8.36	-22.39	-11.64	-26.11	11.12	10.46	-7.92	-5.63
Total Non-Russell group	-1.61	0.39	-2.46	-1.80	0.09	5.79	7.58	13.78*	1.14	2.72
Russell group - Non-Russell group	7.96**	5.42	-2.61	-1.20	14.07***	22.58****	25.73****	27.54****	10.28***	12.44***
Remainers - Leavers	10.44***	10.77***	-0.51	7.03*	12.42***	18.16***	19.32***	23.34***	11.1***	11.9***
yearly ATT	11.62****	0.33	-3.52	8.34	22.47****	36.01****	31.11***	44.35****	106.38***	150.74****

Notes The last two columns contain each university ATT overall the post-treatment years (2008-2014) and adding 2015 (2008-2015), respectively. The last row of the table contains the overall yearly ATT for each year, and note that there are two panels, the top displays the results and subtotal for the Russell Group universities and the panel below for the non-Russell group ones. The last value at the bottom-right corner is the overall ATT for all universities included. The third and second-last rows contain the differences of means of ATTs respectively between the Russell and non-Russell groups and between Remainers and Leavers. Values are marked by *, **, ***, **** if they are significant at a level of, 0.10, 0.05, 0.01 or 0.001, respectively.

is negative in most of the cases, it is not significant. It is only significant for some specific years, for instance, University of Reading for the year 2008, 2009, 2010, 2011 and 2014.

Comparing Russell and non-Russell group, results show a significant difference in ATTs between the two group of about 6.79 and 7.90 publications per year in favour of the Russell group (up to 2014 and 2015, respectively). Even if the ATTs effect between groups is smaller than in the number of publications, it is significant and shows that the Russell group universities benefits more of the REF 2014 than the Non-Russell group. Also, we find a positive and significant difference in ATTs between the universities that remained in the Economics and Econometrics panel against the ones that left, with a difference of 8.9 and 8.8 up to 2014 and 2015, respectively.

4.4.3 Goodness of Fit Measures

As a measure of fit, we compare the distribution of the RMSPEs in the pre-treatment period between treated and untreated units. The more overlapped the two distributions are, the better the overall matching. For the number of publications, the proportion of placebo RMSPEs greater than the average of treated RMSPEs is equal to 0.36 denoting that, overall, the matching is acceptable. Nevertheless, looking at particular universities, Harvard has a very high value of RMSPE (73.8), confirming that this university stands out and is not comparable to any other university in terms of its history in number of publications.

For the number of publications in a 3^*, 4^*, 4^{**} journal the matching is also acceptable, with a p-value of 0.35. Since the outcome variable to be matched in the SCM optimization process illustrated in section 4.3 is different, the set of matching coefficients are allowed to be different from the previous ones. (Table A6 of Appendix A).

Figures 4.2 and 4.3 present respectively the graphical comparison between the distributions of the pre-RMSPE of the number of publications in journals and of the number of publications in top journals. As can be appreciated in the figures, the overlapping of the two distributions is quite good in both cases, eventually leading to a not significant p-value.

4.4.4 Other outcomes

In the sections that follows we present the estimates of the impact of the REF 2014 on additional research outcomes such as the number of publications in Economics and Econometrics in journals graded $3^*, 4^*$, or 4^{**}; the number

Table 4.3: ATT for the REF 2014 by post-treatment year on the number of publications in a scientific journals graded as 3*, 4* or 4**.

Russell group	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**ATT$_{-2015}$**	**ATT**
Cardiff University	0.89	8.95	-1.92	-10.01*	27.94**	16.88*	9.43	36.00***	7.44	11.01
Imperial College London	8.48	-8.07	3.97	5.17	56.83****	42.12**	52.70***	7.89	23.02*	21.13*
King's College London	-1.94	-9.51	-0.39	-29.43*	-11.78	-0.02	1.10	2.28	-7.42	-6.21
LSE	3.14	-10.30	4.36	28.89	-15.41	60.77**	37.34*	49.50***	15.54	19.78
Newcastle University	-5.52	1.38	-7.28	-23.28	8.00	19.38	19.21	57.66****	1.69	8.69
Queen Mary University of London	10.34	5.02	-1.52	6.04	5.42	28.45	17.21	24.28	10.13	11.90
Queen's University Belfast	8.30	7.06	3.54	10.21	22.54	13.24	18.27	13.19	11.87	12.04
University College London	-1.12	-23.78	-27.99*	-5.98	-12.01	2.00	36.53*	15.10	-4.62	-2.15
University of Birmingham	20.65*	-4.96	10.88	22.51	13.74	21.24	22.06	28.05	15.16	16.77
University of Bristol	-1.77	6.52	7.57	1.20	-16.05	20.46	36.60**	14.72	7.78	8.65
University of Cambridge	8.66	36.47**	-2.85	1.38	46.65****	37.05	31.02*	26.43	22.62*	23.10*
University of Durham	-8.11	0.61	-29.57*	-2.26	8.61	17.41	15.81	34.96	0.35	4.68
University of Edinburgh	-14.76	-12.42	-20.09	-29.04*	2.98	-12.24	19.12	20.63	-9.49	-5.73
University of Exeter	-31.31***	-8.95	-11.28	-3.25	20.96**	19.69	13.10	7.77	-0.14	0.84
University of Glasgow	-17.33	0.14	-7.82	-1.35	21.55	38.54*	26.36	34.62*	8.58	11.83
University of Leeds	-9.15	-35.14*	-24.77*	-14.86	-3.30	15.18	-1.55	39.49**	-10.51	-4.26
University of Liverpool	-0.98	-8.50	-2.24	-9.73	-11.60	17.98	4.56	6.36	-1.50	-0.52
University of Manchester	-3.77	0.83	-31.28*	-40.74**	31.97**	-3.14	13.20	1.65	-4.70	-3.90
University of Nottingham	-4.30	70.96****	41.82**	-36.93*	1.23	6.62	-14.54	-25.44	9.26	4.92
University of Oxford	30.09***	42.86**	32.77***	6.69	71.87****	86.47****	68.40****	75.80****	48.44****	51.86****
University of Sheffield	-6.72	-0.98	1.14	-12.37	4.35	-14.88	-13.14	13.80	-6.08	-3.60
University of Southampton	-20.50*	-4.12	1.96	4.46	11.43	23.15	42.19***	30.48*	8.36	11.13
University of Warwick	25.52*	8.41	8.78	1.01	19.86	36.52	25.63	64.22****	17.96	23.74*
University of York	-14.46*	-5.43	-4.23	-17.59	-16.17	-21.82	1.35	6.99	-11.19	-8.91
Total Russell group	-1.06	2.37	-2.35	-6.22	12.06*	19.62*	20.08*	24.43**	6.35	8.61*
Non-Russell group	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**ATT$_{-2015}$**	**ATT**
Aberystwyth University	-13.06	-3.42	-7.17	-12.04	0.08	0.94	-3.93	-2.69	-5.51	-5.16
Aston University	-3.63	8.48	4.72	13.22	20.05	41.94*	23.68	36.68**	15.49	18.14
Bangor University	-4.66	2.40	4.82	9.49	18.74	34.88*	32.74*	21.74	14.05	15.02
Birkbeck College	1.15	1.48	-11.77	-7.42	-4.99	-2.60	-4.41	-8.69	-4.08	-4.65
Bournemouth University	-0.10	0.92	-3.13	16.87	17.78	23.77	29.70*	11.74	12.26	12.19
Brunel University London	16.12*	12.71	-7.36	-5.88	11.50	23.62*	-0.69	5.52	7.14	6.94
City University London	5.80	11.19	-13.76	-19.71	15.30	11.64	49.42***	70.70****	8.55	16.32
Coventry University	1.93	-6.92	-9.77	-10.75	-8.04	-4.95	-1.41	11.40	-5.70	-3.56
Cranfield University	-5.44	-20.14	-21.19	-14.58	-27.04	-23.89	-36.84*	-21.76	-21.30	-21.36
De Montfort University	-4.29	-6.93	2.34	-2.73	-8.61	-4.26	-10.84	-18.90	-5.04	-6.77
Edinburgh Napier University	-2.54	-6.46	-5.54	-12.00	-3.16	-0.62	1.65	-4.70	-4.09	-4.17
Glasgow Caledonian University	-4.03	-7.94	-8.22	-12.18	-17.06	-18.47	-22.23	-11.60	-12.87	-12.71
Heriot-Watt University	-7.38	-6.22	0.39	-0.93	3.40	2.46	22.29	7.17	2.00	2.64
Keele University	-11.43	-9.90	-15.77	-21.38	-17.51	-17.75	-23.32	-17.63	-16.72	-16.83
Kingston University	-7.85	-4.83	-3.82	-2.05	26.46*	11.38	3.23	19.48	3.21	5.25
Lancaster University	-5.71	9.32	-9.24	8.41	27.57*	35.71*	63.86****	73.99****	18.56*	25.48**
Leeds Beckett University	-7.14	-7.82	-7.87	-9.17	1.09	-3.17	1.88	0.72	-4.59	-3.93
London Business School	3.49	-15.71	-3.37	-14.35	-12.02	-19.95	-33.50*	-22.74	-13.63	-14.77
London Metropolitan University	-3.75	1.46	-6.31	-2.90	-2.36	-13.90	-6.53	-8.48	-4.89	-5.35
London South Bank University	-12.61	-8.36	-11.40	-14.36	-5.35	-10.81	-11.22	-8.36	-10.58	-10.31
Manchester Metropolitan University	8.91	8.05	-3.34	-4.54	-5.20	-0.08	-3.87	4.63	-0.01	0.57
Middlesex University	-2.29	1.35	-8.14	8.85	18.05	6.35	17.06	37.09**	5.89	9.79
Nottingham Trent University	-2.81	1.33	1.38	-5.98	5.73	3.70	2.00	10.39	0.76	1.96
Open University	20.02*	11.38	5.15	9.71	8.23	11.16	20.05	25.86*	12.24	13.94
Oxford Brookes University	-4.22	-5.27	-0.26	-1.09	11.02	13.50	10.17	4.02	3.40	3.48
Robert Gordon University	0.80	-1.38	-6.69	-8.47	-4.77	-5.23	2.53	-5.11	-3.31	-3.54
Royal Holloway, University of London	-5.42	13.72	0.72	-4.01	9.17	21.31	12.97	5.43	6.92	6.73
Sheffield Hallam University	-5.78	-11.04	-7.20	-7.00	-4.05	-3.24	-11.92	-1.46	-7.17	-6,46
Staffordshire University	-8.99	-8.95	-8.89	-13.93	-3.97	-2.97	-4.07	-5.00	-7.39	-7.09
Swansea University	-1.03	-5.69	-4.63	-1.50	-5.25	-14.44	-16.87	0.54	-7.05	-6.11
University of Aberdeen	-7.68	-11.93	-4.86	-14.83	-14.64	-11.02	-20.48	-9.72	-12.20	-11.89
University of Bath	-21.62*	-6.72	-15.15	-37.45*	-16.79	-10.48	18.71	16.58	-12.78	-9.11
University of Bedfordshire	-8.00	-7.00	-7.00	-9.00	1.00	6.00	1.00	7.00	-3.28	-2.00
University of Bradford	-3.15	-3.36	-3.14	-0.83	4.53	4.86	-0.78	-17.14	-0.26	-2.37
University of Brighton	-3.68	-6.69	-2.75	-7.82	8.27	3.00	2.50	5.83	-1.02	-0.16
University of Central Lancashire	-8.08	-7.78	-5.79	-6.25	4.87	1.49	4.71	5.65	-2.40	-1.39
University of Dundee	-9.29	-2.80	-6.61	-14.59	-12.73	-19.64	-7.74	-6.95	-10.48	-10.04
University of East Anglia	-8.92*	0.87	8.08*	10.27*	42.08****	19.83*	37.71****	32.74****	15.70	17.83
University of East London	-9.00	-9.00	-7.00	-11.00	2.00	2.00	-5.00	-1.00	-5.28	-4.75
University of Essex	-13.88	7.91	-14.13	-4.63	6.00	7.27	21.28	33.16**	1.40	5.37
University of Greenwich	-5.56	-2.04	-5.92	-10.19	5.62	3.54	3.83	3.10	-1.53	-0.95
University of Hertfordshire	-5.56	-3.70	0.07	-11.41	4.42	7.37	7.67	0.84	-0.16	-0.04
University of Hull	7.09	14.12	3.92	5.94	13.42	22.90	23.66*	28.46**	13.00	14.94
University of Kent	7.10	19.75	11.47	17.78	23.95*	40.74*	30.12*	64.77****	21.55*	26.96**
University of Leicester	-6.02*	0.57	2.66	-12.77*	4.58	3.59	11.85	10.73*	0.63	1.89
University of Northumbria at Newcastle	-10.03	-6.55	-7.62	-12.22	-0.40	-6.72	2.28	6.45	-5.89	-4.35
University of Plymouth	-5.66	-5.02	-12.62	-9.98	13.12	3.04	8.78	3.76	-1.18	-0.57
University of Portsmouth	-15.93	-14.82	-5.01	-6.66	-3.56	0.41	7.23	10.98	-5.47	-3.42
University of Reading	-11.97*	-18.89**	-22.74**	-26.52**	3.76	-2.44	-7.85*	25.17**	-12.37	-7.68
University of Salford	16.41*	2.49	7.11	5.45	3.35	11.82	-5.35	1.16	5.89	5.30
University of South Wales	-10.39	-7.77	-13.19	-19.20	-13.90	-14.88	-11.56	-10.67	-12.98	-12.69
University of St Andrews	2.38	23.90	19.37*	5.34	1.18	18.26	10.93	17.33	11.62	12.34
University of Stirling	-4.54**	1.41**	-5.92**	7.45***	-4.42**	10.15**	24.29****	25.11****	4.06	6.69
University of Strathclyde	4.74	-8.29	-11.33	27.14*	8.17	4.90	33.76*	-1.41	8.44	7.21
University of Sunderland	-8.00	-9.00	-11.00	-14.00	-4.00	-8.00	-4.00	-4.00	-8.28	-7.75
University of Surrey	16.13	-4.70	6.55	-13.43	41.50***	7.48	21.28	39.73**	10.68	14.31
University of Sussex	1.24	-10.99	-6.68*	-1.01	21.45**	23.55*	65.99****	35.49****	13.36	16.13
University of the West of England, Bristol	-1.38	-2.54	7.25	8.52	6.24	-5.46	8.71	12.45	3.04	4.22
University of Ulster	-2.98	-11.97	-5.36	-15.24	-17.88*	-15.31	-13.51	-15.43	-11.74	-12.21
University of Westminster	-6.07	-1.70	-0.42	-6.62	9.60	0.08	13.60	6.43	1.20	1.86
University of Wolverhampton	-6.00	-4.00	-11.00	-12.00	-5.00	-6.00	0.00	-3.00	-6.28	-5.87
Total Non-Russell group	-3.34	-2.44	-4.65	-5.44	3.28	3.25	6.31	8.74*	-0.43	0.71
Russell group - Non-Russell group	2.27	4.83*	2.30	-0.77	8.78***	16.37****	13.76***	15.68***	6.79***	7.90***
Remainers - Leavers	3.52	10.9***	5.5**	1.7	10***	14.4****	16.6****	16.8****	8.9***	8.8****
yearly ATT	-4.80**	0.67	-6.53	6.22***	-4.12	9.69**	23.13****	25.11****	24.26**	49.38***

Notes The last two columns contain each university ATT overall the post-treatment years (2008-2014) and adding 2015 (2008-2015), respectively. The last row of the table contains the overall yearly ATT for each year, and note that there are two panels, the top displays the results and subtotal for the Russell Group universities and the panel below for the non-Russell group ones. The last value at the bottom-right corner is the overall ATT for all universities included. The third and second-last rows contain the differences of means of ATTs respectively between the Russell and non-Russell groups and between Remainers and Leavers. Values are marked by *, **, ***, **** if they are significant at a level of, 0.10, 0.05, 0.01 or 0.001, respectively.

Figure 4.2: Distribution of pre-treatment RMSPE for placebos (US) and UK universities for the assessment of the quality of SCM matches for the total number of publications. Red and green lines refer to US and UK, respectively.

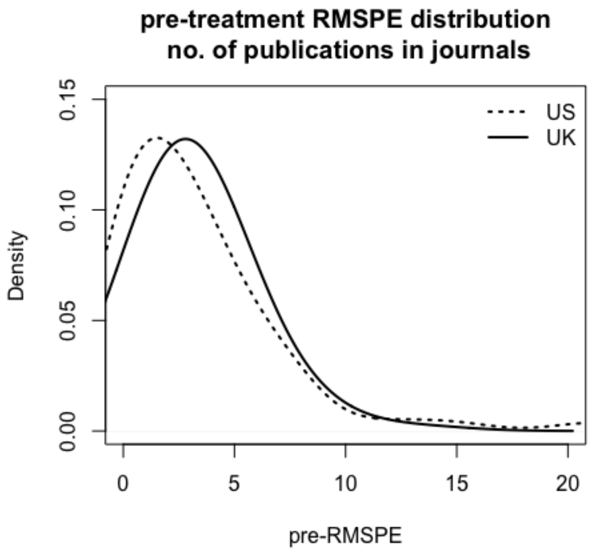
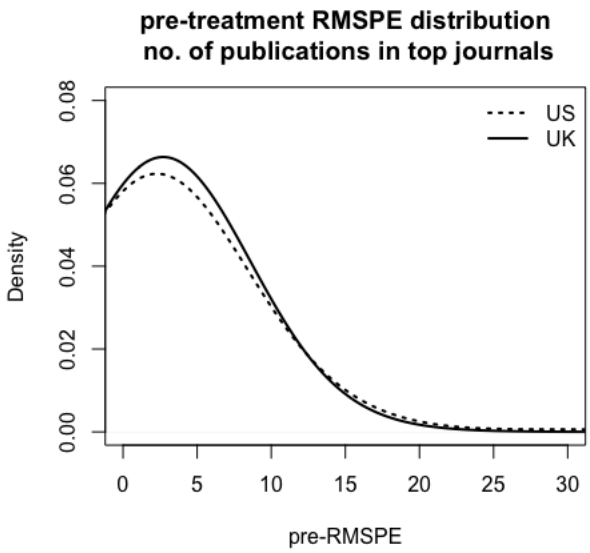


Figure 4.3: Distribution of pre-treatment RMSPE for placebos (US) and UK universities for the assessment of the quality of Synthetic Control Method (SCM) matches for the number of publications in a 3*, 4*, 4** journal. Red and green lines refer to US and UK, respectively.



of publications in Finance/Management in journals graded $3^*, 4^*$, or 4^{**}; the number of publications per author; the number of publications per author in journals graded $3^*, 4^*$, or 4^{**}; the number of publications in Economics and Econometrics journals graded as $3^*, 4^*$, or 4^{**} per author; and, the number of publications in Finance/Management in journals graded as $3^*, 4^*$, or 4^{**} per author. We also present these same measures in proportions rather than numbers, i.e. the proportion of publications in Economics and Econometrics journals; the proportion of publications in Finance/Management journals; the proportion of publications in journals graded as $3^*, 4^*$, or 4^{**}; the proportion of publications in Economics and Econometrics journals graded as $3^*, 4^*$, or 4^{**}; and, finally, the proportion of publications in Finance/Management journals graded as $3^*, 4^*$, or 4^{**}.

Table A7 presents results for all universities together, Table A8 the results for universities in the Russell Group and Table A9 for those that are not. The two top rows in these tables repeat information from Tables 4.2 and 4.3 to ease comparisons. With these extensions we want to explore whether or not the REF2014 affected the relative weight of these outcomes by sub-field and/or type of university.

As can be seen in Table A7, although the number of publications (ATT of 106.38**** and 150.74**** up to 2014 and to 2015, respectively) and the number of publications in $3^*, 4^*$, or 4^{**} journals (ATT of 24.26** and of 49.38*** for up to 2014 and to 2015, respectively) increased significantly since 2008, the number of publications per author decreased very significantly (ATT of -.53**** and -.60**** up to both years) and had done so on each individual year since 2009. At the same time, the number of publications in $3^*, 4^*$, or 4^{**} journals in Finance/Management per author increased slightly (insignificant ATT of 0.008 up to 2014 but significant ATT of 0.164*** up to 2015) while the proportion of publications in Economics and Econometrics went down (-0.126* and -0.150** respectively).

All other outcomes did not significantly change due to the REF2014. Thus, one hypothesis is that, while the total number of publications and those in $3^*, 4^*$, or 4^{**} journals went up overall, it was due to the increase in the number of publications per capita in $3^*, 4^*$, or 4^{**} journals in Finance/Management but not in Economics and Econometrics.

It is somehow surprising that the *number* of publications in $3^*, 4^*$, or 4^{**} journals for both Economics and Econometrics and Finance/Management did not change significantly (although the latter estimate is twice that of the former) and that the aggregate number of publications in $3^*, 4^*$, or 4^{**} journals per author did not change for Economics and Econometrics but increased significantly for Finance/Management.

We also observe that although the *proportion* of publications in Economics and Econometrics decreased, the *proportion* of publications in $3^*, 4^*$, or 4^{**} journals did not change significantly, neither aggregately nor by subfield.

To understand from where the above results stem from, we analyse the outcomes by separately for universities in the Russell Group and universities that are not.

What stands out in Table A8 below for the Russell Group is the fact that, although the number of publications (significant ATT of 11.42* up to 2014 and of 15.16** up to 2015) and the number of publications in top journals increased significantly (insignificant ATT of 6.35 up to 2014 but significant and of 8.61* up to 2015), nor the relative number of publications in 3^*, 4^*, or 4^{**} journals by subfield, nor per author changed significantly. Similarly, the *proportion* of publications by sub-field, overall and per author, did not change significantly. The number in Finance and Management in top journals per author has some significant changes (-0.039** in 2008, 0.038* in 2014 and 0.085*** in 2015) but the overall ATTs are insignificant.

In contrast, Table A9 provides a very different picture. For universities that did not belong to the Russell Group, the number of publications and those in $3^*, 4^*$, or 4^{**} journals do not change significantly (as reported in Tables 4.2 and 4.3) but, looking in more detail, we see that number of publications per author (-0.07* and -0.06*, respectively) and per author in $3^*, 4^*$, or 4^{**} journals declined (-.08* and -0.07*, respectively). At the same time, the *proportion* of publications in Economics and Econometrics journals declined (-.070* and -0.073*, respectively) while the *proportion* of those in Finance and Management increased (0.069* and 0.072*, respectively).

4.4.5 Extension: *Remainers* versus *Leavers*. The survival of the fittest or the sinking of the weakest?

Because since the beginning of the introduction of the RAE/REFs in the UK the absolute number of Economics departments submitting to the Economics/Econometrics Panel has decreased (from 41 in 2001 to 28 in 2014), we examine the impact of the REF 2014 on the research productivity separately for universities that submitted to this panel both in 2008 and 2014 (*remainers*) from those that submitted in 2008 to the Economics/Econometrics panel but switched to other panels in 2014 (*leavers*).

From Tables A10 and A11 below, we observe that the universities that remain in the Economics/Econometrics panel in 2014 increase significantly their number of publications (ATT of 11.48** up to 2014 and of 14.69*** up to 2015), their publications in good journals (7.46*** and 9.59***, respectively), as well as publications in top journals in Finance/Management (4.07** and 4.08**, respectively). On some of the years they experience significant increases or decreases in other outcomes intermittently but the overall ATTs are not significant.

The *leavers* experience a very different fate: from 2008 to 2014, the total number of publications and the publications in top journals do not change; but all other measures decrease significantly, most importantly the number of publications in top journals in Economics/Econometrics (-2.71** and -2.99**, respectively) while those in Finance/Management top journals do not change significantly (-0.21).

Every other indicator is significantly negative, the number of publications per author and in $3^*, 4^*$, or 4^{**} journals per author; per author and for each subfield, total and in top journals; as well as for all the proportions in each subfield, total, in top journals, total and per author (which means that the proportion of unclassified must have increased).

The difference in the fate of these two groups (Tables 4.2 and 4.3) is even more striking when we calculate the difference in the average effects in number of publications: 11.9 (p-value=0.004). Moreover, this difference is 10.44***, 10.77***, -0.51, 7.03*, 12.42***, 18.16***, 19.32***, 23.34***, respectively for the years 2008 through 2015.

Consistently, the difference in the estimated effects between the two groups of universities for the total number of publications in top journals is again positive and significant for the overall period 8.87 (p-value=0.0003), being 3.52, 10.9***, 5.5**, 1.7, 10***, 14.4****, 16.6****, and 16.8**** the differences in the average effects, for the years 2008 through 2015.

4.5 Conclusion and future research

A plausible interpretation of our results is that the overall increase in the number of publications and the number of publications in top journals due to the REF2014 stems from an increase in the number of publications in Finance/Management and a decrease in the proportion of Economics and Econometrics publications steered mainly by universities in the Russell Group that remained in the Economics and Econometrics panel.

The REF2014 did increase the total number of publications and those in top journals at the expense of the number of publications in top journals in Economics/Econometrics, the proportion of publications in Economics and Econometrics and the decrease in overall productivity of the Non-Russell group and the decay in the results of universities that left the Economics/Econometrics panel. This is counter-balanced by an increase in the proportion of publications in Finance/Management, in absolute and relative numbers. The number of publications per author did not increase. Therefore, the RAE/REFs have reinforced the strong position of the already strong departments in economics and *depressed* the weaker ones.

In fact, the REF may have introduced changes in the way academics work, incentivizing collaborative research and/or created distortions in the way universities recruit academics. Scientific collaboration is widely assumed to enhance the quality and impact of scientific research. Individuals with many links to others may have access to a larger pool of available ideas, methods, and resources, which allows cost sharing and time saving as a result of division of labour.

A potential future research would be to explore whether academics are becoming more connected to others similar to them, creating links within clusters, thus working more efficiently but not necessarily doing better research, or if they are bridging communities, achieving competitive advantage from inter-cluster weak ties, thus becoming empowered to tackle more important and difficult, possibly interdisciplinary, problems. The aim would be then to explore if the policy response mechanisms are sustainable or if they may induce negative feedback on research productivity in the longer term (such as if research excellence becomes more and more concentrated within few institutions); if the mechanisms at play are different for different disciplines or for different types of academics and if these mechanisms are gender (or otherwise) biased.

In particular, my idea is to focus on the development of a new interdisciplinary approach to evaluate the impact of policy interventions on agents that belong to connected communities. The new approach would challenge standard academic thinking in the way policies are assessed, by considering both direct and indirect effects stemming from spillovers that the policy may have on the behaviour of the community of interest, and its feedback on the variable directly targeted by the policy. In particular, the new approach would integrate state-of-the-art concepts and methodologies from two distinct fields of knowledge, Economics and Network Sciences (a field which draws theory and methods from computer science, physics and statistics), creating a new interdisciplinary methodological space.

Therefore, it's my intention to apply the proposed methods for the study of direct and indirect (unintended) effects, if any, due to REF on the dynamics of mobility (universities' hiring decisions) and collaboration (preferential attachment of authors for joint research) networks of researchers in the UK and the impact of these changes on research productivity.

Chapter 5

The doom-loop: financial correlation networks based on Credit Default Swaps

Abstract

We analyse the interdependence between sovereigns and financial institutions in terms of risk transmission. In particular, we analyse CDS data issued by sovereigns and financial institutions between 2009 and 2016 to infer spillover effects in the global financial system.

We introduce a SVN approach, which is novel in this context, and show that traditional approaches to compute spillover effects can benefit when used in companion with SVNs.

Specifically, we bring forth two benefits: 1) overcome the problem observed in the orthogonalized FEVD related to the dependence of the results on the order of the variables in the VAR model, and 2) prove both formally and empirically that the generalized FEVD is not suitable for the description of pure spillover effects, since its coefficients reflect both a synchronous part—due to the co-movement of variables in the system (R-squared)—and an asynchronous component that represents the pure spillover effect.

We derive pure spillover effects from the generalized FEVD to then construct SVNs, which provide insights on which preferential patterns risk transmits across the agents of the financial system.

5.1 Introduction

Sovereigns are exposed to bank risk and, at the same time, banks are exposed to sovereign risk. During the euro-area sovereign debt crisis started in

2010, this two-way risk exposure generated a "vicious circle", also known as the "doom loop" [66]. At a point when government bonds were considered risky assets, euro-area banks faced with both balance sheet and reputational risks, making it hard to compete with their non-euro area counterparts, forcing to tight their exposure to sovereign credit risk, thus igniting the most disruptive financial crisis has ever jeopardized the Euro currency system. Understanding the relationship between sovereign and banking risk is therefore fundamental to deploy policies and regulatory measures aimed at reducing the probability and impact of financial crises.

We focus our analysis on the interdependence between sovereigns and financial institutions in terms of their risk transmission. We termed with "interdependence" the bidirectional relationship between the risk profile of a government and of owned financial groups over time. Notice that a "feedback loop" is a special case of such interdependence, when risk factors for either banks or sovereigns lead to a self-reinforcing deterioration of credit risk.

There is a growing body of theoretical studies that illustrate how increasing interconnectedness can pose a serious threat to the stability of a financial system due to contagion and amplification effects ([6, 62], [81, 80]). For example, Acemoglu et al. (2012) [5] show that intersectorial input-output linkages between firms can give rise to aggregate (or economy-wide) fluctuations when idiosyncratic or sector shocks propagate, thus leading to network effects that impact the aggregate economy. Covi and Eydam (2017) [47] analysed a panel data on European banks and sovereigns ranging from 2012 and 2016 in order to test the effects of the Bank Recovery and Resolution Directory (BRRD) on the two-way feedback process, finding that there was a pronounced feedback loop between banks and sovereigns from 2012 to 2014, which disappeared after the implementation of the new regulatory framework. Acharya et al. 2013 [8] analyse CDS rates on European sovereigns and banks for 2007-11, showing that bailouts triggered the rise of sovereign credit risk, highlighting how post-bailout changes in sovereign CDS explain changes in bank CDS even after controlling for aggregate and bank-level determinants of credit spreads, confirming the sovereign-bank loop. Diebold and Yielmaz (2012) [56] use a generalized VAR framework and FEVD coefficients that are invariant to the variables ordering, proposing some measures of volatility spillovers to characterize daily volatility spillovers across US stocks, bonds, foreign exchange and commodities markets from 1999 to 2010, showing that cross-market volatility spillovers were quite limited until the global financial crisis began in 2007, and as the crisis intensified, the volatility spillovers did too. Acemoglu et al. 2015 [6] highlight that dense networks facilitate propagation of shocks, leading to a more fragile financial system, and that the same factors that contribute to the resilience under certain conditions may function as significant sources of systemic risk under others.

We introduce a validated network approach, which is novel in this context. In particular, SVNs [183] allow one to assess the "excess" of risk transmission among the agents of the financial system, therefore going beyond the observed interconnections due to the "physiological" heterogeneity that characterizes the system. The new approach allows one to better highlight the nodes and patterns in the network that are less resilient when risk propagates in the system. We study spillover effects among sovereigns and financial institutions in the global financial market. Specifically, we analyse CDS data issued by sovereigns and financial institutions between April 2009 and July 2016 to infer their risk transmission. Also, we deal with the estimation of high-dimensional regularized VAR models by using the Least Absolute Shrinkage and Selection Operator (LASSO) and post-LASSO method, which leads to regularized networks. We then resort to the FEVD [129, 157] to compute the spillover effects. We show that traditional approaches to compute spillover effects can benefit when used in companion with SVNs. Specifically, we bring forth two benefits: 1) overcome the problem observed in the orthogonalized FEVD related to the dependence of the results on the order of the variables in the VAR model, and 2) prove both formally and empirically that the generalized FEVD is not suitable for the description of pure spillover effects, since its coefficients reflect both a synchronous part—due to the co-movement of variables in the system (correlations or R-squared)—and an asynchronous component that represents the pure spillover effects. We derive pure spillover effects from the generalized FEVD to then construct SVNs, which eventually, show an overlap with the SVNs constructed using the orthogonalized FEVD.

5.2 Data and Methods

5.2.1 Data

We downloaded the data from the Thomson Reuters Eikon Database. The data refer to 147 daily CDS spread with a maturity of 5-years and issued by sovereigns and financial companies all across the globe (see Tables B1 and B2 of Appendix B for the list of financials and sovereigns considered in the study).

5.2.2 Variance Decomposition for high-dimensional problems

We study financial networks in which we have n nodes. A subset of these nodes includes sovereigns, while the remaining nodes are financial companies. We study the links among these sovereigns and financial companies focusing on their CDS. Therefore, we compute the CDS returns of the n nodes at time t, which we include in a $n \times 1$ vector $\mathbf{R}_t = [R_{1,t} \ \cdots \ R_{n,t}]'$, for $t = 1, \cdots, T$. Following [89], we use the Generalized Dynamic Factor (GDF) model (see [73, 71, 70] and [72]) to separate common shocks from idiosyncratic shocks. We then compute the following decomposition:

$$R_{j,t} = C_{j,t} + X_{j,t} = b_{j,1}(L)u_{1,t} + \cdots b_{j,q}(L)u_{q,t} + X_{j,t}, \tag{5.1}$$

where $C_{j,t}$ and $X_{j,t}$ denote, respectively, the common and the idiosyncratic components of $R_{j,t}$, for $j = 1, \cdots, n$, $\mathbf{u}_t = [u_{1,t} \cdots u_{q,t}]$ is an unobservable q-dimensional orthonormal white noise with square-summable filters $b_{j,1}(L), \cdots,$ $b_{j,q}(L)$, whereas L is the lag operator.[1]

We adopt the decomposition in Eq. (5.1) because we focus on the so-called 'pure' contagion risk component of systemic risk; that is, we filter the shock arising from a given node which subsequently propagates towards other nodes within the network [89]. Following [55], [53] and [89], we use the FEVD method to measure the spillover effects among the n nodes. The FEVD, in turn, builds on the estimation of the following covariance stationary VAR model:

$$\mathbf{X}_t = \boldsymbol{\nu} + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \mathbf{X}_{t-i} + \boldsymbol{\epsilon}_t, \tag{5.2}$$

where $\mathbf{X}_t = [X_{1,t} \ \cdots \ X_{n,t}]'$, $\boldsymbol{\Phi}_i$ is an $n \times n$ parameter matrix, $\boldsymbol{\nu}$ is a $n \times 1$ vector of intercept terms and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, with $\mathbb{E}(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_s') = 0$, for $s \neq t$.[2]

Under the stability assumption, the model in Eq. (5.2) admits the following infinite Moving Average (MA) representation [129]:

$$\mathbf{X}_t = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \mathbf{A}_i \boldsymbol{\epsilon}_{t-i}, \tag{5.3}$$

where the coefficient matrix $\mathbf{A}_i$ can be iteratively computed as $\mathbf{A}_i = \boldsymbol{\Phi}_1 \mathbf{A}_{i-1} + \boldsymbol{\Phi}_2 \mathbf{A}_{i-2} + \cdots + \boldsymbol{\Phi}_p \mathbf{A}_{i-p}$, for $i = 1, 2, \cdots$, whereas $\mathbf{A}_0 = \mathbf{I}_N$ and $\mathbf{A}_i = 0$ for $i < 0$.

An alternative way to compute $\mathbf{A}_i$ in Eq. (5.3) takes the following form

[1] Following [70] and [89], we employ the method of [91] to estimate the optimal value of q. This rule suggests $q = 1$, which is consistent with the findings of [89].

[2] Following [89], we set $p = 2$ in our empirical analysis.

[129]:

$$\mathbf{A}_i = \mathbf{J}\mathbf{\Phi}^i\mathbf{J}', \tag{5.4}$$

where

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi}_1 & \mathbf{\Phi}_2 & \cdots & \mathbf{\Phi}_{p-1} & \mathbf{\Phi}_p \\ \mathbf{I}_n & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_n & \mathbf{0} \end{bmatrix} \tag{5.5}$$

is an $np \times np$ matrix, $\mathbf{J} = [\mathbf{I}_n\ \mathbf{0}\ \cdots\ \mathbf{0}]$ and $\mathbf{I}_n$ is an $n \times n$ identity matrix.

Our method should be flexible in dealing with large values of n. However, the coefficients derived from the standard VAR model in Eq. (5.2) are affected by serious issues related to the accumulation of estimation errors when n takes large values. Furthermore, we do not know a priori which of the variables in Eq. (5.2) have a significant impact on $\mathbf{X}_t$. Our method would suffer from overfitting problems when using too many covariates. On the other hand, we run the risk of an omitted variable bias when shrinking the set of such regressors. We deal with the curse of dimensionality using a well–known variable selection and regularization method; that is, the LASSO introduced by [181]. This method consists of adding an ℓ_1-norm penalty to the Ordinary Least Squares (OLS) loss function. As a result, we estimate the parameters from the following optimization problem:

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}_j}{\operatorname{argmin}} \left(\sum_{t=p+1}^{T} (X_{j,t} - \nu_j - \sum_{i=1}^{p} \boldsymbol{\phi}_{i,j}\mathbf{X}_{t-i})^2 + \lambda_j \sum_{i=1}^{p} |\boldsymbol{\phi}_{i,j}| \right), \tag{5.6}$$

for $j = 1, \cdots, n$, where $\boldsymbol{\beta}_j = [\nu_j\ \boldsymbol{\phi}_{1,j} \cdots \boldsymbol{\phi}_{p,j}]$, $\boldsymbol{\phi}_{i,j}$ is the j-th row of $\mathbf{\Phi}_i$, ν_j is the j-th element in $\boldsymbol{\nu}$ and $\lambda_j > 0$ is a tuning parameter.

λ_j determines the intensity of the penalization in (5.6). The larger λ_j is, the larger the number of coefficients that approach zero, providing a sparser solution. We select the optimal value of λ_j by employing the 10-fold cross-validation method, which is widely used in the statistical and econometric literature (see, e.g., [96]).

We differ from [89] who, instead, used the elastic net shrinkage method in place of the LASSO. Indeed, according to [89], the elastic net penalty has the advantage of being relatively less aggressive in reducing the number of selected variables. Nevertheless, on the other hand, this method leads to denser networks, in which it could be difficult to identify the relevant transmission channels among the n nodes. In contrast, we prefer the LASSO because it leads to sparse solutions, selecting the nodes that have a stronger impact on the entire network. Moreover, we also differ from [89] because we do not directly use the coefficients computed from the penalized problem in (5.6) to build our network, but we improve the accuracy of the estimates by implementing a further exercise. Indeed, penalized regression models suffer from some limitations. For instance, they typically provide biased estimates, overshrinking the values of the selected variables. In this study, we address this issue by using the post-LASSO method, which is described as follows. We solve in a first step the problem in (5.6) and select the regressors whose coefficients are, in absolute value, greater that a given threshold η.[3] We include the selected regressors in $\mathbf{X}_{t-i}^{(s)}$ and solve, in a second step, the following problem, which does not include any penalty function:

$$\widehat{\boldsymbol{\beta}}_j^{(s)} = \underset{\boldsymbol{\beta}_j^{(s)}}{\operatorname{argmin}} \sum_{t=p+1}^{T} \left(X_{j,t} - \nu_j - \sum_{i=1}^{p} \boldsymbol{\phi}_{i,j}^{(s)} \mathbf{X}_{t-i}^{(s)} \right)^2 . \tag{5.7}$$

We finally obtain the estimate of $\mathbf{\Phi}_i$, denoted as $\widehat{\mathbf{\Phi}}_i$, for $i = 1, \cdots, p$. Note that the coefficients in $\widehat{\mathbf{\Phi}}_i$ are classified in two groups: i) the coefficients of the covariates that are LASSO-selected in the first step, which are computed from (5.7); and ii) the coefficients of the covariates that are not LASSO-selected in the first step (i.e., the ones whose absolute value is lower than or equal to η), which we set equal to zero. Notably, the post-LASSO method provides superior estimates (see, e.g.,[65], [23] and [98])[4].

After estimating the coefficients of the penalized VAR model, we compute the FEVD; that is, the proportion of the h-step ahead forecast error variance of variable i that is accounted for by the innovations in variable j [129]. We first define the orthogonalized FEVD [129], which takes the following form:

$$\theta_{i,j}^{o}(H) = \frac{\sum_{h=0}^{H-1} \left(\mathbf{e}_i' \mathbf{A}_h \mathbf{P} \mathbf{e}_j\right)^2}{\sum_{h=0}^{H-1} \left(\mathbf{e}_i' \mathbf{A}_h \mathbf{\Sigma} \mathbf{A}_h' \mathbf{e}_i\right)}, \tag{5.8}$$

where $\mathbf{e}_j$ is an $n \times 1$ selection vector with unity as its j-th element and zeros elsewhere, whereas $\mathbf{P}$ is computed from the Cholesky decomposition of $\mathbf{\Sigma}$: $\mathbf{P}\mathbf{P}' = \mathbf{\Sigma}$.

Despite being widely used in many statistical applications, the orthogonalized FEVD suffers from an important limitation. Indeed, the results depend

[3] We set $\eta = 0.000001$ in our empirical analysis.

[4] We also evaluate the statistical significance of the coefficients resulting from (5.7), comparing the results with the ones obtained with the elastic net. We checked that the LASSO provides a relevant percentage of selected variables that are also statistically significant at the 1% level. In contrast, a relevant percentage of variables that are selected by the elastic net are not statistically significant at the 5% level. This evidence further supports our choice of using the LASSO.

on the ordering of the variables in the VAR model. We attempt to overcome this gap by randomly permuting the positions of the variables W times and, for each permutation, we estimate the penalized VAR along with the corresponding FEVD. We then select the vertices that are stable with a given frequency.

We compare the orthogonalized FEVD with the generalized FEVD introduced by [157], which is defined as follows:

$$\theta^g_{i,j}(H) = \frac{\sigma_{jj}^{-2} \sum_{h=0}^{H-1} \left(\mathbf{e}_i' \mathbf{A}_h \boldsymbol{\Sigma} \mathbf{e}_j\right)^2}{\sum_{h=0}^{H-1} \left(\mathbf{e}_i' \mathbf{A}_h \boldsymbol{\Sigma} \mathbf{A}_h' \mathbf{e}_i\right)}, \tag{5.9}$$

where σ_{ij} is the element placed in the i-th row and in the j-th column of $\boldsymbol{\Sigma}$, for $i, j = 1, \cdots, n$.

Note that $\sum_{j=1}^{n} \theta^o_{i,j}(H) = 1$, whereas $\sum_{j=1}^{n} \theta^g_{i,\,j}(H) \neq 1$ in general.
As in [89], we then normalize $\theta^g_{i,j}(H)$ by computing the following quantity:

$$\gamma_{i,j} = \frac{\theta^g_{i,j}(H)}{\sum_{j=1}^{n} \theta^g_{i,j}(H)} \times 100. \tag{5.10}$$

The main advantage of the generalized FEVD is that it provides results that are invariant to the ordering of the variables in the VAR model. However, we check that this method often produces values of $\gamma_{i,j}$ which are clearly distant from zero even if the variable j is not LASSO-selected as a relevant regressor to explain variable i. In contrast, the values of $\gamma_{i,j}$ are mainly driven by synchronous correlations. In this case, we could simply compute a correlation matrix in place of the generalized FEVD to obtain similar information. We formally show this evidence in Section 5.2.4.

5.2.3 SVN of spillover effects

To construct the SVN, the idea is to discretize the FEVD coefficients as follows. We first perform a bootstrap algorithm on the starting date to generate the sampling distribution of FEVD coefficients. By doing so, we are able to find a threshold based on the results deriving from the generated bootstrap samples. We define this threshold as the standard deviation of the sampling distribution: $\theta^*_{i,j} = 0.005$.

Therefore, we use this threshold to associate each $\theta_{i,j}$ with a positive integer k_{ij}, taken as the greatest integer less than or equal to the ratio between the observed coefficient $\theta_{i,j}$ and the threshold $\theta^*_{i,j}$:

$$\left\lfloor \frac{\theta_{i,j}}{\theta^*_{i,j}} \right\rfloor = k_{ij} \tag{5.11}$$

We then statistically validate link between i and j if:

$$\text{p-value}(k_{ij}) = 1 - \mathrm{P}_{hyper}\left(k_{ij}; \sum_j k_{ij}, \sum_i k_{ij}, \sum_i \sum_j k_{ij}\right) < \frac{0.01}{\#tests} \quad (5.12)$$

using a Bonferroni correction for multiple tests with $\#tests = n(n-1)$, that is, two tests per pair.

5.2.4 A simple model to describe how synchronous and lagged correlation among variables influence FEVD coefficients

In this subsection, we show that, although the generalized FEVD does not depend on the ordering of variables in the VAR model, it is biased when it comes to measure pure spillover effects. Indeed, we show that it involves two parts, a part due to synchronous correlations, and another part due to asynchronous correlations, which measure the pure spillover effects among the variables. Assuming the matrix of coefficients of lag 1 of a VAR process as being the result of the following model, which depends on parameters $(\lambda_1, \lambda_2) \in [0,1]^2 : \lambda_1 + \lambda_2 = 1$

$$\mathbf{\Phi}(\lambda_1, \lambda_2) = (\lambda_1 - \lambda_2)\mathbf{I} + \lambda_2 \mathbf{U} \quad (5.13)$$

where $\mathbf{I}$ is the identity matrix of dimension n, and $\mathbf{U}$ is the all-ones matrix of dimension n, with n the number of variables.
λ_1 is an intensity parameter of the auto-correlation process, while λ_2 the one of the process of lagged cross-correlations between variables.
For simplicity, we assume that any lag greater than one is not statistically significant: $\mathbf{\Phi}_i = \mathbf{0}_{nxn}, \quad \forall i > 1$.

We model the process as the sum of two effects: an effect due to lagged auto-correlations of features; and an effect due to lagged cross-correlations between features.
So, if only auto-correlations are present, we would have

$$\mathbf{\Phi}_1(\lambda_1, \lambda_2 = 0) = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_1 \end{bmatrix}$$

and with only lagged cross-correlations,

$$\mathbf{\Phi}_1(\lambda_1 = 0, \lambda_2) = \begin{bmatrix} 0 & \lambda_2 & \lambda_2 \\ \lambda_2 & 0 & \lambda_2 \\ \lambda_2 & \lambda_2 & 0 \end{bmatrix}$$

The coefficients of the MA representation of the generic VAR model can be written recursively as follows

$$\mathbf{A}_i = \sum_{j=1}^{i} \mathbf{\Phi}_j \mathbf{A}_{i-j}, \quad \forall i = 1, 2, \ldots$$

Proposition: Assuming the process being a VAR(1) with coefficients $\mathbf{\Phi}_i = \mathbf{0}_{nxn}, \quad \forall i > 1$; By construction, $\mathbf{A}_0 = \mathbf{I}$. Also, for lag=1: $\mathbf{A}_1 = \mathbf{\Phi}_1 A_0 = \mathbf{\Phi}_1 \mathbf{I} = \mathbf{\Phi}_1$; for lag=2: $\mathbf{A}_2 = \mathbf{\Phi}_1 \mathbf{A}_1 + \mathbf{\Phi}_2 \mathbf{A}_0 = \mathbf{\Phi}_1^2$
So, in general, the coefficients of the MA representation of VAR(m) are $\mathbf{A}_m = \mathbf{\Phi}_1^m$

Proof
By mathematical induction: assuming $\mathbf{A}_{m-1} = \mathbf{\Phi}_1^{m-1}$ as true.
For $m = 1$: $\mathbf{A}_1 = \mathbf{\Phi}_1 \mathbf{A}_0 = \mathbf{\Phi}_1$
$\mathbf{A}_m = \mathbf{\Phi}_1 \mathbf{A}_{m-1} + \mathbf{\Phi}_2 \mathbf{A}_{m-2} + \cdots + \mathbf{\Phi}_m \mathbf{A}_0 = \mathbf{\Phi}_1 \mathbf{A}_{m-1} = \mathbf{\Phi}_1^m, \quad$ since $\mathbf{\Phi}_i = \mathbf{0}_{nxn}, \quad \forall i > 1$

Therefore, to evaluate $\mathbf{A}_m = \mathbf{\Phi}_1^m$, it is necessary to evaluate the matrix

$$\mathbf{\Phi}_1^m = [(\lambda_1 - \lambda_2)\mathbf{I} + \lambda_2 \mathbf{U}]^m$$

By indicating $\lambda_1 - \lambda_2 = \Delta\lambda$ and using the Binomial theorem (which can be used since matrices $\mathbf{I}$ and $\mathbf{U}$ commute), we can derive the following equations[5]:

$$\begin{aligned} \mathbf{\Phi}_1^m &= (\Delta\lambda \mathbf{I} + \lambda_2 \mathbf{U})^m \\ &= \sum_{k=0}^{m} \binom{m}{k} \Delta\lambda^k \mathbf{I}^k \cdot \lambda_2^{m-k} \mathbf{U}^{m-k} = \frac{\mathbf{U}}{n}[(\Delta\lambda + n\lambda_2)^m - \Delta\lambda^m] + \Delta\lambda^m \mathbf{I} = \mathbf{A}_m \end{aligned} \tag{5.14}$$

If $\lambda_1 = \lambda_2 = \lambda \;\Rightarrow\; \Delta\lambda = 0$ and $\quad \mathbf{A}_m = \frac{\mathbf{U}}{n} n^m \lambda^m = n^{m-1}\lambda^m \mathbf{U}$

Let $\mathbf{\Sigma} = \{\sigma_{ij}\}_{i,j=1,2,\ldots,n}$ be the variance-covariance matrix of model residuals

[5] In Appendix B we report the detailed steps leading to the result

and $\mathbf{U\Sigma U} = S_T\mathbf{U}$, where $S_T = \mathbf{U\Sigma} = \sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}$

It's also useful to introduce the notation: $\bar{\sigma}_i = \frac{\sum_{j=1}^{n}\sigma_{ji}}{n}$.

Following the notation of Demirer et al. 2018 [53] (and correcting the typo $\sigma_{jj}^{-1} \to \sigma_{jj}^{-2}$), the H-step-ahead generalized forecast error variance $\theta_{ij}^{g}(H)$ is:

$$\theta_{ij}^{g}(H) = \frac{\sigma_{jj}^{-2}\sum_{h=0}^{H-1}(\mathbf{e}_i^T\mathbf{A}_h\mathbf{\Sigma}\mathbf{e}_j)^2}{\sum_{h=0}^{H-1}(\mathbf{e}_i^T\mathbf{A}_h\mathbf{\Sigma}\mathbf{A}_h^T\mathbf{e}_i)}, \quad H = 1, 2, \ldots \tag{5.15}$$

It can be shown by using some algebra (in Appendix B we report the detailed steps leading to the result) that:

- if H=1, then

$$\theta_{ij}^{g}(H=1) = \frac{\sigma_{ij}^2}{\sigma_{ii}^2\sigma_{jj}^2} = R_{ij}^2 \tag{5.16}$$

- if H=2, then

$$\theta_{ij}^{g}(H=2) = R_{ij}^2\left\{\frac{1 + (\frac{n\lambda_2\bar{\sigma}_j}{\sigma_{ij}} + \Delta\lambda)^2}{1 + \Delta\lambda^2 + \frac{S_T}{\sigma_{ii}^2}\lambda_2^2 + 2n\lambda_2\Delta\lambda\frac{\bar{\sigma}_i}{\sigma_{ii}^2}}\right\} \tag{5.17}$$

Therefore, with VAR(1) and H=1, FEVD coefficients just reflect the synchronous relationships among variables. On the contrary, Eq. 5.17 highlights that for H=2 the FEVD can be written as the combination of two effects: a synchronous effect, summarized by coefficient R_{ij}^2, and another one that accounts for pure spillover effects due to asynchronous relationships among variables (the outcome of VAR). In particular, spillover effects will depend on $\mathbf{\Sigma}$, the number of variables n, and parameters λ_1 and λ_2. It's worth to note that if $\lambda_2 = 0$ then, according to the VAR, the system does not reveal significant lagged cross-correlations, and, as a consequence, the FEVD leads again to the R^2.

5.3 Empirical results

In this section, we first show empirically that the spillover effects that derive from the generalized FEVD are biased, since they involve both an asynchronous and a synchronous component. Indeed, by randomly shuffling the original data only with respect to time dimension—in order to completely remove the asynchronous effects—results remain the same as in the case of original data.

Secondly, we show that the orthogonalized FEVD better describes spillover effects in the system, and we use SVNs to overcome the problem of dependence of the method on the ordering of the variables in the phase of estimation of the VAR model.

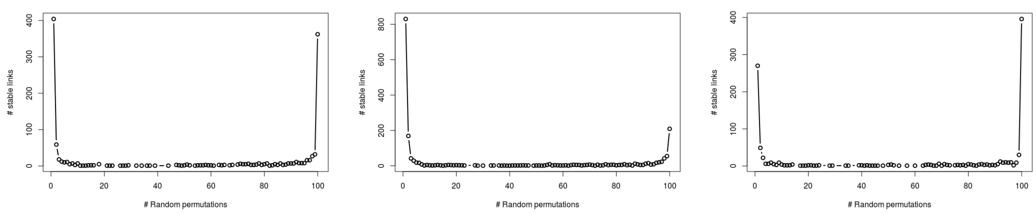
Thirdly, we construct the SVN using: (i) the biased generalized FEVD; (ii) the unbiased generalized FEVD; and (iii) the orthogonalized FEVD. We also compute Jaccard index values and overlap coefficients to compare the resulting SVNs in the different cases. In fact, we show that there is a good overlap between the SVNs using the unbiased version of the generalized FEVD and the ones using the orthogonalized FEVD.

5.3.1 Construction of the SVN using biased generalized FEVD coefficients

We construct the SVN according to the method introduced in subsection 5.2.3. In Figures 5.5, 5.6 and 5.7 we show the SVN respectively for the first, second, and third sub-period. In all the sub-periods, most of the relationships between nodes are bidirectional, meaning that almost only synchronous effects are highlighted by the method.

We then randomly shuffle the data with respect to time dimension and construct the SVNs. We repeat the procedure 100 times and find that, despite the shuffled time dimension, some of the links are still persistent in the networks. Fig. 5.1 describes a bimodal distribution of stable links for all sub-periods; it shows that the resulting SVNs have a peak of stable links even beyond 90 time permutations, which are clearly due to synchronous effects (R^2) among variables. Moreover, in Figure 5.4 we show the values of the Jaccard index to quantify the overlap between the networks resulting from the shuffling procedure. For all the sub-periods Jaccard values are quite high, meaning that, no matter the temporal ordering is, the networks keep showing the synchronous component contained in the data, which dominates the asynchronous one.

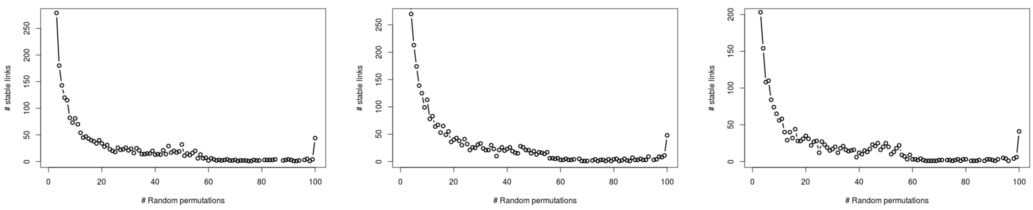
Figure 5.1: Link stability of SVNs constructed using biased generalized FEVD coefficients and *data with shuffled time dimension*: sub-periods 04/2009-08/2011 (left), 09/2011-01/2013 (middle), 02/2013-07/2016 (right)



5.3.2 Construction of the SVN using orthogonalized FEVD coefficients

Unlike the generalized FEVD, the orthogonalized FEVD treats shocks as orthogonal to each other and, therefore, it allows to write the variance of the total forecasting error as a sum of variances of the single shocks (as the covariance terms are zero following the orthogonality property of structural shocks) and most importantly, it is not affected by synchronous effects. Nevertheless, one crucial drawback of the orthogonalized FEVD is that it depends on the order of the variables defined in phase of estimation of the VAR model. We use SVNs to overcome the problem. Therefore, we run many permutations of variable orderings and build the SVNs using the respective estimated orthogonalized FEVD coefficients. For the final networks we consider links that are stable in more than 90 random permutations. Fig. 5.2 shows the peak of stable links in correspondence of 100 random permutations. Figures 5.8, 5.9, and 5.10 show the SVNs for the first, second, and third sub-period, respectively. Morover, the networks don't show bidirectional links, suggesting the absence of synchronous effects.

Figure 5.2: Link stability of SVNs constructed using orthogonalized FEVD coefficients and original data: sub-periods 04/2009-08/2011 (left), 09/2011-01/2013 (middle), 02/2013-07/2016 (right).



5.3.3 Construction of the SVN using unbiased generalized FEVD coefficients

We remove the first addend (referring to $h = 0$) in both the numerator and denominator of Formula 5.15 to "clean" the total effects from the synchronous component and obtain the pure spillover effects. We construct the SVNs using the resulting unbiased FEVD coefficients. This time, stability of links in the SVNs obtained by randomizing the data with respect to time converges towards 0 right after 5 permutations (see Fig. 5.3).

We show the networks for the three sub-periods in Figures 5.11, 5.12, and 5.13, respectively. The networks show a dense interconnectedness in all the sub-periods. They also show that the most interconnected nodes (and therefore

systemically important) are sovereigns and financial companies from Europe, and this “doom-loop” interplay is seen since the “explosion” of the European sovereign debt crisis, which started at the end of 2009.
These networks show a greater number of statistically significant spillover effects compared to the other types of networks, actually highlighting a global interdependence in the system. Nevertheless, it is worth to note how the SVN using unbiased generalized FEVD and SVN using orthogonalized FEVD share a common source of information about the system interconnectedness.

Indeed, we quantify the overlap between the SVNs obtained using the unbiased generalized FEVD and the ones obtained using the orthogonalized FEVD to see if they actually attempt to measure the same thing. Since the networks being compared have a different number of nodes, the overlap coefficient (also called Szymkiewicz–Simpson coefficient) is preferred to the Jaccard coefficient. The overlap coefficient amounts for 0.53, 0.56, 0.52 (statistically significant in all three cases through a hypergeometric test) for the first, second, and third sub-period, respectively, denoting that the two networks share the information on spillover effects among the agents of the system.

Figure 5.3: Link stability of SVNs constructed using unbiased generalized FEVD coefficients and *data with shuffled time dimension*: sub-periods 04/2009-08/2011 (left), 09/2011-01/2013 (middle), 02/2013-07/2016 (right).

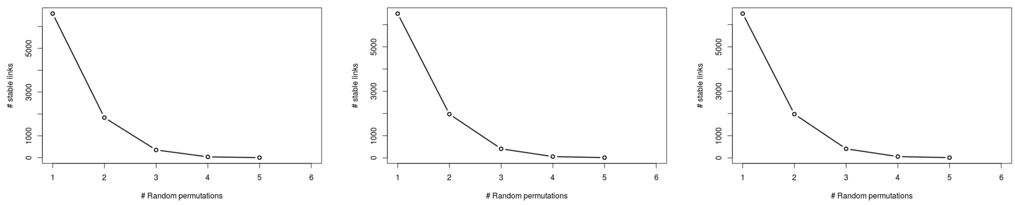
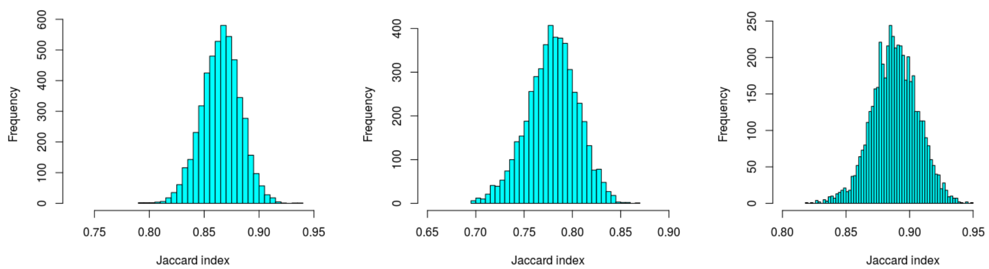


Figure 5.4: Histograms of Jaccard index values comparing the set of links of the SVNs constructed using biased generalized FEVD coefficients and randomly shuffling time dimension: sub-periods 04/2009-08/2011 (left), 09/2011-01/2013 (middle), 02/2013-07/2016 (right).



5.4 Conclusions

We have studied the interdependence between the agents of the financial system through CDS data over the period that goes from April 2009 to July 2016. We have introduced an approach based on SVNs to measure such interdependencies. Through a SVN we have assessed the "excess" of risk transmission among the agents of the system compared to the case of random connectedness while controlling for the heterogeneity in the system.

We have shown both formally and empirically that the generalized FEVD needs to be modified when used for measuring pure spillover effects. Therefore, we have untangled the asynchronous relationships from the synchronous relationships, and we have proved the validity of the approach through the application of SVNs on data randomly shuffled with respect to time dimension.

Also, we have overcome the problem of the dependence of spillover effects on the ordering of variables defined in the VAR model when the orthogonalized FEVD is used. In particular, we have constructed as many SVNs as the number of generated random permutations of variable ordering, and we have selected the links that persisted beyond a given threshold of the number of permutations.

Finally, we have found a statistically significant overlap between the SVNs constructed using unbiased generalized FEVD and the ones constructed using orthogonalized FEVD, meaning that there is consistency in what they aim to measure.

Therefore, the new approach properly highlights the patterns of the network that are less resilient when it comes to risk propagation.

Figure 5.5: SVNs using biased generalized FEVD coefficients and original data: sub-period 04/2009-08/2011.
Squared nodes are financial companies. Colours: Yellow=Asia; Blue=EU and UE (EU with own currency); Pink=Middle East; Light red= Other Asian; Green=Oceania; Light blue= European not in EU; Brown=Russia; Red=US.

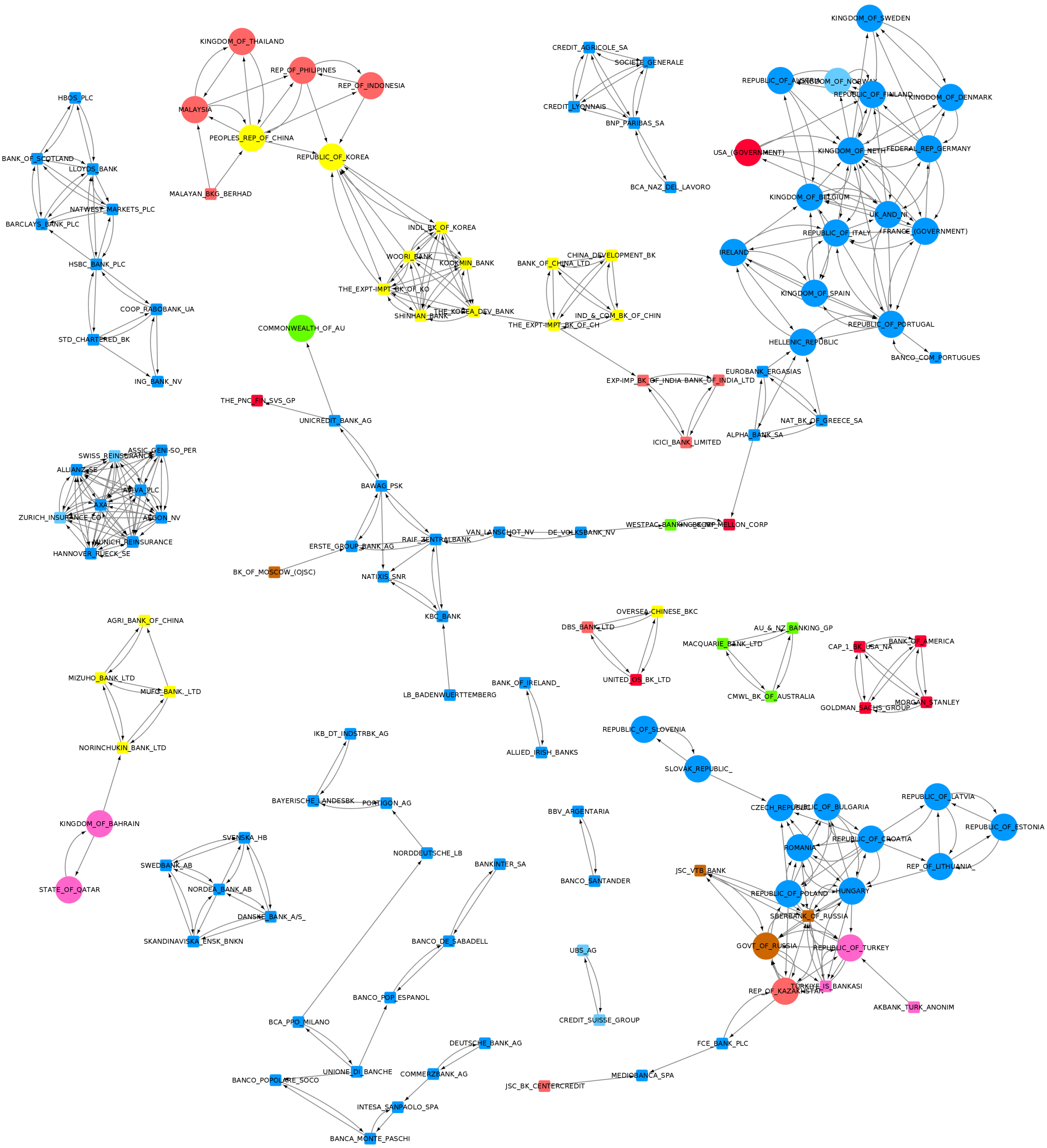


Figure 5.6: SVNs using biased generalized FEVD coefficients and original data: sub-period 09/2011-01/2013.
Squared nodes are financial companies. Colours: Yellow=Asia; Blue=EU and UE (EU with own currency); Pink=Middle East; Light red= Other Asian; Green=Oceania; Light blue= European not in EU; Brown=Russia; Red=US.

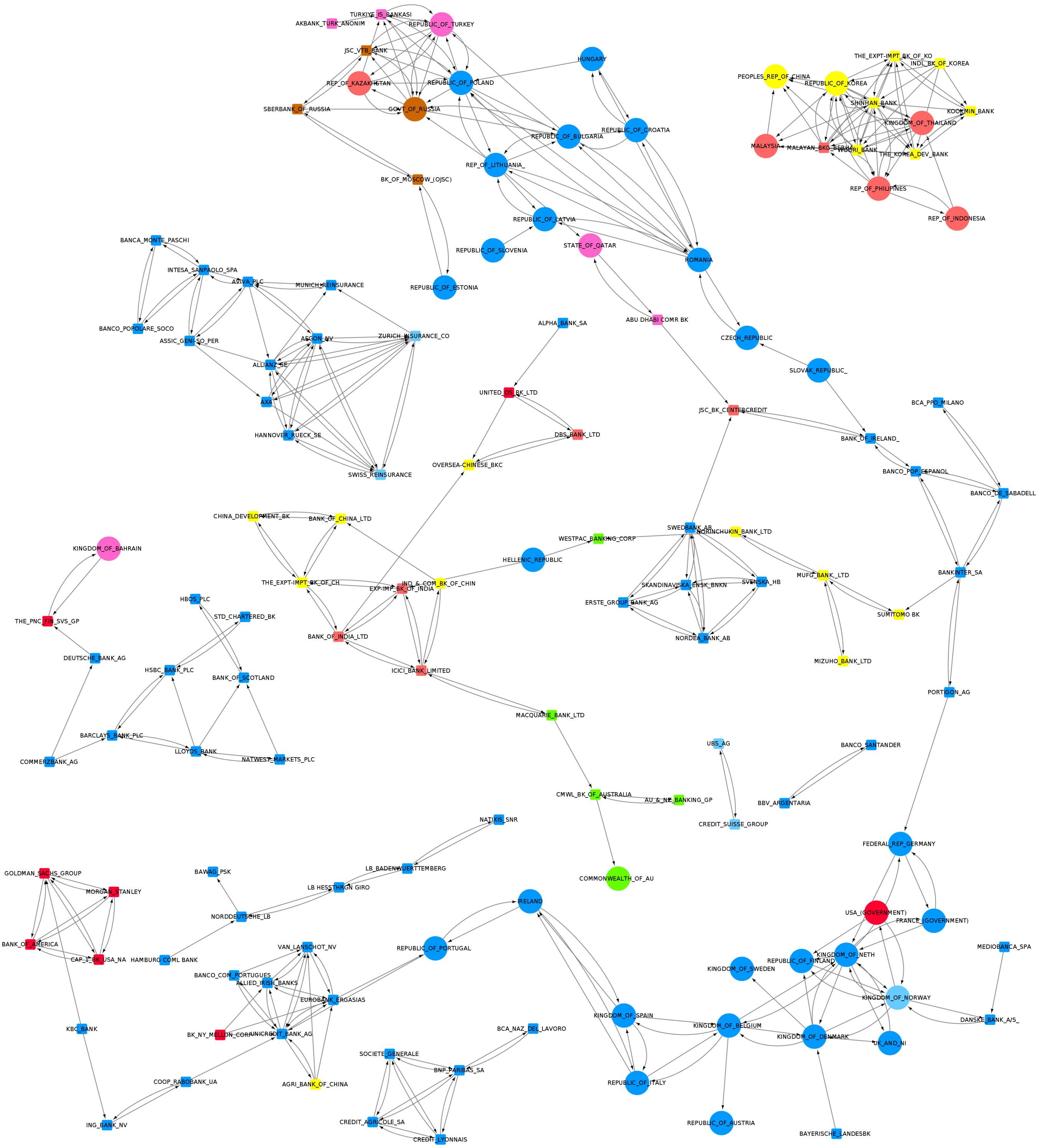


Figure 5.7: SVNs using biased generalized FEVD coefficients and original data: sub-period 02/2013-07/2016.
Squared nodes are financial companies. Colours: Yellow=Asia; Blue=EU and UE (EU with own currency); Pink=Middle East; Light red= Other Asian; Green=Oceania; Light blue= European not in EU; Brown=Russia; Red=US.

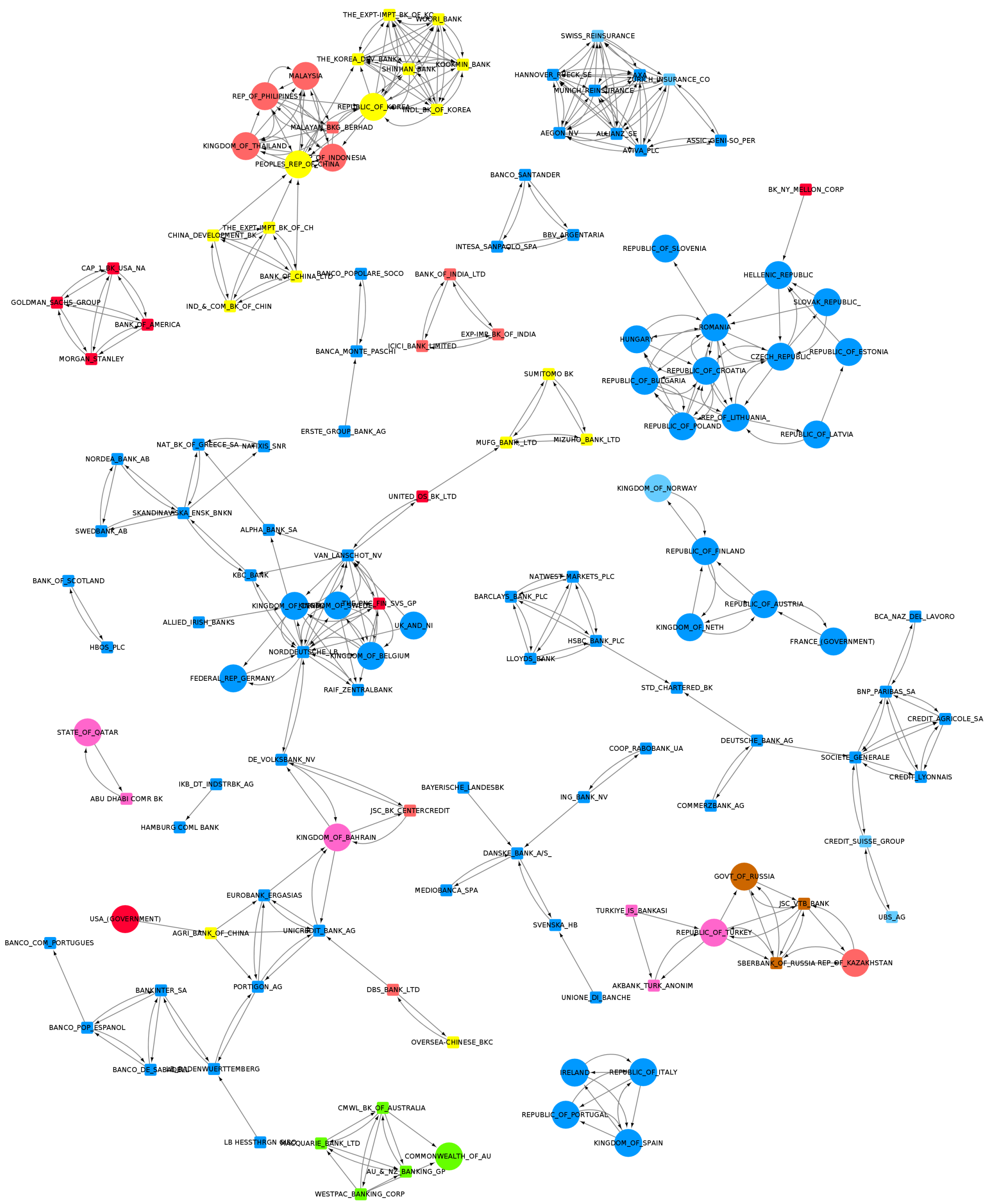


Figure 5.8: SVNs using orthogonalized FEVD coefficients: links stable in at least 90 permutations: sub-period 04/2009-08/2011.
Squared nodes are financial companies. Colours: Yellow=Asia; Blue=EU and UE (EU with own currency); Pink=Middle East; Light red= Other Asian; Green=Oceania; Light blue= European not in EU; Brown=Russia; Red=US.

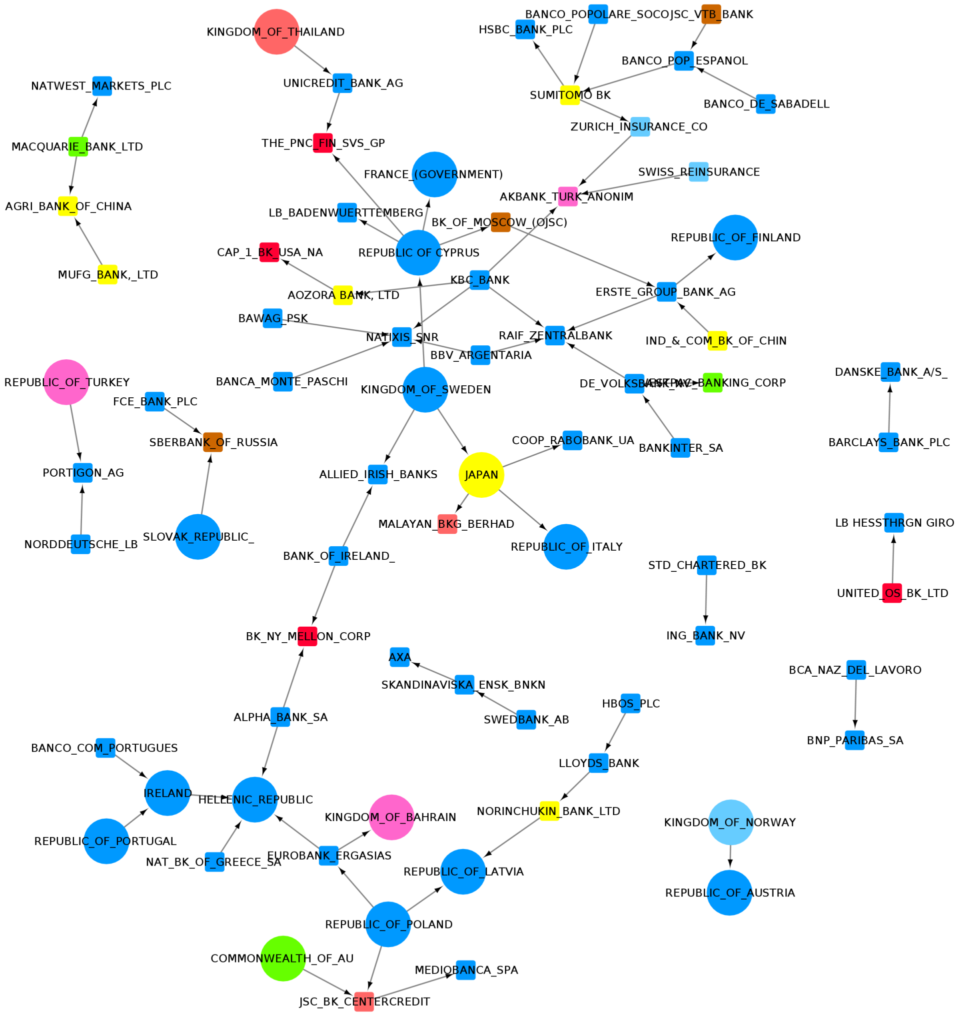


Figure 5.9: SVNs using orthogonalized FEVD coefficients: links stable in at least 90 permutations: sub-period 09/2011-01/2013.
Squared nodes are financial companies. Colours: Yellow=Asia; Blue=EU and UE (EU with own currency); Pink=Middle East; Light red= Other Asian; Green=Oceania; Light blue= European not in EU; Brown=Russia; Red=US.

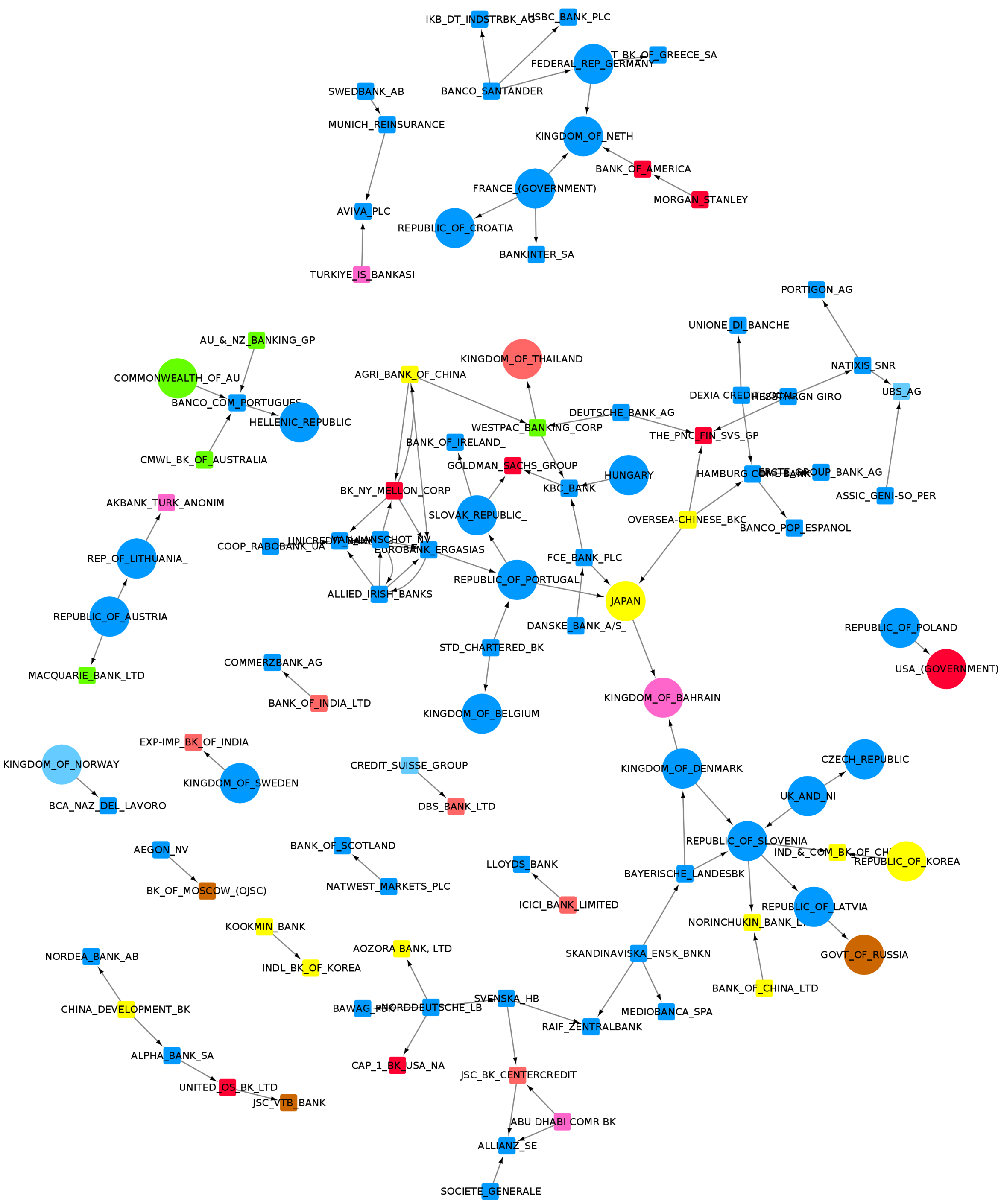


Figure 5.10: SVNs using orthogonalized FEVD coefficients: links stable in at least 90 permutations: sub-period 02/2013-07/2016.
Squared nodes are financial companies. Colours: Yellow=Asia; Blue=EU and UE (EU with own currency); Pink=Middle East; Light red= Other Asian; Green=Oceania; Light blue= European not in EU; Brown=Russia; Red=US.

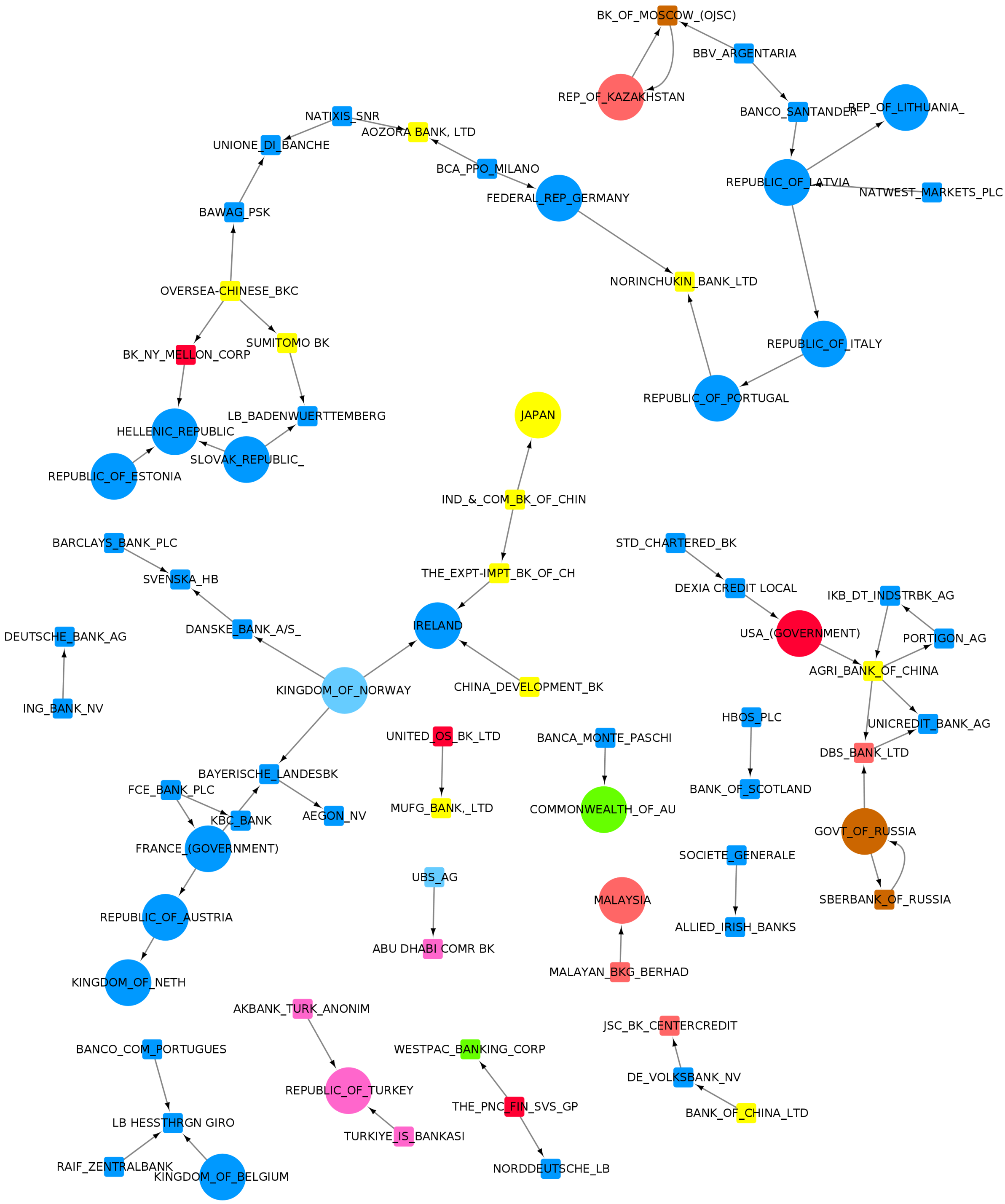


Figure 5.11: SVNs using unbiased generalized FEVD coefficients and original data: sub-period 05/2009-08/2011.
Squared nodes are financial companies. Colours: Yellow=Asia; Blue=EU and UE (EU with own currency); Pink=Middle East; Light red= Other Asian; Green=Oceania; Light blue= European not in EU; Brown=Russia; Red=US.

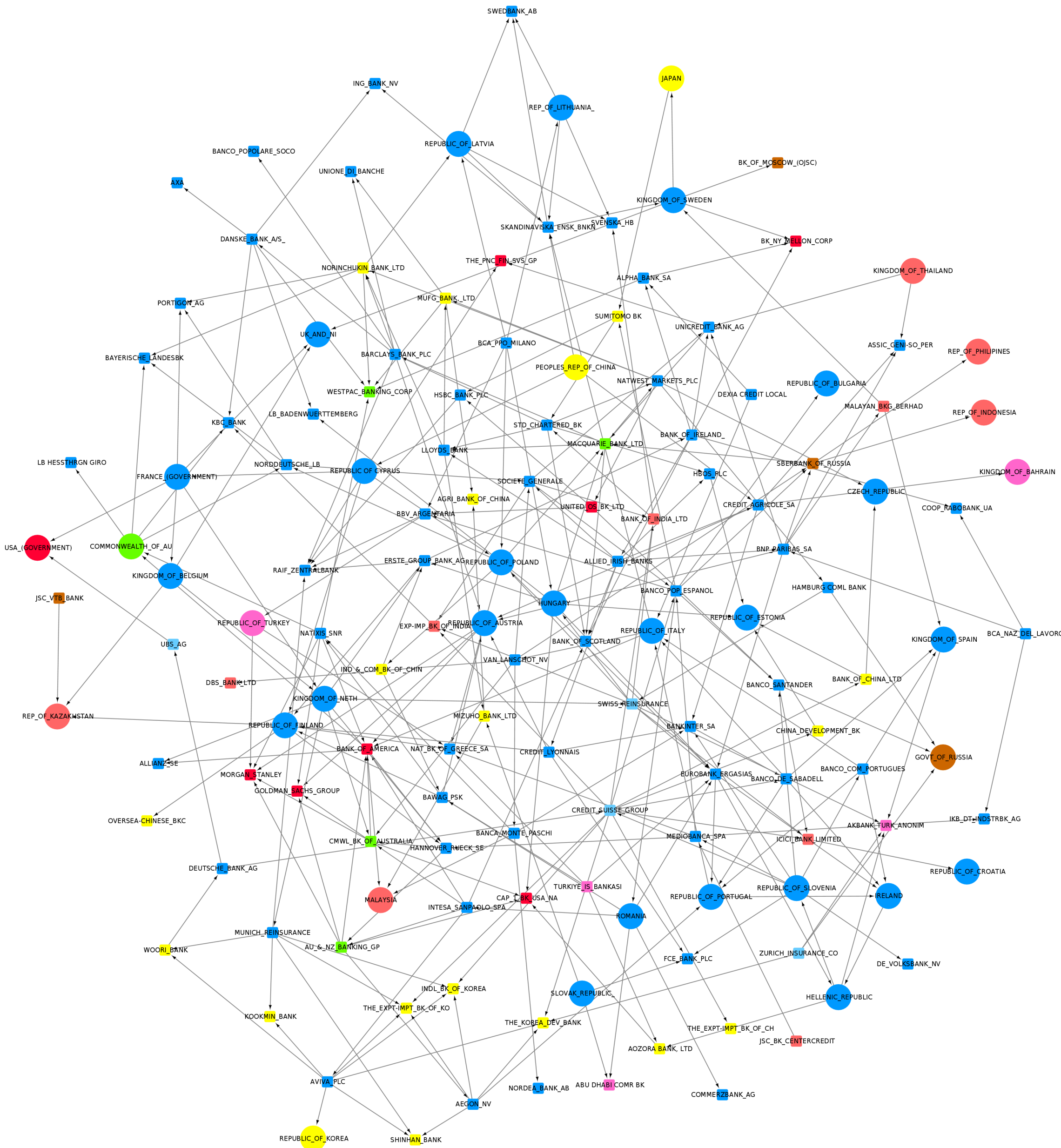


Figure 5.12: SVNs using unbiased generalized FEVD coefficients and original data: sub-period 09/2011-01/2013.
Squared nodes are financial companies. Colours: Yellow=Asia; Blue=EU and UE (EU with own currency); Pink=Middle East; Light red= Other Asian; Green=Oceania; Light blue= European not in EU; Brown=Russia; Red=US.

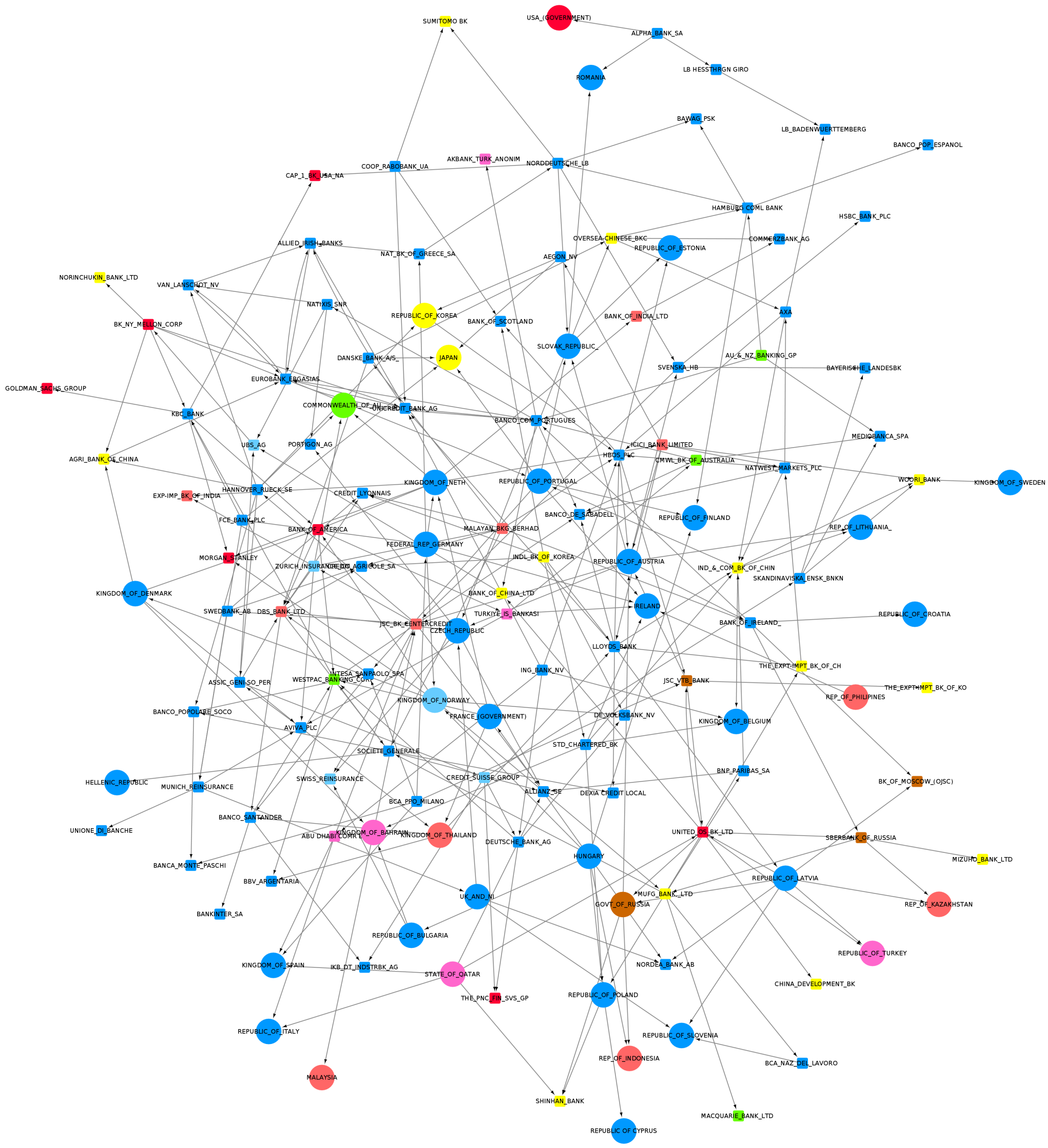
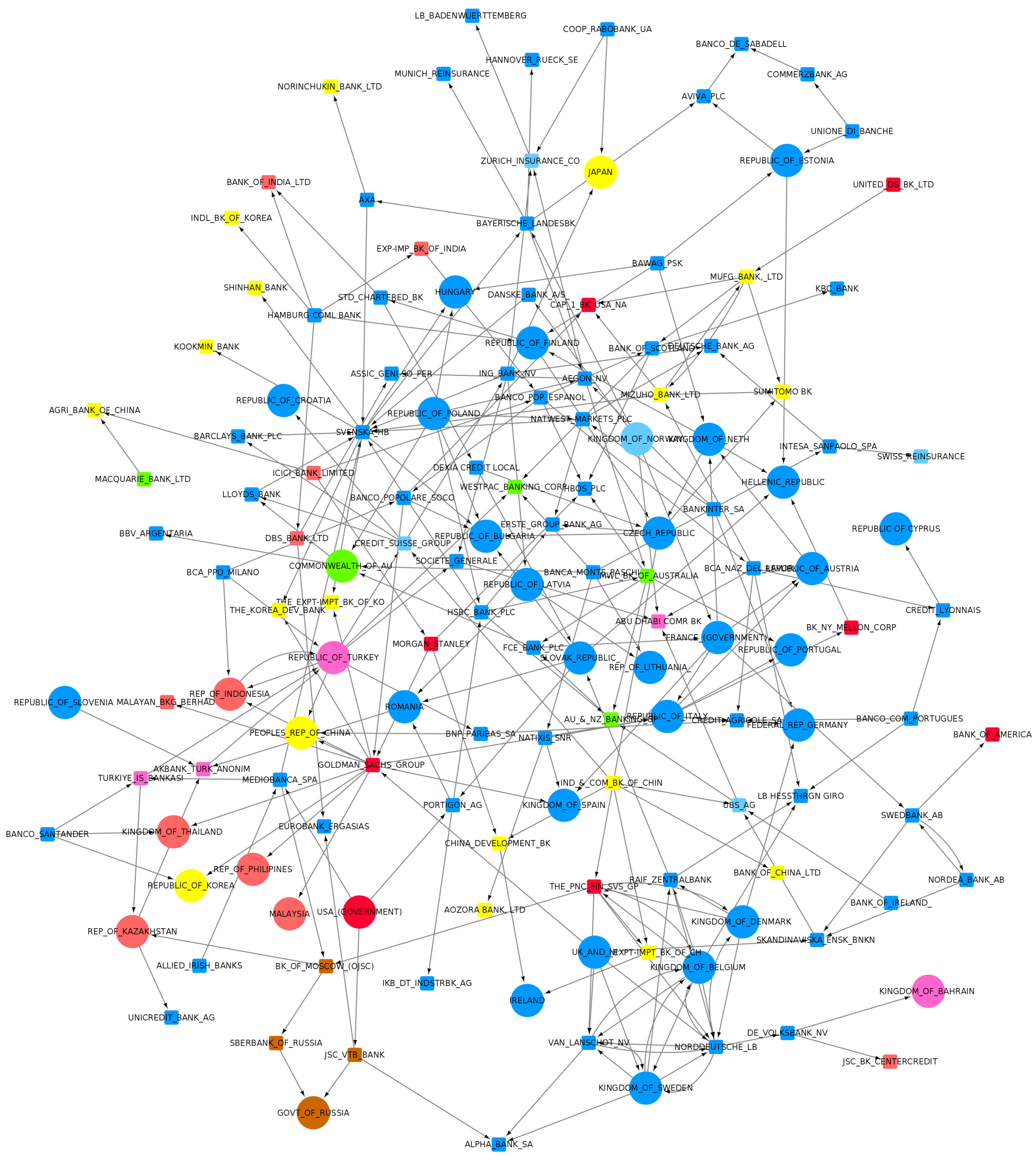


Figure 5.13: SVNs using unbiased generalized FEVD coefficients and original data: sub-period 02/2013-07/2016.
Squared nodes are financial companies. Colours: Yellow=Asia; Blue=EU and UE (EU with own currency); Pink=Middle East; Light red= Other Asian; Green=Oceania; Light blue= European not in EU; Brown=Russia; Red=US.



Appendix A

Table A1: US universities: means of the outcome variables of Tab. 4.1 *over the pre-treatment years (2001-2007)*. Universities are listed in a decreasing order according to the pre-treatment mean of the number of publications.

Universities of US	**N. publications in journals** (Means)	**N. publications in $3^*, 4^*, 4^{**}$ journals** (Means)	**Average N. of coauthors per paper** (Means)	**N. of affiliated authors-Interpolated** (Means)	**N. of papers per author** (Means)
Harvard University	297.29	241.00	2.20	587.72	1.66
University of California-Berkeley	241.00	189.29	2.18	425.68	1.74
University of Pennsylvania	238.14	203.71	2.26	388.48	1.72
University of Michigan	213.43	168.29	2.37	495.12	1.45
Columbia University	206.29	154.29	2.19	350.27	1.71
New York University (NYU)	203.43	164.86	2.15	340.20	1.68
Pennsylvania State University	197.43	132.00	2.43	469.41	1.37
Cornell University	194.71	137.43	2.24	370.92	1.49
Massachusetts Institute of Technology (MIT)	190.57	156.71	2.26	337.10	1.76
University of Illinois at Urbana-Champaign	188.00	134.86	2.31	397.12	1.49
Stanford University	186.00	145.14	2.17	369.01	1.61
University of Maryland	179.00	142.14	2.38	358.30	1.55
Texas A&M University	173.00	116.00	2.44	367.03	1.45
Michigan State University	167.71	139.14	2.42	359.96	1.50
Northwestern University	162.86	128.00	2.19	257.38	1.69
Ohio State University	152.86	94.71	2.29	295.42	1.52
Rutgers University-New Brunswick	151.86	100.86	2.30	333.17	1.51
University of Wisconsin-Madison	151.43	106.00	2.19	303.88	1.50
University of California-Los Angeles	151.14	116.00	2.23	277.48	1.68
University of Chicago	151.00	126.71	2.07	233.93	1.82
University of Texas-Austin	148.43	126.86	2.41	308.32	1.52
Indiana University	145.86	96.14	2.28	333.52	1.45
Arizona State University	139.57	109.29	2.61	319.03	1.36
Purdue University	139.43	94.29	2.54	316.83	1.37
University of Florida	135.00	99.57	2.63	269.07	1.36
Duke University	131.43	112.14	2.35	243.15	1.56
Yale University	130.00	89.29	2.17	215.34	1.88
University of Minnesota	124.29	85.14	2.44	269.67	1.56
University of Southern California	119.71	97.14	2.36	243.44	1.54
University of Washington	119.43	82.86	2.43	273.48	1.41
University of North Carolina-Chapel-Hill	115.86	84.43	2.48	239.66	1.48
University of Georgia	113.43	67.71	2.47	240.31	1.36
Georgia Institute of Technology	106.71	86.29	2.65	259.43	1.28
Iowa State University	104.57	72.71	2.33	220.67	1.30
Georgia State University	102.86	73.57	2.37	190.27	1.56
Carnegie Mellon University	101.29	91.14	2.51	255.13	1.40
University of California-Davis	100.14	66.71	2.33	180.32	1.63
North Carolina State University	100.00	58.43	2.49	224.28	1.33
University of Arizona	99.29	71.86	2.57	196.50	1.48
Princeton University	99.00	77.71	2.11	172.32	1.64
George Mason University	96.57	47.57	2.08	189.35	1.66
City University of New York	89.57	52.14	2.03	212.46	1.54
Florida State University	89.29	53.29	2.43	186.06	1.41
University of Connecticut	87.86	69.71	2.55	168.31	1.38
George Washington University	86.14	40.29	2.01	183.33	1.48
University of California-Irvine	80.43	56.14	2.10	150.65	1.61
University of Central Florida	79.86	55.29	2.46	180.64	1.46
Boston University	78.14	63.29	2.21	165.89	1.55
University of Colorado at Boulder	74.71	47.00	2.42	135.32	1.68
University of Pittsburgh	73.86	51.29	2.28	176.43	1.48
University of Missouri	73.29	42.29	2.36	157.02	1.34
Louisiana State University	73.14	43.71	2.38	157.93	1.41
University of Virginia	73.00	48.29	2.34	161.21	1.59
Auburn University	71.71	33.14	2.36	157.30	1.37
Syracuse University	70.86	44.57	2.21	145.44	1.51
University of South Carolina	69.86	49.14	2.39	147.20	1.43
Georgetown University	68.00	46.29	2.10	136.93	1.53
Emory University	67.14	51.86	2.38	124.64	1.59
Boston College	65.29	52.14	2.23	119.18	1.66
University of California-San Diego	64.71	44.86	2.13	97.66	1.96
University of Illinois at Chicago	64.57	34.43	2.26	154.29	1.44
University of Houston	63.00	42.57	2.43	138.91	1.58
University of Iowa	62.86	50.29	2.38	144.89	1.34
University of Alabama-Tuscaloosa	60.43	34.00	2.50	143.22	1.48
Vanderbilt University	59.86	34.57	2.12	100.82	1.67
Johns Hopkins University	59.14	34.29	2.36	151.12	1.34

Universities of US	N. publications in journals (Means)	N. publications in $3^*, 4^*, 4^{**}$ journals (Means)	Average N. of coauthors per paper (Means)	N. of affiliated authors-Interpolated (Means)	N. of papers per author (Means)
University of Kentucky	59.00	37.71	2.48	132.94	1.45
University of Texas-Dallas	57.71	54.00	2.54	95.09	1.62
University of Miami	57.43	37.00	2.29	107.47	1.70
Washington University in St. Louis	56.43	45.29	2.28	109.82	1.64
Dartmouth College	55.57	42.57	2.18	86.95	1.74
University of Notre Dame	55.43	38.14	2.15	118.02	1.38
Colorado State University	55.00	23.29	2.35	117.50	1.47
University of Oklahoma	55.00	40.29	2.72	122.67	1.29
Rensselaer Polytechnic Institute	54.57	42.43	2.38	94.94	1.67
Temple University	54.43	38.14	2.31	104.51	1.65
Clemson University	53.86	37.00	2.52	116.35	1.41
University of Rochester	52.29	46.14	2.18	103.05	1.51
University of Tennessee-Knoxville	52.00	23.14	2.40	108.61	1.47
State University of New York-Buffalo	51.29	41.86	2.32	109.76	1.42
Southern Methodist University	49.86	37.57	2.31	85.94	1.49
University of Delaware	49.00	24.71	2.23	106.58	1.34
Rice University	48.29	38.57	2.31	89.76	1.52
Case Western Reserve University	48.14	38.29	2.16	94.02	1.52
University of Massachusetts-Amherst	48.14	27.29	2.33	148.07	1.34
Drexel University	46.57	29.43	2.37	99.87	1.36
Oklahoma State University	45.71	24.57	2.51	89.37	1.53
Brigham Young University	43.86	33.43	2.33	120.26	1.13
Brown University	42.71	32.00	2.07	77.87	1.82
Florida Atlantic University	41.71	24.29	2.54	82.31	1.55
University of California-Santa Barbara	41.57	24.14	2.16	88.33	1.58
American University	41.29	19.86	2.20	78.13	1.64
University of Oregon	40.57	26.00	2.29	79.60	1.59
University of California-Riverside	40.43	23.29	2.32	75.66	1.56
University of Kansas	39.14	23.00	2.29	83.64	1.44
University of Wyoming	38.00	29.71	2.23	50.45	2.00
University of Hawaii-Manoa	35.43	16.57	2.14	77.91	1.58
West Virginia University	35.29	13.14	2.24	88.36	1.37
State University of New York-Binghamton	35.00	22.71	2.31	70.77	1.53
Virginia Commonwealth University	33.43	17.29	2.44	84.62	1.30
California Institute of Technology	33.43	25.86	2.52	53.69	2.09
Utah State University	31.29	16.29	2.61	64.53	1.43
Tufts University	30.86	16.71	2.02	54.05	1.77
University of Colorado at Denver	30.57	21.43	2.37	67.71	1.40
Fordham University	29.29	12.14	1.91	52.97	1.75
Tulane University	28.86	22.29	2.31	65.93	1.45
College of William & Mary	28.71	20.43	2.11	58.53	1.46
State University of New York-Albany	28.00	15.57	2.07	82.34	1.35
University of Nevada-Reno	27.86	14.29	2.26	62.11	1.43
Baylor University	27.29	20.14	2.44	52.47	1.37
DePaul University	26.14	17.00	2.28	68.81	1.37
Santa Clara University	25.29	17.00	2.12	48.44	1.46
University of North Carolina-Greensboro	25.00	14.00	2.23	62.98	1.30
University of California-Santa Cruz	23.86	15.43	2.21	35.06	2.02
Stony Brook University	21.43	11.71	2.31	38.47	1.60
University of Maryland-Baltimore County	18.43	10.29	2.54	51.75	1.31
Appalachian State University	18.00	7.43	2.28	46.24	1.26
Brandeis University	16.00	12.43	2.66	34.73	1.56
Middlebury College	13.57	3.86	1.83	23.55	1.79
Williams College	11.86	7.29	2.11	24.77	1.58
Claremont McKenna College	11.43	8.14	2.06	18.28	1.60

Table A2: UK universities: means of the outcome variables of Tab. 4.1 *over the pre-treatment years (2001-2007)*. Universities are listed in a decreasing order according to the mean of the number of publications.

Universities of UK	**N. publications in journals** (Means)	**N. publications in** $3^*, 4^*, 4^{**}$ **journals** (Means)	**Average N. of coauthors per paper** (Means)	**N. of affiliated authors-Interpolated** (Means)	**N. of papers per author** (Means)
University of Manchester	181.00	126.86	2.28	427.56	1.40
London School of Economics and Political Science	168.57	115.14	2.00	341.11	1.72
University of Warwick	148.43	109.86	2.16	279.70	1.63
University of Oxford	147.71	89.86	1.97	293.14	1.80
University of Nottingham	147.43	115.57	2.32	273.41	1.53
University of Cambridge	127.57	88.29	2.09	284.88	1.74
Cardiff University	110.29	85.57	2.27	233.36	1.33
University College London	89.29	62.71	2.51	189.63	1.66
Imperial College London	83.86	56.43	2.49	184.81	1.39
University of Leeds	81.29	57.14	2.20	200.73	1.35
Lancaster University	79.86	62.29	2.16	192.84	1.38
University of Strathclyde	78.00	50.86	2.28	162.22	1.47
University of Birmingham	77.86	38.57	2.11	187.41	1.64
University of Sheffield	75.86	35.86	2.44	177.22	1.46
London Business School	74.14	57.29	2.11	124.21	1.72
University of Southampton	73.14	55.29	2.50	172.68	1.34
City University London	70.86	55.00	2.24	158.48	1.47
Cranfield University	68.57	43.86	2.28	156.98	1.22
University of Reading	68.29	35.43	2.14	150.86	1.58
University of York	68.14	37.71	2.16	141.64	1.52
Brunel University London	68.00	32.57	2.47	149.81	1.44
University of Bath	67.43	49.86	2.28	141.71	1.38
University of Edinburgh	58.00	38.71	2.25	139.28	1.50
Aston University	55.29	38.00	2.51	87.04	1.64
Newcastle University	55.00	29.71	2.43	149.20	1.29
University of Exeter	52.43	35.43	2.24	110.13	1.50
University of Surrey	51.71	28.14	2.42	121.56	1.47
University of Essex	51.29	42.57	2.04	115.11	1.50
University of Leicester	51.00	27.86	2.17	114.10	1.50
University of Glasgow	49.00	28.43	2.29	120.46	1.44
University of Durham	47.86	28.43	2.10	105.02	1.53
University of Bristol	46.43	33.14	2.45	114.86	1.41
University of East Anglia	43.14	32.00	2.23	108.31	1.39
University of Ulster	42.14	15.57	2.46	98.45	1.33
University of Aberdeen	40.29	24.57	2.32	103.63	1.39
University of Sussex	39.57	25.29	1.91	122.16	1.67
King's College London	39.43	30.86	2.54	95.29	1.49
Queen Mary University of London	37.86	25.57	2.41	62.75	2.04
University of Salford	37.71	14.00	2.29	112.40	1.12
Royal Holloway, University of London	36.86	22.00	2.14	72.03	1.60
Manchester Metropolitan University	36.57	12.86	2.12	96.77	1.35
University of Stirling	36.29	21.86	2.18	72.77	1.58
University of Bradford	36.14	17.57	2.24	79.48	1.63
Open University	35.14	13.57	2.04	115.80	1.30
University of Kent	34.57	20.29	2.18	70.34	1.61
Birkbeck College	34.43	22.86	2.05	65.05	1.93
University of Liverpool	34.43	19.14	2.47	101.33	1.30
Queen's University Belfast	33.86	17.00	2.29	84.48	1.36
Heriot-Watt University	32.71	10.43	2.29	84.68	1.39
University of Hull	32.00	11.14	1.99	68.76	1.50
Middlesex University	30.86	10.43	2.30	66.71	1.46
Swansea University	29.29	14.14	1.90	52.52	1.57
University of St Andrews	29.29	17.86	2.10	53.52	1.66
University of Portsmouth	29.14	16.14	2.54	67.58	1.21
University of the West of England, Bristol	27.14	12.00	2.03	79.77	1.39
Glasgow Caledonian University	25.14	10.43	2.29	76.76	1.25
University of Dundee	24.00	18.14	2.48	62.50	1.07
University of Plymouth	24.00	10.29	2.42	68.85	1.21
London Metropolitan University	23.71	8.86	1.78	61.71	1.42
Oxford Brookes University	22.71	6.71	1.93	53.48	1.58
De Montfort University	21.57	16.29	2.64	54.67	1.50
Sheffield Hallam University	20.57	8.86	2.04	61.78	1.46
Kingston University	20.29	8.57	2.09	61.59	1.37
Nottingham Trent University	19.71	8.43	1.98	54.09	1.53
Keele University	18.43	10.57	1.84	41.27	1.68
University of Westminster	17.86	6.29	2.40	48.38	1.37
Edinburgh Napier University	16.86	4.00	2.35	48.49	1.27
Leeds Beckett University	16.43	3.71	2.11	45.73	1.50
Coventry University	16.29	8.00	2.48	43.71	1.39
University of South Wales	16.14	6.00	2.70	50.21	1.34
London South Bank University	15.71	6.86	2.18	42.27	1.64
University of Northumbria at Newcastle	15.00	6.00	1.94	51.40	1.36
University of Wolverhampton	14.29	2.14	2.07	43.96	1.34
University of Brighton	14.29	7.14	2.39	39.38	1.61
Aberystwyth University	13.57	7.43	2.27	33.46	1.47
University of Greenwich	13.29	6.14	2.24	36.36	1.59
University of Hertfordshire	12.71	7.57	1.95	36.97	2.13
Bournemouth University	12.57	3.57	1.93	38.38	1.31
Bangor University	12.43	6.29	2.25	20.60	1.76
Robert Gordon University	11.57	6.57	2.24	31.53	1.28
University of Central Lancashire	11.14	4.57	1.63	23.83	2.71
University of Sunderland	8.71	1.71	2.03	19.86	1.51
University of Bedfordshire	8.43	2.71	2.46	19.85	1.62
University of East London	8.43	2.29	2.28	20.97	1.55
Staffordshire University	8.00	2.71	1.94	23.70	1.47

Table A3: US universities: means of the outcome variables of Tab. 4.1 *over the post-treatment years (2008-2015)*. Universities are listed in a decreasing order according to mean of the number of publications.

Universities of US	**N. publications in journals** (Means)	**N. publications in $3^*, 4^*, 4^{**}$ journals** (Means)	**Average N. of coauthors per paper** (Means)	**N. of affiliated authors-Interpolated** (Means)	**N. of papers per author** (Means)
Harvard University	454.38	309.25	2.37	748.65	1.89
University of Michigan	360.75	228.88	2.58	634.60	1.63
Pennsylvania State University	350.25	190.25	2.80	635.10	1.59
Texas A&M University	343.50	177.38	2.80	613.65	1.57
University of California-Berkeley	341.12	216.75	2.42	540.37	1.93
Stanford University	330.50	211.75	2.58	553.75	1.83
Columbia University	330.50	208.50	2.39	483.57	2.02
University of Pennsylvania	324.12	236.75	2.55	500.74	1.91
Cornell University	307.00	177.38	2.47	496.00	1.77
New York University (NYU)	294.62	194.88	2.41	476.98	1.88
Indiana University	287.50	159.88	2.59	515.77	1.69
University of Illinois at Urbana-Champaign	280.88	151.00	2.66	517.14	1.61
Michigan State University	274.88	172.50	2.83	479.50	1.54
Massachusetts Institute of Technology (MIT)	273.50	201.62	2.63	444.14	1.86
Arizona State University	259.38	161.12	2.82	453.68	1.63
Purdue University	252.00	141.50	2.77	473.63	1.53
Rutgers University-New Brunswick	241.75	130.00	2.57	434.40	1.66
University of Maryland	239.38	167.75	2.74	462.81	1.68
Northwestern University	238.75	178.62	2.40	368.70	1.74
University of Chicago	237.50	173.38	2.37	316.60	2.03
University of Florida	236.62	131.12	2.81	415.05	1.58
Ohio State University	234.38	133.38	2.72	398.25	1.62
Duke University	231.75	169.88	2.71	356.12	1.73
University of Texas-Austin	228.38	147.62	2.69	407.71	1.59
University of Wisconsin-Madison	221.38	126.25	2.67	400.15	1.66
University of Washington	216.00	116.88	2.83	399.07	1.63
Yale University	202.12	126.62	2.60	275.79	2.22
University of Southern California	198.25	121.75	2.45	356.74	1.67
University of Georgia	194.88	105.12	2.72	343.00	1.54
University of California-Los Angeles	191.88	111.62	2.45	305.30	1.88
Georgia Institute of Technology	182.50	129.62	2.86	373.13	1.50
University of North Carolina-Chapel-Hill	178.38	115.25	2.75	317.70	1.62
City University of New York	169.88	73.12	2.34	353.14	1.54
George Mason University	168.88	70.88	2.27	314.92	1.84
Georgia State University	167.50	99.50	2.68	249.64	1.87
University of Minnesota	166.75	96.62	2.72	287.28	1.78
Florida State University	158.75	94.62	2.83	244.64	1.64
Princeton University	158.50	97.12	2.29	247.64	1.96
North Carolina State University	158.38	75.62	2.80	309.97	1.48
Iowa State University	155.75	80.25	2.75	301.10	1.37
Carnegie Mellon University	153.25	118.62	2.79	310.83	1.51
University of California-Davis	152.62	103.62	2.61	250.65	1.72
George Washington University	150.38	66.00	2.54	269.15	1.69
University of Connecticut	139.75	80.25	2.70	223.03	1.60
Temple University	136.25	81.62	2.61	194.43	1.84
University of South Carolina	134.25	78.00	2.76	232.61	1.56
Boston University	133.12	77.62	2.37	217.85	1.72
University of Arizona	130.62	80.50	2.91	234.27	1.76
University of Central Florida	123.75	62.75	2.65	229.03	1.52
University of Virginia	122.38	79.88	2.48	229.51	1.62
University of Alabama-Tuscaloosa	121.25	56.88	2.78	213.06	1.65
University of California-San Diego	120.00	76.25	2.37	184.77	1.89
University of Texas-Dallas	119.38	102.12	2.91	159.69	1.91
Auburn University	118.62	46.62	2.78	209.47	1.48
Johns Hopkins University	118.12	57.62	2.67	220.34	1.70

Universities of US	**N. publications in journals** (Means)	**N. publications in** $3^*, 4^*, 4^{**}$ **journals** (Means)	**Average N. of coauthors per paper** (Means)	**N. of affiliated authors-Interpolated** (Means)	**N. of papers per author** (Means)
University of Houston	115.00	67.88	2.70	216.79	1.66
University of California-Irvine	114.50	65.50	2.39	193.42	1.81
Clemson University	112.50	58.12	2.86	204.23	1.48
Syracuse University	112.00	59.75	2.54	190.32	1.76
University of Pittsburgh	111.88	74.12	2.76	233.19	1.45
University of Colorado at Boulder	109.50	59.25	2.64	195.11	1.67
Boston College	104.50	74.62	2.42	153.51	1.85
University of Tennessee-Knoxville	104.25	41.38	2.88	172.58	1.74
Colorado State University	103.00	42.00	2.82	197.35	1.50
University of Iowa	102.00	61.25	2.64	177.91	1.57
University of Kentucky	101.25	43.12	2.74	207.16	1.41
University of Missouri	99.25	45.62	2.61	173.13	1.77
Louisiana State University	99.12	38.38	2.73	185.51	1.58
University of Massachusetts-Amherst	96.75	45.38	2.52	197.47	1.64
University of Illinois at Chicago	94.88	45.62	2.67	197.27	1.61
Vanderbilt University	92.75	53.62	2.49	155.10	1.62
Georgetown University	91.00	53.50	2.46	167.60	1.67
Washington University in St. Louis	90.62	67.00	2.53	150.24	1.73
University of Oklahoma	89.38	53.62	2.97	187.30	1.49
Drexel University	88.25	45.62	2.70	153.84	1.67
State University of New York-Buffalo	86.12	50.38	2.71	168.60	1.53
University of Miami	84.38	54.50	2.81	155.67	1.59
Emory University	81.38	54.25	2.68	154.14	1.62
Brigham Young University	80.38	46.75	2.75	177.25	1.30
American University	79.88	36.25	2.34	133.82	1.76
Rice University	79.00	53.12	2.46	110.74	1.82
University of Notre Dame	78.88	53.88	2.37	135.83	1.71
University of Kansas	77.38	37.38	2.67	149.34	1.51
West Virginia University	76.12	26.12	2.55	139.94	1.55
Fordham University	75.00	37.50	2.38	100.32	1.84
University of Hawaii-Manoa	74.00	25.88	2.53	129.88	1.76
University of Rochester	73.00	49.50	2.33	120.39	1.55
University of California-Santa Barbara	72.50	28.75	2.66	131.61	1.78
University of Delaware	71.12	30.00	2.67	148.86	1.43
Oklahoma State University	70.50	33.38	2.92	123.28	1.74
Southern Methodist University	70.25	38.75	2.38	111.00	1.76
Dartmouth College	69.25	52.12	2.55	101.27	1.85
University of Oregon	65.88	32.62	2.34	124.32	1.70
Virginia Commonwealth University	64.62	28.62	2.77	123.74	1.42
University of California-Riverside	62.88	32.75	2.56	86.83	1.90
Florida Atlantic University	62.75	27.50	2.93	134.62	1.48
Brown University	62.12	34.62	2.37	98.80	1.87
DePaul University	60.62	25.50	2.48	134.49	1.39
State University of New York-Albany	59.00	31.50	2.48	120.58	1.46
State University of New York-Binghamton	57.25	37.12	2.76	108.04	1.73
University of North Carolina-Greensboro	56.38	21.25	2.58	93.65	1.74
Rensselaer Polytechnic Institute	54.88	38.88	2.59	95.51	1.62
Utah State University	52.12	25.62	2.67	98.08	1.56
University of Colorado at Denver	52.00	23.88	2.55	99.96	1.55
Case Western Reserve University	50.00	27.50	2.62	84.63	1.88
University of Wyoming	46.62	24.75	2.69	69.02	1.71
Santa Clara University	45.75	28.75	2.31	71.94	1.67
Baylor University	45.62	24.88	2.80	87.00	1.42
University of California-Santa Cruz	44.75	22.62	2.44	61.93	2.20
Appalachian State University	44.38	12.38	2.64	83.60	1.42
College of William & Mary	43.12	23.00	2.44	82.58	1.48
California Institute of Technology	39.88	27.75	2.61	67.24	1.84
Tulane University	39.62	17.50	2.50	68.37	1.93
Stony Brook University	39.00	14.12	2.64	60.70	1.90
University of Nevada-Reno	36.12	11.25	2.71	81.66	1.50
Tufts University	35.00	17.38	2.18	66.92	1.97
University of Maryland-Baltimore County	28.50	8.88	2.44	59.57	1.63
Brandeis University	26.12	12.00	2.46	46.32	1.84
Claremont McKenna College	22.38	12.75	2.31	29.29	1.78
Middlebury College	19.75	8.50	2.15	28.95	2.13
Williams College	12.62	6.38	2.14	23.57	1.62

Table A4: UK universities: means of the outcome variables of Tab. 4.1 *over the post-treatment years (2008-2015)*. Universities are listed in a decreasing order according to the mean of the number of publications.

Universities of UK	**N. publications in journals** (Means)	**N. publications in 3*, 4*, 4** journals** (Means)	**Average N. of coauthors per paper** (Means)	**N. of affiliated authors-Interpolated** (Means)	**N. of papers per author** (Means)
University of Manchester	323.38	179.00	2.49	615.38	1.79
University of Oxford	322].00	157.62	2.34	559.23	1.98
London School of Economics and Political Science	298.38	171.88	2.22	518.33	1.94
University of Cambridge	281.25	145.88	2.46	518.55	1.92
University of Warwick	271.50	166.00	2.45	429.53	1.84
University of Nottingham	233.50	167.00	2.76	418.17	1.63
Cardiff University	178.38	116.12	2.51	306.05	1.70
University College London	175.38	84.62	2.63	334.96	1.87
Lancaster University	175.00	122.25	2.58	312.01	1.65
University of Leeds	154.00	87.38	2.68	300.88	1.63
City University London	152.12	100.38	2.51	225.32	1.90
University of Southampton	150.12	84.25	2.64	286.33	1.56
University of Birmingham	149.62	59.25	2.36	267.89	1.84
Imperial College London	143.00	93.12	2.83	262.14	1.73
University of Bath	133.00	69.50	2.57	215.05	1.76
University of Sheffield	129.75	61.88	2.86	238.90	1.79
University of Strathclyde	126.75	75.88	2.67	228.35	1.74
Brunel University London	122.25	59.25	2.76	223.77	1.68
University of Edinburgh	116.62	54.38	2.57	242.65	1.69
University of Reading	110.38	53.12	2.49	237.42	1.69
University of Essex	108.50	71.12	2.52	183.72	1.65
University of Surrey	100.75	56.25	2.67	173.78	1.78
University of Glasgow	100.38	51.38	2.46	226.86	1.68
Cranfield University	100.00	51.88	2.83	217.50	1.37
University of East Anglia	98.62	56.00	2.80	182.15	1.77
University of Durham	98.50	48.62	2.61	173.35	1.84
University of Kent	97.75	53.12	2.53	159.40	1.75
University of York	97.62	44.75	2.55	186.74	1.73
London Business School	96.38	68.00	2.32	135.66	1.86
Aston University	95.00	64.50	2.78	138.85	1.91
University of Leicester	95.00	41.62	2.20	157.71	1.73
University of Bristol	94.75	49.62	2.39	176.91	1.64
Newcastle University	94.38	45.75	2.60	203.63	1.47
University of Exeter	93.25	53.00	2.77	182.60	1.70
University of Sussex	91.75	49.62	2.49	188.53	1.87
University of Liverpool	90.88	39.38	2.76	164.06	1.68
King's College London	89.38	38.12	2.48	183.31	1.92
Queen Mary University of London	87.38	46.00	2.38	138.86	1.88
Royal Holloway, University of London	77.75	39.62	2.35	126.73	1.93
Open University	75.38	29.88	2.40	188.80	1.50
University of Stirling	74.00	33.38	2.39	118.93	1.93
University of St Andrews	69.62	38.88	2.41	105.55	1.93
Queen's University Belfast	68.12	35.38	2.80	127.92	1.64
University of Hull	66.25	29.12	2.58	109.01	1.85
Heriot-Watt University	66.12	21.88	2.73	127.04	1.53
Middlesex University	61.88	26.38	2.38	122.68	1.72
Bournemouth University	60.75	20.88	2.44	112.79	1.72
University of the West of England, Bristol	60.25	18.62	2.53	139.15	1.59
University of Salford	55.25	20.38	2.68	131.49	1.57
Swansea University	53.25	20.75	2.62	93.96	1.60
University of Portsmouth	53.12	19.12	2.59	113.33	1.51
Manchester Metropolitan University	51.88	13.00	2.50	121.68	1.55
University of Ulster	51.62	17.88	2.91	146.60	1.20
University of Plymouth	50.12	13.88	2.77	113.89	1.38
Nottingham Trent University	50.00	14.62	2.33	101.90	1.63
Birkbeck College	49.00	20.75	2.34	91.20	1.99
University of Northumbria at Newcastle	47.12	9.88	2.33	117.43	1.55
University of Aberdeen	47.12	18.50	3.00	112.49	1.54
University of Bradford	46.88	23.50	2.57	108.02	1.56
Kingston University	45.75	17.38	2.55	93.40	1.79
Oxford Brookes University	45.62	14.38	2.37	94.72	1.66
University of Westminster	41.62	12.38	2.76	100.58	1.57
Leeds Beckett University	40.00	5.25	2.19	82.85	1.69
Bangor University	39.50	24.62	3.05	59.19	1.75
De Montfort University	36.50	13.50	2.68	78.33	1.62
London Metropolitan University	35.00	10.38	1.99	67.57	1.98
Sheffield Hallam University	34.38	8.38	2.32	92.82	1.48
University of Hertfordshire	33.88	14.12	2.17	71.23	1.72
University of Central Lancashire	32.50	8.12	2.54	68.15	1.79
University of Greenwich	31.38	10.00	2.52	73.32	1.69
Coventry University	31.25	7.75	2.93	90.44	1.62
University of Brighton	28.88	8.25	2.31	71.02	1.54
Glasgow Caledonian University	28.50	6.38	2.69	87.20	1.29
University of Dundee	27.88	12.50	2.83	65.25	1.40
University of South Wales	26.88	3.12	3.20	68.13	1.59
Edinburgh Napier University	26.12	5.38	2.41	67.26	1.48
University of East London	23.00	3.75	2.31	51.26	1.83
Robert Gordon University	22.75	6.50	2.53	57.88	1.78
Aberystwyth University	22.62	6.75	2.89	44.61	1.78
University of Wolverhampton	21.75	2.62	2.42	75.39	1.36
Keele University	21.25	7.12	2.08	52.07	1.99
University of Bedfordshire	20.38	6.50	2.67	53.66	1.45
London South Bank University	13.25	1.75	1.99	37.18	2.11
Staffordshire University	12.00	1.38	2.49	27.69	1.79
University of Sunderland	6.62	0.75	2.06	20.96	1.97

Table A5: SCM estimated coefficients: number of publications in journals.

Treated	Synthetic control composition
Aberystwyth University	Brandeis University (0.452), Claremont McKenna College (0.432), University of Maryland-Baltimore (0.104), Baylor University (0.013)
Aston University	Florida Atlantic University (0.791), University of Georgia (0.209)
Bangor University	Middlebury College (0.330), Claremont McKenna College (0.294), University of Maryland-Baltimore (0.231), Williams College (0.144)
Birkbeck College	Middlebury College (0.393), West Virginia University (0.269), Syracuse University (0.199), University of Rochester (0.056), University of Massachusetts-Amherst (0.046), Brandeis University (0.036)
Bournemouth University	Williams College (0.430), University of Maryland-Baltimore (0.240), Middlebury College (0.225), Appalachian State University (0.105)
Brunel University London	University of Texas-Dallas (0.777), Arizona State University (0.088), State University of New York-Buffalo (0.061), Purdue University (0.044), University of North Carolina-Greensboro (0.023), Florida Atlantic University (0.007)
Cardiff University	University of Chicago (0.274), Washington University in St. Louis (0.216), Vanderbilt University (0.197), New York University (0.158), Oklahoma State University (0.093), University of Illinois at Urbana-Champaign (0.038), University of Texas-Dallas (0.025)
City University London	University of Delaware (0.381), Florida Atlantic University (0.292), University of Georgia (0.254), Northwestern University (0.073)
Coventry University	Claremont McKenna College (0.423), Middlebury College (0.198), Fordham University (0.197), Appalachian State University (0.148), University of Maryland-Baltimore (0.024), Brandeis University (0.010)
Cranfield University	Florida Atlantic University (0.652), University of Georgia (0.288), Texas A&M University (0.040), University of Arizona (0.021)
De Montfort University	University of Nevada-Reno (0.423), Claremont McKenna College (0.352), University of North Carolina-Greensboro (0.133), College of William & Mary (0.091)
Edinburgh Napier University	Appalachian State University (0.631), University of Maryland-Baltimore (0.139), Claremont McKenna College (0.118), Baylor University (0.112)
Glasgow Caledonian University	Appalachian State University (0.562), Florida Atlantic University (0.293), Baylor University (0.116), West Virginia University (0.029)
Heriot-Watt University	Baylor University (0.410), West Virginia University (0.334), Florida Atlantic University (0.103), University of North Carolina-Greensboro (0.092), University of Alabama-Tuscaloosa (0.060)
Imperial College London	University of California-Santa Barbara (0.278), University of North Carolina-Greensboro (0.247), Stanford University (0.187), Georgia State University (0.180), University of California-Davis (0.068), University of California-Los Angeles (0.039)
Keele University	Brandeis University (0.548), University of Maryland-Baltimore (0.345), Fordham University (0.106)
King's College London	Baylor University (0.304), University of North Carolina-Greensboro (0.304), Syracuse University (0.193), Williams College (0.100), Temple University (0.070), Harvard University (0.017), Florida Atlantic University (0.011)
Kingston University	Appalachian State University (0.271), Williams College (0.270), University of Maryland-Baltimore (0.250), Florida Atlantic University (0.105), Baylor University (0.057), Oklahoma State University (0.047)
Lancaster University	University of Texas-Dallas (0.626), University of Georgia (0.274), Texas A&M University (0.066), Florida Atlantic University (0.034)
Leeds Beckett University	Brandeis University (0.558), University of Maryland-Baltimore (0.242), Baylor University (0.106), Claremont McKenna College (0.095)
London Business School	State University of New York-Buffalo (0.755), Boston College (0.152), Harvard University (0.087), University of Oklahoma (0.006)
London Metropolitan University	University of Nevada-Reno (0.590), Middlebury College (0.204), Baylor University (0.148), Brandeis University (0.053)
LSE	Harvard University (0.336), University of Georgia (0.296), University of Connecticut (0.213), MIT (0.150)
London South Bank University	Williams College (0.504), Middlebury College (0.201), Fordham University (0.198), Brandeis University (0.091), Claremont McKenna College (0.007)
Manchester Metropolitan University	Middlebury College (0.264), Boston College (0.233), Baylor University (0.170), University of North Carolina-Greensboro (0.153), Florida Atlantic University (0.100), Syracuse University (0.078)
Middlesex University	Florida Atlantic University (0.275), University of Maryland-Baltimore (0.243), University of Nevada-Reno (0.168), Claremont McKenna College (0.135), University of Alabama-Tuscaloosa (0.131), Baylor University (0.048)
Newcastle University	University of Nevada-Reno (0.499), Baylor University (0.150), Tufts University (0.085), Princeton University (0.064), Rutgers University-New Brunswick (0.064), Auburn University (0.061), Syracuse University (0.040), Harvard University (0.037)
Nottingham Trent University	University of Colorado at Denver (0.345), Claremont McKenna College (0.253), Middlebury College (0.235), University of North Carolina-Greensboro (0.089), Williams College (0.052), University of Maryland-Baltimore (0.025)
Open University	Brandeis University (0.322), University of California-Riverside (0.248), University of Oklahoma (0.184), Baylor University (0.113), University of Iowa (0.098), University of Maryland-Baltimore (0.035)
Oxford Brookes University	Middlebury College (0.483), Baylor University (0.180), Florida Atlantic University (0.144), Oklahoma State University (0.109), University of Maryland-Baltimore (0.084)
Queen Mary University of London	University of Maryland-Baltimore (0.293), Florida Atlantic University (0.282), University of Tennessee-Knoxville (0.252), University of North Carolina-Greensboro (0.089), University of Alabama-Tuscaloosa (0.070), University of Georgia (0.013)
Queen's University Belfast	University of North Carolina-Greensboro (0.588), University of California-Santa Barbara (0.227), University of Maryland-Baltimore (0.095), University of Texas-Dallas (0.043), Purdue University (0.036), Florida Atlantic University (0.010)
Robert Gordon University	Middlebury College (0.516), Claremont McKenna College (0.433), University of North Carolina-Greensboro (0.033), Williams College (0.018)
Royal Holloway, University of London	University of California-Santa Cruz (0.533), Florida Atlantic University (0.215), University of California-Santa Barbara (0.137), City University of New York (0.078), University of Texas-Dallas (0.029), Georgia State University (0.009)
Sheffield Hallam University	Brandeis University (0.437), University of Maryland-Baltimore (0.301), Baylor University (0.177), West Virginia University (0.085)
Staffordshire University	Claremont McKenna College (0.786), Williams College (0.214)
Swansea University	Fordham University (0.622), Appalachian State University (0.205), University of Texas-Dallas (0.102), University of Maryland-Baltimore (0.041), West Virginia University (0.030)
University College London	University of Chicago (0.315), Rice University (0.273), City University of New York (0.273), Fordham University (0.093), University of California-Santa Barbara (0.046)
University of Aberdeen	Brigham Young University (0.350), State University of New York-Albany (0.258), Stony Brook University (0.140), Washington University in St. Louis (0.091), Fordham University (0.070), University of Iowa (0.060), University of Minnesota (0.030)
University of Bath	Florida Atlantic University (0.488), University of Georgia (0.284), University of Alabama-Tuscaloosa (0.210), University of Michigan (0.010), West Virginia University (0.008)
University of Bedfordshire	Claremont McKenna College (0.702), Middlebury College (0.298)
University of Birmingham	Stanford Uni (0.237), Rensselaer Polytechnic Institute (0.196), Uni of California-Santa Cruz (0.186), Uni of Colorado at Denver (0.180), Uni of Rochester (0.116), Georgia State Uni (0.050), Temple Uni (0.034)

Treated	Synthetic control composition
University of Bradford	University of Tennessee-Knoxville (0.473), Fordham University (0.204), Baylor University (0.152), University of Maryland-Baltimore (0.086), Claremont McKenna College (0.085)
University of Brighton	Williams College (0.596), University of Maryland-Baltimore (0.388), Florida Atlantic University (0.016)
University of Bristol	West Virginia University (0.393), University of Delaware (0.231), Brandeis University (0.170), Iowa State University (0.123), Syracuse University (0.053), Boston College (0.031)
University of Cambridge	University of California-Santa Barbara (0.433), Harvard University (0.261), Rensselaer Polytechnic Institute (0.150), MIT (0.069), Georgia State University (0.045), University of California-Los Angeles (0.042)
University of Central Lancashire	Claremont McKenna College (0.521), University of Maryland-Baltimore (0.433), Appalachian State University (0.045)
University of Dundee	West Virginia University (0.493), Middlebury College (0.353), University of Maryland-Baltimore (0.152)
University of Durham	University of Tennessee-Knoxville (0.533), University of Alabama-Tuscaloosa (0.237), Fordham University (0.117), Baylor University (0.059), University of Maryland-Baltimore (0.054)
University of East Anglia	Appalachian State University (0.337), Syracuse University (0.255), University of North Carolina-Greensboro (0.155), Oklahoma State University (0.132), University of Texas-Dallas (0.061), Iowa State University (0.046), University of Rochester (0.013)
University of East London	Claremont McKenna College (0.674), Middlebury College (0.315), University of Maryland-Baltimore (0.011)
University of Edinburgh	Claremont McKenna College (0.323), University of Texas-Dallas (0.294), Georgia State University (0.225), University of California-Santa Barbara (0.108), MIT (0.050)
University of Essex	University of Maryland-Baltimore (0.288), Georgia Institute of Technology (0.181), Oklahoma State University (0.164), University of Wyoming (0.158), University of Rochester (0.077), University of California-Santa Barbara (0.070), Iowa State University (0.056), University of Massachusetts-Amherst (0.006)
University of Exeter	University of Maryland-Baltimore (0.263), University of Delaware (0.250), University of California-Riverside (0.163), Arizona State University (0.157), Baylor University (0.115), University of Iowa (0.039), North Carolina State University (0.014)
University of Glasgow	University of Maryland-Baltimore (0.322), University of Massachusetts-Amherst (0.257), Iowa State University (0.169), University of North Carolina-Greensboro (0.100), George Washington University (0.086), Oklahoma State University (0.038), University of Tennessee-Knoxville (0.025)
University of Greenwich	Claremont McKenna College (0.747), University of North Carolina-Greensboro (0.126), Appalachian State University (0.069), University of Maryland-Baltimore (0.059)
University of Hertfordshire	Claremont McKenna College (0.737), University of Maryland-Baltimore (0.232), Baylor University (0.017), Fordham University (0.014)
University of Hull	University of Maryland-Baltimore (0.428), Oklahoma State University (0.375), Florida Atlantic University (0.193)
University of Kent	University of California-Santa Cruz (0.483), Florida Atlantic University (0.214), University of California-Santa Barbara (0.148), University of Texas-Dallas (0.134), University of Maryland-Baltimore (0.021)
University of Leeds	University of North Carolina-Greensboro (0.290), University of Georgia (0.264), University of Florida (0.226), University of Texas-Dallas (0.210), Arizona State University (0.010)
University of Leicester	University of North Carolina-Greensboro (0.364), Williams College (0.225), Dartmouth College (0.195), Boston College (0.095), Harvard University (0.066), Rensselaer Polytechnic Institute (0.055)
University of Liverpool	University of California-Santa Cruz (0.405), College of William & Mary (0.290), University of Texas-Dallas (0.171), Claremont McKenna College (0.068), City University of New York (0.066)
University of Manchester	Pennsylvania State University (0.565), Texas A&M University (0.205), Purdue University (0.165), Northwestern University (0.064)
University of Northumbria at Newcastle	Middlebury College (0.691), Baylor University (0.174), Brandeis University (0.135)
University of Nottingham	Texas A&M University (0.739), Florida Atlantic University (0.124), Syracuse University (0.066), Columbia University (0.048), City University of New York (0.023)
University of Oxford	Harvard University (0.477), University of Chicago (0.231), Georgia State University (0.141), City University of New York (0.119), Stanford University (0.032)
University of Plymouth	University of Maryland-Baltimore (0.460), University of Nevada-Reno (0.377), Claremont McKenna College (0.074), University of Alabama-Tuscaloosa (0.061), Baylor University (0.028)
University of Portsmouth	University of North Carolina-Greensboro (0.586), Florida Atlantic University (0.193), University of California-Santa Barbara (0.110), Stony Brook University (0.100), University of Maryland-Baltimore (0.011)
University of Reading	Florida Atlantic University (0.298), Oklahoma State University (0.286), University of Florida (0.225), Syracuse University (0.190)
University of Salford	University of Maryland-Baltimore (0.552), Stony Brook University - SUNY (0.181), University of Chicago (0.102), Oklahoma State University (0.088), University of Texas-Dallas (0.077)
University of Sheffield	Oklahoma State University (0.547), University of Georgia (0.453)
University of Southampton	University of Maryland-Baltimore (0.625), Iowa State University (0.208), University of California-Berkeley (0.157), University of Illinois at Urbana-Champaign (0.011)
University of South Wales	Claremont McKenna College (0.289), University of Maryland-Baltimore (0.243), Stony Brook University (0.184), Brandeis University (0.177), Middlebury College (0.078), Fordham University (0.029)
University of St Andrews	University of Maryland-Baltimore (0.521), Colorado State University (0.135), Baylor University (0.125), California Institute of Technology (0.115), University of Texas-Dallas (0.074), Stony Brook University (0.030)
University of Stirling	University of California-Santa Barbara (0.391), University of North Carolina-Greensboro (0.361), Florida Atlantic University (0.229), City University of New York (0.009), University of Texas-Dallas (0.007)
University of Strathclyde	University of California-Santa Barbara (0.399), University of Virginia (0.357), University of California-Los Angeles (0.160), University of Minnesota (0.030), University of Pittsburgh (0.023), Stanford University (0.019), University of Illinois at Urbana-Champaign (0.011)
University of Sunderland	Claremont McKenna College (0.834), Middlebury College (0.166)
University of Surrey	Temple University (0.485), Syracuse University (0.245), West Virginia University (0.189), University of North Carolina-Greensboro (0.080)
University of Sussex	Fordham University (0.555), University of Texas-Dallas (0.259), University of Maryland-Baltimore (0.098), Iowa State University (0.049), Appalachian State University (0.026), University of Rochester (0.016)
University of the West of England, Bristol	Appalachian State University (0.311), University of Maryland-Baltimore (0.308), Florida Atlantic University (0.216), University of California-Santa Barbara (0.097), Oklahoma State University (0.069)
University of Ulster	Virginia Commonwealth University (0.434), Middlebury College (0.278), Boston College (0.144), Baylor University (0.054), Harvard University (0.033), Syracuse University (0.032), Florida Atlantic University (0.025)
University of Warwick	Pennsylvania State University (0.297), Yale University (0.257), Purdue University (0.231), University of Georgia (0.120), Florida State University (0.058), University of Chicago (0.037)
University of Westminster	University of Maryland-Baltimore (0.426), Middlebury College (0.248), Appalachian State University (0.183), University of North Carolina-Greensboro (0.105), Florida Atlantic University (0.038)
University of Wolverhampton	Middlebury College (0.535), Appalachian State University (0.304), University of Maryland-Baltimore (0.081), Williams College (0.080)
University of York	Dartmouth College (0.643), Princeton University (0.287), Boston College (0.059), University of North Carolina-Greensboro (0.011)

Table A6: SCM estimated coefficients: number of papers published in a 3*, 4*, 4** journal

Treated	Synthetic control composition
Aberystwyth University	Claremont McKenna College(0.512), Middlebury College(0.257), Brandeis University(0.153), Appalachian State University(0.046), Baylor University(0.032)
Aston University	University of Missouri(0.592), Baylor University(0.347), Texas A&M University(0.06)
Bangor University	Middlebury College(0.497), University of Maryland-Baltimore County(0.428), University of North Carolina-Greensboro(0.074)
Birkbeck College	College of William & Mary(0.491), University of Nevada-Reno(0.317), University of Kentucky(0.144), Oklahoma State University(0.024), Arizona State University(0.022)
Bournemouth University	Middlebury College(0.793), Fordham University(0.104), West Virginia University(0.103)
Brunel University London	University of Hawaii-Manoa(0.34), University of Texas-Dallas(0.27), University of Nevada-Reno(0.225), State University of New York-Buffalo (SUNY)(0.059), University of Washington(0.051), Temple University(0.044), University of Arizona(0.012)
Cardiff University	Rice University(0.279), University of Pennsylvania(0.227), Baylor University(0.19), City University of New York (CUNY)(0.156), Michigan State University(0.087), University of Delaware(0.04), New York University (NYU)(0.021)
City University London	University of Georgia(0.543), Baylor University(0.274), Temple University(0.066), University of Arizona(0.064), University of Michigan(0.03), University of Washington(0.024)
Coventry University	Middlebury College(0.574), University of Maryland-Baltimore County(0.211), University of Colorado at Denver(0.137), Claremont McKenna College(0.055), University of Hawaii-Manoa(0.023)
Cranfield University	State University of New York-Buffalo (SUNY)(0.327), Florida State University(0.253), Temple University(0.173), University of Texas-Dallas(0.137), Florida Atlantic University(0.097), Arizona State University(0.012)
De Montfort University	Tufts University(0.358), University of Nevada-Reno(0.322), University of Hawaii-Manoa(0.215), Fordham University(0.094), University of Wisconsin-Madison(0.011)
Edinburgh Napier University	Middlebury College(0.730), Appalachian State University(0.270)
Glasgow Caledonian University	Williams College(0.528), Fordham University(0.315), Baylor University(0.157)
Heriot-Watt University	Middlebury College(0.364), West Virginia University(0.34), Baylor University(0.246), University of Hawaii-Manoa(0.038), Brandeis University(0.012)
Imperial College London	Emory University(0.494), University of Washington(0.235), City University of New York (CUNY)(0.177), Fordham University(0.074), Georgia State University(0.02)
Keele University	West Virginia University(0.408), Middlebury College(0.287), Fordham University(0.245), University of Delaware(0.038), University of Colorado at Denver(0.023)
King's College London	University of Hawaii-Manoa(0.41), Brandeis University(0.269), Arizona State University(0.14), Syracuse University(0.098), Baylor University(0.082)
Kingston University	Middlebury College(0.324), Williams College(0.316), Florida Atlantic University(0.187), Claremont McKenna College(0.173)
Lancaster University	University of Texas-Dallas(0.431), Baylor University(0.357), University of Michigan(0.17), University of Washington(0.042)
Leeds Beckett University	Middlebury College(0.667), Williams College(0.300), Fordham University(0.029), West Virginia University(0.004)
London Business School	State University of New York-Buffalo (SUNY)(0.489), Florida State University(0.3), Pennsylvania State University(0.085), Dartmouth College(0.08), Stanford University(0.043)
London Metropolitan University	Middlebury College(0.672), Baylor University(0.186), Fordham University(0.07), West Virginia University(0.045), Syracuse University(0.026)
LSE	University of Michigan(0.413), University of Minnesota(0.253), University of California-Riverside(0.214), Harvard University(0.063), Duke University(0.036), Baylor University(0.021)
London South Bank University	Middlebury College(0.704), Tufts University(0.173), Santa Clara University(0.095), Brandeis University(0.028)
Manchester Metropolitan University	University of Nevada-Reno(0.443), Brandeis University(0.262), Williams College(0.236), University of Massachusetts-Amherst(0.039), Dartmouth College(0.02)
Middlesex University	Middlebury College(0.396), Baylor University(0.289), Claremont McKenna College(0.147), University of Hawaii-Manoa(0.14), West Virginia University(0.017), Williams College(0.01)
Newcastle University	Brandeis University(0.714), University of Alabama-Tuscaloosa(0.152), University of Michigan(0.044), Baylor University(0.042), University of Pennsylvania(0.035), University of California-Riverside(0.013))
Nottingham Trent University	Middlebury College(0.679), Tufts University(0.179), University of Colorado at Denver(0.072), Santa Clara University(0.054), University of Delaware(0.015)
Open University	Middlebury College(0.317), University of Maryland-Baltimore County(0.306), University of Wyoming(0.222), Colorado State University(0.113), Williams College(0.041)
Oxford Brookes University	Middlebury College(0.629), University of Maryland-Baltimore County(0.213), University of Hawaii-Manoa(0.093), Brandeis University(0.039), State University of New York-Albany (SUNY)(0.018), Baylor University(0.009)
Queen Mary University of London	Florida Atlantic University(0.904), University of Texas-Dallas(0.067), Temple University(0.029)
Queen's University Belfast	University of North Carolina-Greensboro(0.440), University of California-Santa Cruz (UCSC)(0.213), Florida Atlantic University(0.178), University of Maryland-Baltimore County(0.102), State University of New York-Buffalo (SUNY)(0.067)
Robert Gordon University	Middlebury College(0.661), University of Nevada-Reno(0.3), Baylor University(0.035)
Royal Holloway, University of London	Claremont McKenna College(0.551), University of California-Santa Cruz (UCSC)(0.245), University of Washington(0.135), Florida State University(0.039), Florida Atlantic University(0.027)
Sheffield Hallam University	Middlebury College(0.569), Santa Clara University(0.17), Williams College(0.103), University of North Carolina-Greensboro(0.075), California Institute of Technology(0.061), Rice University(0.022)
Staffordshire University	Middlebury College(0.786), Williams College(0.214)
Swansea University	Fordham University(0.412), University of Maryland-Baltimore County(0.256), Oklahoma State University(0.182), College of William & Mary(0.111), Claremont McKenna College(0.039)
University College London	University of Virginia(0.349), Rice University(0.202), Michigan State University(0.201), College of William & Mary(0.147), Georgia State University(0.102)
University of Aberdeen	University of Delaware(0.358), Brandeis University(0.318), West Virginia University(0.172), University of Minnesota(0.105), Baylor University(0.047)
University of Bath	University of Georgia(0.464), Baylor University(0.309), University of Texas-Dallas(0.088), University of Washington(0.082), Temple University(0.038), University of Hawaii-Manoa(0.019)
University of Bedfordshire	Middlebury College(0.400), Baylor University(0.389), Claremont McKenna College(0.211)
University of Birmingham	Emory University(0.315), Tulane University(0.303), Dartmouth College(0.264), University of Notre Dame(0.114)
University of Bradford	Claremont McKenna College(0.4), Florida Atlantic University(0.324), University of Hawaii-Manoa(0.194), University of Texas-Dallas(0.062), Baylor University(0.019)

Treated	Synthetic control composition
University of Brighton	Williams College(0.641), Middlebury College(0.228), Baylor University(0.063), Brandeis University(0.054), Claremont McKenna College(0.014)
University of Bristol	College of William & Mary(0.415), University of Delaware(0.354), Syracuse University(0.135), Michigan State University(0.069), Oklahoma State University(0.028)
University of Cambridge	University of Southern California(0.468), University of Michigan(0.163), University of Texas-Dallas(0.137), City University of New York (CUNY)(0.124), Baylor University(0.059), Temple University(0.05)
University of Central Lancashire	Middlebury College(0.609), Claremont McKenna College(0.226), University of Maryland-Baltimore County(0.165)
University of Dundee	Brandeis University(0.471), Baylor University(0.301), Oklahoma State University(0.201), Syracuse University(0.014), Arizona State University(0.011)
University of Durham	Baylor University(0.701), University of Texas-Dallas(0.16), Fordham University(0.086), University of Florida(0.053)
University of East Anglia	University of Nevada-Reno(0.412), University of Kentucky(0.175), State University of New York-Buffalo (SUNY)(0.137), Syracuse University(0.106), University of Miami(0.097), Georgetown University(0.041), Arizona State University(0.033)
University of East London	Middlebury College (0.621), West Virginia University (0.379)
University of Edinburgh	Temple University(0.393), College of William & Mary(0.195), University of Nevada-Reno(0.149), University of Miami(0.114), University of Wisconsin-Madison(0.08), University of Arizona(0.069)
University of Essex	Washington University in St. Louis(0.437), University of Texas-Dallas(0.205), Baylor University(0.169), University of Hawaii-Manoa(0.127), Rutgers University-New Brunswick(0.062)
University of Exeter	University of California-Riverside(0.636), Arizona State University(0.114), University of Hawaii-Manoa(0.09), Florida State University(0.079), Dartmouth College(0.042), Baylor University(0.039)
University of Glasgow	University of Miami(0.403), University of California-Santa Cruz (UCSC)(0.197), State University of New York-Buffalo (SUNY)(0.18), University of North Carolina-Greensboro(0.166), University of Maryland-Baltimore County(0.037), University of Nevada-Reno(0.017)
University of Greenwich	Middlebury College(0.599), Williams College(0.166), University of Hawaii-Manoa(0.141), Appalachian State University(0.081), Brandeis University(0.014)
University of Hertfordshire	Claremont McKenna College(0.747), West Virginia University(0.138), Middlebury College(0.112), Baylor University(0.004)
University of Hull	Claremont McKenna College(0.429), University of Maryland-Baltimore County(0.344), Baylor University(0.221), West Virginia University(0.007)
University of Kent	University of Maryland-Baltimore County(0.58), Claremont McKenna College(0.197), Clemson University(0.091), University of Wisconsin-Madison(0.047), University of Washington(0.033), University of Texas-Dallas(0.027), University of Hawaii-Manoa(0.026)
University of Leeds	University of Texas-Dallas(0.371), Baylor University(0.222), Arizona State University(0.192), Oklahoma State University(0.112), University of Florida(0.104)
University of Leicester	University of Nevada-Reno(0.32), State University of New York-Binghamton (SUNY)(0.219), University of Notre Dame(0.205), Temple University(0.136), University of Missouri(0.055), University of Rochester(0.054), Syracuse University(0.01)
University of Liverpool	Fordham University(0.633), University of Notre Dame(0.202), Appalachian State University(0.057), State University of New York-Binghamton (SUNY)(0.053), University of Missouri(0.04), Dartmouth College(0.015)
University of Manchester	Texas A&M University(0.416), Northwestern University(0.347), Stanford University(0.07), University of Washington(0.068), Pennsylvania State University(0.052), Harvard University(0.047)
University of Northumbria at Newcastle	Middlebury College(0.692), Fordham University(0.175), Brandeis University(0.099), Tufts University(0.034)
University of Nottingham	Duke University(0.527), University of Southern California(0.271), Northwestern University(0.072), University of Maryland(0.059), Michigan State University(0.037), Harvard University(0.033)
University of Oxford	University of Arizona(0.737), University of California-Berkeley(0.097), Northwestern University(0.052), University of Wisconsin-Madison(0.045), Arizona State University(0.035), Purdue University(0.034)
University of Plymouth	Middlebury College(0.510), University of Colorado at Denver(0.317), University of Maryland-Baltimore County(0.137), Fordham University(0.035)
University of Portsmouth	University of California-Santa Cruz (UCSC)(0.416), Middlebury College(0.307), State University of New York-Buffalo (SUNY)(0.199), Williams College(0.075)
University of Reading	Oklahoma State University(0.397), Temple University(0.218), University of Texas-Dallas(0.16), DePaul University(0.102), Florida State University(0.05), University of California-Santa Barbara (UCSB)(0.042), Arizona State University(0.029)
University of Salford	University of Maryland-Baltimore County(0.652), College of William & Mary(0.148), University of Colorado at Denver(0.096), University of Miami(0.046), Santa Clara University(0.029), Middlebury College(0.029)
University of Sheffield	Temple University(0.384), University of Maryland-Baltimore County(0.353), University of Texas-Dallas(0.16), Arizona State University(0.089), University of Hawaii-Manoa(0.014)
University of Southampton	University of Miami(0.283), State University of New York-Buffalo (SUNY)(0.277), University of Wisconsin-Madison(0.145), University of Nevada-Reno(0.139), Arizona State University(0.135), University of Texas-Dallas(0.021)
University of South Wales	Middlebury College(0.705), Fordham University(0.19), West Virginia University(0.102)
University of St Andrews	University of Maryland-Baltimore County(0.732), University of Texas-Dallas(0.179), University of Hawaii-Manoa(0.046), Claremont McKenna College(0.043)
University of Stirling	University of California-Santa Cruz (UCSC)(0.318), University of Maryland-Baltimore County(0.212), University of Hawaii-Manoa(0.13), University of Nevada-Reno(0.097), Tufts University(0.092), North Carolina State University(0.056), Iowa State University(0.045), University of Colorado at Denver(0.026), University of Wisconsin-Madison(0.025)
University of Strathclyde	Florida State University(0.384), University of Wyoming(0.34), University of Virginia(0.118), University of California-Los Angeles (UCLA)(0.112), University of California-Santa Cruz (UCSC)(0.039), University of Illinois at Urbana-Champaign(0.007)
University of Sunderland	West Virginia University(0.625), Middlebury College(0.375)
University of Surrey	Baylor University(0.378), Fordham University(0.318), City University of New York (CUNY)(0.118), Emory University(0.113), University of Arizona(0.072)
University of Sussex	College of William & Mary(0.448), University of Hawaii-Manoa(0.415), University of Arizona(0.083), University of Texas-Dallas(0.041), Arizona State University(0.008), University of Miami(0.006)
University of the West of England, Bristol	Williams College(0.365), Brandeis University(0.169), University of Hawaii-Manoa(0.164), DePaul University(0.137), University of Nevada-Reno(0.115), Middlebury College(0.036), State University of New York-Buffalo (SUNY)(0.014)
University of Ulster	Fordham University(0.36), Virginia Commonwealth University(0.229), Baylor University(0.167), West Virginia University(0.132), California Institute of Technology(0.08), Williams College(0.032)
University of Warwick	MIT(0.315), Florida State University(0.189), University of Illinois at Urbana-Champaign(0.179), University of Washington(0.147), University of Wisconsin-Madison(0.094), University of Central Florida(0.076)
University of Westminster	Middlebury College(0.740), Williams College(0.202), State University of New York-Buffalo (SUNY)(0.058)
University of Wolverhampton	Middlebury College(0.636), Williams College(0.364)
University of York	George Washington University(0.307) , University of Delaware(0.215), Baylor University(0.186), University of Hawaii-Manoa(0.107), Syracuse University(0.086), University of Michigan(0.061), Brandeis University(0.039)

Table A7: Yearly and overall effects considering the outcomes and some extensions

Outcomes & Extensions	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	ATT_{-2015}	ATT
number of publications in journals	11.62****	0.33	-3.52	8.34	22.47****	36.01****	31.11***	44.35****	106.38***	150.74****
number of publications in journals 3*, 4*, 4**	-4.80**	0.67	-6.53	6.22***	-4.12	9.69**	23.13****	25.11****	24.26**	49.38***
number of publications in Economics/Econometrics journals graded as 3*, 4*, 4** stars	1.23	5.53	4.08	0.19	-6.46	0.13	0.03	0.54	4.74	5.28
number of publications in Finance/Management journals graded as 3*, 4*, 4** stars	-4.12	-4.36**	0.23	-8.22	5.39	3.15	4.52*	13.63	-3.40	10.23
number of publications in journals per author	-0.049	-0.098***	-0.078***	-0.077***	-0.105****	-0.049	-0.078**	-0.065*	-0.53****	-0.60****
number of publications in journals 3*, 4*, 4** per author	-0.078**	0.004	-0.005	-0.069	-0.010	-0.002	-0.024	0.111*	-0.185	-0.074
number of publications in Economics/Econometrics journals graded as 3*, 4*, 4** stars per author	0.010	0.011	0.025**	0.007	0.021	0.025	0.039	0.002	0.140	0.143
number of publications in Finance/Management journals graded as 3*, 4*, 4** stars per author	-0.074***	-0.051	0.019	-0.039*	0.021	0.049**	0.082****	0.156****	0.008	0.164***
proportion of publications in Economics/Econometrics journals	-0.076	0.037	-0.029	0.013	0.016	-0.023	-0.037*	-0.024*	-0.126*	-0.150**
proportion of publications in Finance/Management journals	0.074**	0.023	0.022	0.019	-0.089	0.024	0.008	0.029	0.082	0.112
proportion of publications in journals graded as 3*, 4*, 4** stars	-0.048	-0.051	0.031	-0.078	0.065***	0.056	0.031	0.080	0.006	0.086
proportion of publications in Economics/Econometrics journals graded as 3*, 4*, 4** stars	0.004	-0.012	0.005	-0.002	0.047	0.031	0.026	-0.006	0.090	0.084
proportion of publications in Finance/Management journals graded as 3*, 4*, 4** stars	0.072	-0.045*	-0.071****	-0.048	-0.028	0.017	0.028	0.064	-0.070	-0.009

Values are marked by *, **, ***, **** if they are significant at a level of 0.10, 0.05, 0.01 or 0.001, respectively.

Table A8: Averaged yearly and overall ATTs considering the outcomes and some extensions for the *Russell group*.

Russell group: Outcomes & Extensions	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	ATT_{-2015}	ATT
number of publications in journals	6.34	5.81	-5.07	-3.01	14.17*	28.38*	33.31**	41.32***	11.42*	15.16**
number of publications in journals 3*, 4*, 4**	-1.06	2.37	-2.35	-6.22	12.06*	19.62*	20.08*	24.43**	6.35	8.61*
number of publications in Economics/Econometrics journals graded as 3*, 4*, 4** stars	0.55	2.16	-0.03	-2.98	3.98	4.58	4.61	3.27	1.84	2.02
number of publications in Finance/Management in top journals	-1.57	-0.56	-1.33	-3.70	7.94	14.00*	14.76*	20.98**	4.22	6.31
number of publications in journals per author	-0.04	-0.01	-0.03	-0.07	-0.03	0.01	0.01	0.02	-0.02	-0.02
number of publications in top journals per author	-0.06*	-0.01	-0.02	-0.06*	0.02	0.03*	0.007	0.07**	-0.04	-0.01
number of publications in Economics/Econometrics in top journals per author	-0.01	-0.006	0.009	-0.01	0.008	0.001	0.001	-0.005	-0.002	-0.003
number of publications in Finance/Management in top journals per author	-0.039**	-0.014	-0.024	-0.020	0.004	0.038*	0.032	0.085***	-0.003	0.007
proportion of publications in Economics/Econometrics journals	-0.044	0.007	-0.010	-0.028	0.006	-0.043	-0.012	-0.032	-0.017	-0.019
proportion of publications in Finance/Management journals	0.027	0.001	0.002	0.024	-0.017	0.039	0.011	0.038	0.012	0.015
proportion of publications in top journals	-0.056	-0.011	-0.034	-0.040	0.016	0.011	0.008	0.041	-0.015	-0.008
proportion of publications in Economics/Econometrics in top journals	-0.014	0.013	0.015	-0.007	0.026	-0.001	-0.001	-0.005	0.004	0.003
proportion of publications in Finance/Management in top journals	-0.017	-0.009	-0.030	-0.006	-0.001	0.043	0.019	0.047	-0.001	0.005

Values are marked by *, **, ***, **** if they are significant at a level of, respectively, 0.10, 0.05, 0.01 or 0.001.

Table A9: Averaged yearly and overall ATTs considering the outcomes and some extensions for the *Non-Russell group*

Non-Russell group: Outcomes & Extensions	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	ATT_{-2015}	ATT
number of publications in journals	-1.61	0.39	-2.46	-1.80	0.09	5.79	7.58	13.78*	1.14	2.72
number of publications in journals 3*, 4*, 4**	-3.34	-2.44	-4.65	-5.44	3.28	3.25	6.31	8.74*	-0.43	0.71
number of publications in Economics/Econometrics journals graded as 3, 4, 4* stars	0.16	-4.26	-1.62	-2.78	-1.75	-2.35	-3.12	-4.08	-2.24	-2.47
number of publications in Finance/Management journals graded as 3, 4, 4* stars	-2.29	-1.96	-3.32	-1.66	1.81	3.14	4.39	7.00*	0.01	0.88
number of publications in journals per author	-0.09**	-0.06	-0.07*	-0.08*	-0.07*	-0.06*	-0.03	-0.03	-0.07*	-0.06*
number of publications in journals 3*, 4*, 4** per author	-0.13***	-0.10**	-0.11**	-0.12*	-0.03	-0.02	-0.03	-0.01	-0.08*	-0.07**
number of publications in Economics/Econometrics journals graded as 3, 4, 4* stars per author	-0.008	0.030	-0.014	-0.023	-0.011	-0.021	-0.030	-0.043	-0.020	-0.023
number of publications in Finance/Management journals graded as 3, 4, 4* stars per author	-0.043**	-0.031*	-0.048**	-0.051**	-0.006	0.006	-0.001	0.046**	-0.025	-0.016
proportion of publications in Economics/Econometrics journals	-0.082*	-0.074*	-0.071*	-0.081**	-0.027	-0.076	-0.080	-0.092**	-0.070*	-0.073*
proportion of publications in Finance/Management journals	0.080**	0.073**	0.073**	0.080**	0.026	0.074**	0.078**	0.091**	0.069*	0.072*
proportion of publications in journals graded as 3*, 4*, 4** stars	-0.136*	-0.162**	-0.129*	-0.152**	-0.002	-0.041	-0.042	-0.029	-0.095	-0.086
proportion of publications in Economics/Econometrics journals graded as 3*, 4*, 4** stars	-0.011	-0.047	-0.023	-0.031	-0.015	-0.017	-0.033	-0.041	-0.025	-0.027
proportion of publications in Finance/Management journals graded as 3*, 4*, 4** stars	-0.023	-0.057*	-0.076***	-0.047*	-0.012	0.014	-0.018	0.001	-0.031	-0.026

Values are marked by *, **, ***, **** if they are significant at a level of, respectively, 0.10, 0.05, 0.01 or 0.001.

Table A10: Averaged yearly and overall ATTs considering the outcomes and some extensions for the *Remainers*

Remainers: Outcomes & Extensions	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	ATT_{-2015}	ATT
number of publications in journals	7.64*	9.14**	-3.54	2.57	12.40***	24.35****	27.81****	37.21****	11.48**	14.69***
number of publications in journals 3*, 4*, 4**	-0.34	5.88*	-0.26	-4.47	12.48****	17.59****	21.37****	24.49****	7.46***	9.59***
number of publications in Economics/Econometrics journals graded as 3, 4, 4* stars	1.13	2.83	-0.44	-3.14	4.39**	4.98**	5.76**	3.93	2.22	2.43
number of publications in Finance/Management journals graded as 3, 4, 4* stars	-2.18	1.30	1.10	-2.41	7.46****	10.12***	13.20****	19.88****	4.07**	4.08**
number of publications in journals per author	-0.034*	0.025	-0.022	-0.048***	-0.031	0.008	0.024	0.031	-0.011	-0.005
number of publications in journals 3*, 4*, 4** per author	-0.053**	0.008	-0.018	-0.056***	0.023	0.027	0.030	0.085****	-0.005	0.005
number of publications in Economics/Econometrics journals graded as 3, 4, 4* stars per author	-0.001	0.000	-0.002	-0.022	0.014	-0.001	0.001	-0.007	-0.002	-0.003
number of publications in Finance/Management journals graded as 3, 4, 4* stars per author	-0.035***	0.006	-0.003	-0.018	0.008	0.024	0.035**	0.097****	0.002	0.003
proportion of publications in Economics/Econometrics journals	-0.041*	-0.006	-0.002	-0.031	0.021*	-0.047***	-0.013	-0.043**	-0.017	-0.020
proportion of publications in Finance/Management journals	0.024	0.007	-0.001	0.026	-0.033	0.039**	0.009	0.046**	0.009	0.010
proportion of publications in journals graded as 3*, 4*, 4** stars	-0.031	0.022	-0.021	-0.033	0.028	0.001	0.028	0.053***	0.000	0.006
proportion of publications in Economics/Econometrics journals graded as 3*, 4*, 4** stars	-0.013	0.015	-0.001	-0.011	0.034*	0.001	0.004	-0.009	0.004	0.002
proportion of publications in Finance/Management journals graded as 3*, 4*, 4** stars	-0.014	0.001	-0.015	-0.019	-0.011	0.025	0.017	0.051***	-0.002	-0.002

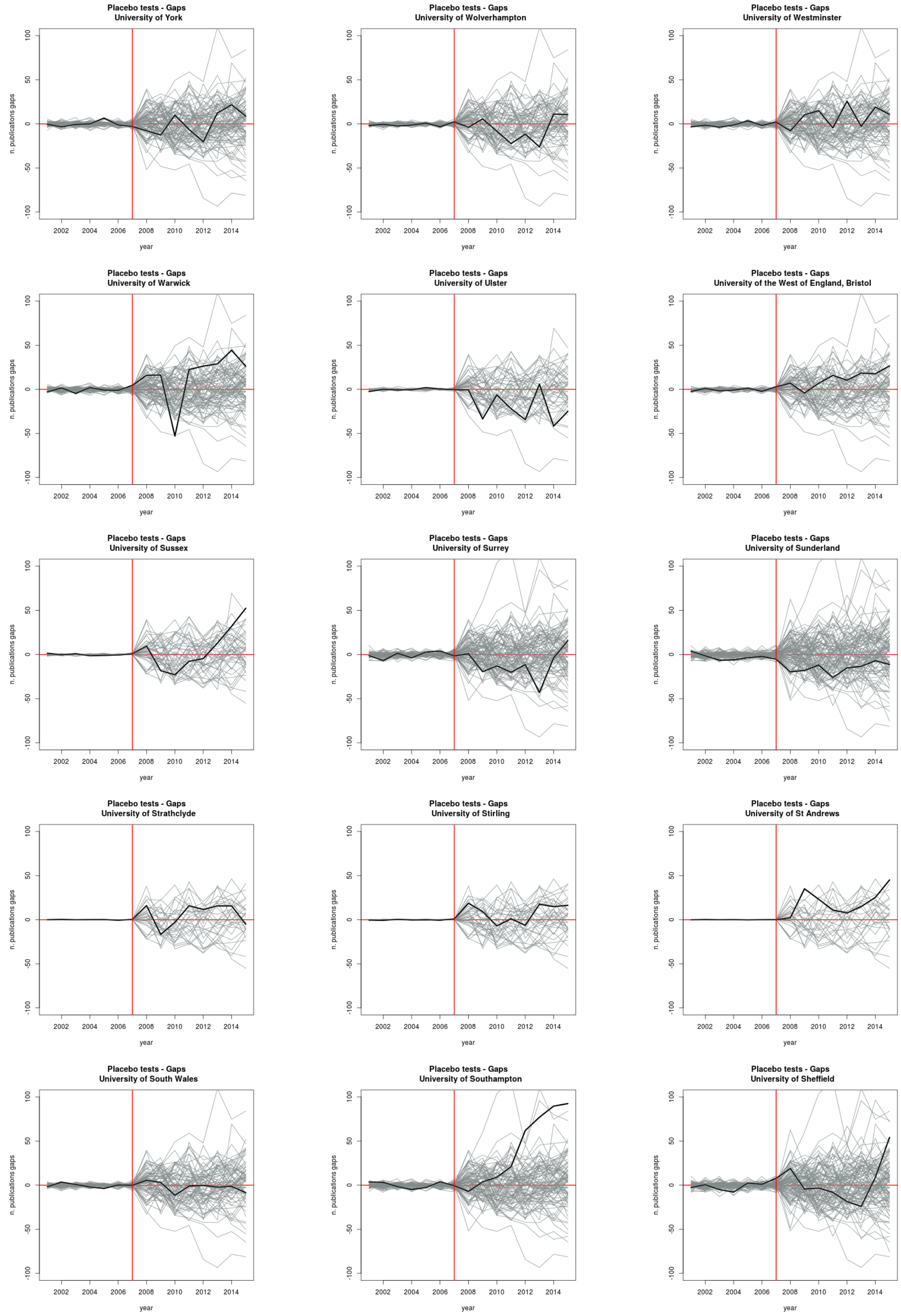
Values are marked by *, **, ***, **** if they are significant at a level of, respectively, 0.10, 0.05, 0.01 or 0.001.

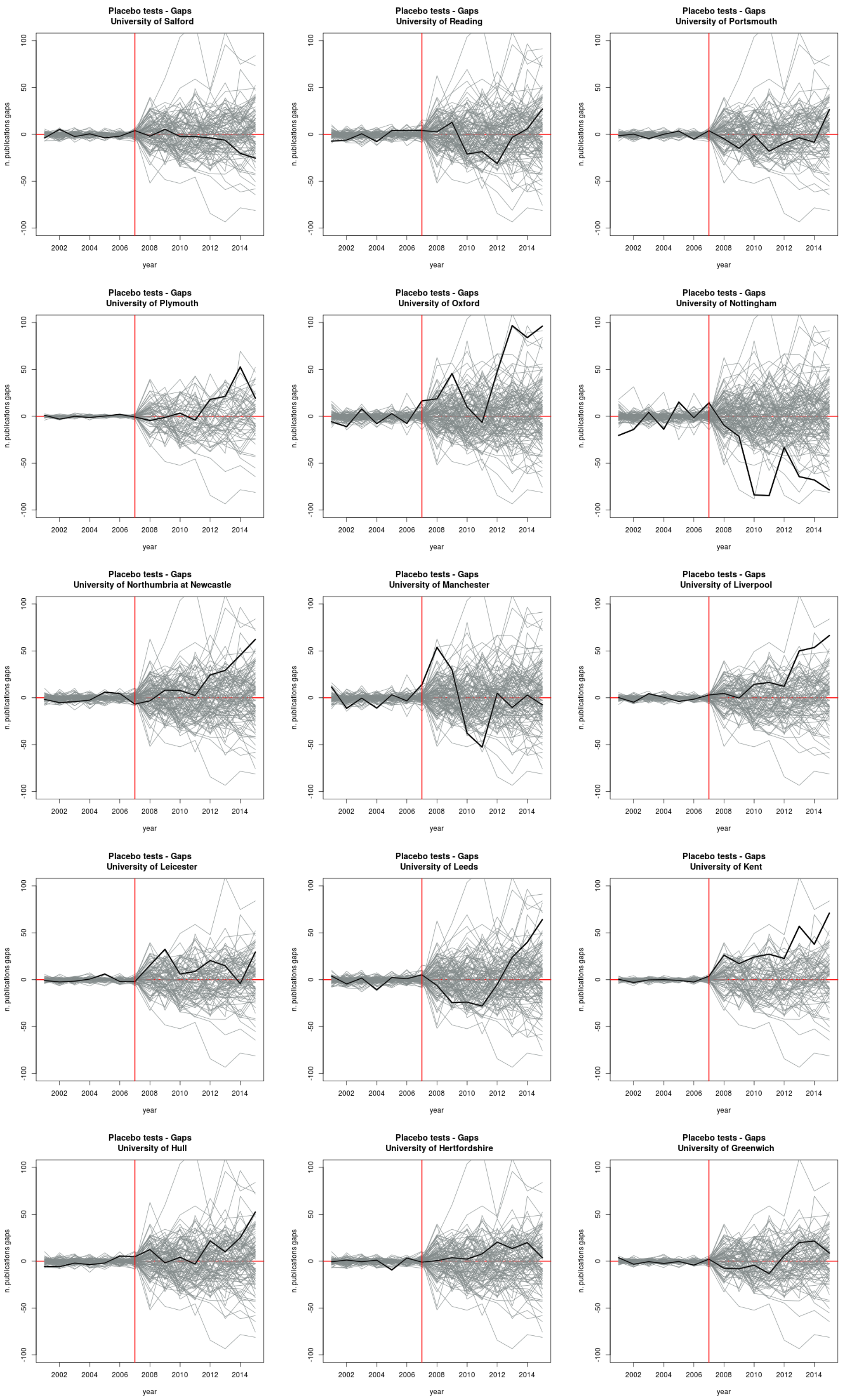
Table A11: Averaged yearly and overall ATTs considering the outcomes and some extensions for the *Leavers*

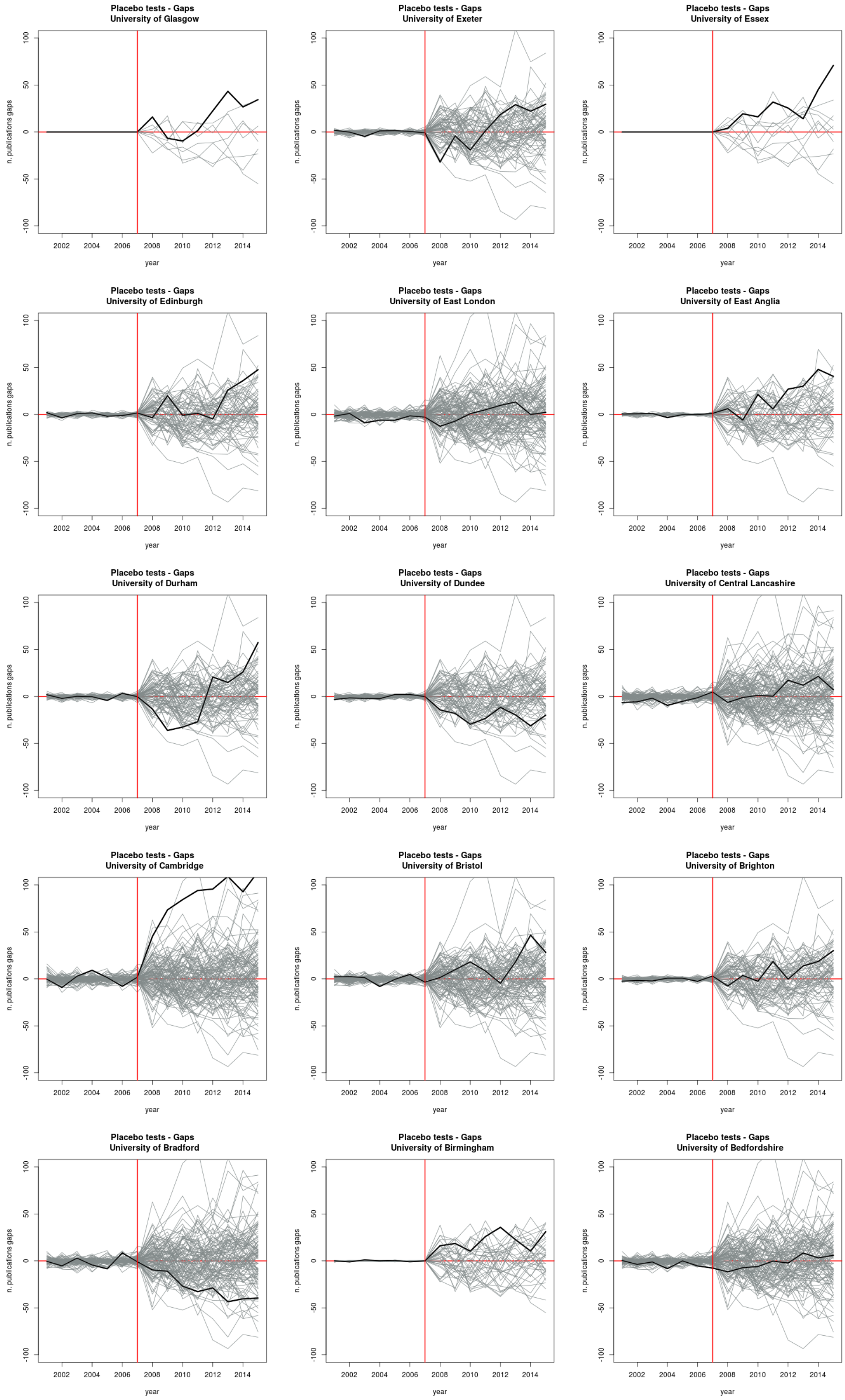
Non-Remainers group: Outcomes & Extensions	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	ATT_{-2015}	ATT
number of publications in journals	-2.81	-1.62	-3.03	-4.46	-0.02	6.19	8.48*	13.87***	0.38	2.07
number of publications in journals 3*, 4*, 4**	-3.86*	-4.51*	-5.84****	-6.24****	2.46	3.10	4.71**	7.61****	-1.45	-0.32
number of publications in Economics/Econometrics journals graded as 3, 4, 4* stars	-0.14	-5.04****	-1.53	-2.69*	-2.35*	-3.03**	-4.23****	-4.92***	-2.71**	-2.99**
number of publications in Finance/Management journals graded as 3, 4, 4* stars	-2.04*	-2.98	-4.65***	-2.15	1.61	4.28**	4.43***	6.56****	-0.22	-0.21
number of publications in journals per author	-0.100****	-0.081****	-0.079****	-0.101****	-0.078****	-0.066****	-0.044***	-0.034**	-0.078***	-0.073***
number of publications in journals 3*, 4*, 4** per author	-0.142****	-0.125****	-0.126****	-0.133****	-0.033**	-0.034**	-0.043****	-0.024	-0.091****	-0.082****
number of publications in Economics/Econometrics journals graded as 3, 4, 4* stars per author	-0.008	-0.036****	-0.013	-0.019**	-0.015	-0.021***	-0.033****	-0.045****	-0.021***	-0.024***
number of publications in Finance/Management journals graded as 3, 4, 4* stars per author	-0.046***	-0.043***	-0.060***	-0.053***	-0.009	0.011	-0.004	0.037***	-0.030**	-0.029**
proportion of publications in Economics/Econometrics journals	-0.086****	-0.073****	-0.079****	-0.083****	-0.037**	-0.076****	-0.084****	-0.090****	-0.074****	-0.076****
proportion of publications in Finance/Management journals	0.085****	0.075****	0.080****	0.083****	0.037*	0.076****	0.083****	0.090****	0.073****	0.074****
proportion of publications in journals graded as 3*, 4*, 4** stars	-0.154****	-0.189****	-0.142****	-0.163****	-0.009	-0.039***	-0.054****	-0.040****	-0.107****	-0.099****
proportion of publications in Economics/Econometrics journals graded as 3*, 4*, 4** stars	-0.011	-0.052****	-0.017	-0.030***	-0.022	-0.020	-0.038****	-0.041****	-0.027***	-0.029***
proportion of publications in Finance/Management journals graded as 3*, 4*, 4** stars	-0.024**	-0.066****	-0.087****	-0.043***	-0.008	0.020	-0.020*	-0.004	-0.033***	-0.032***

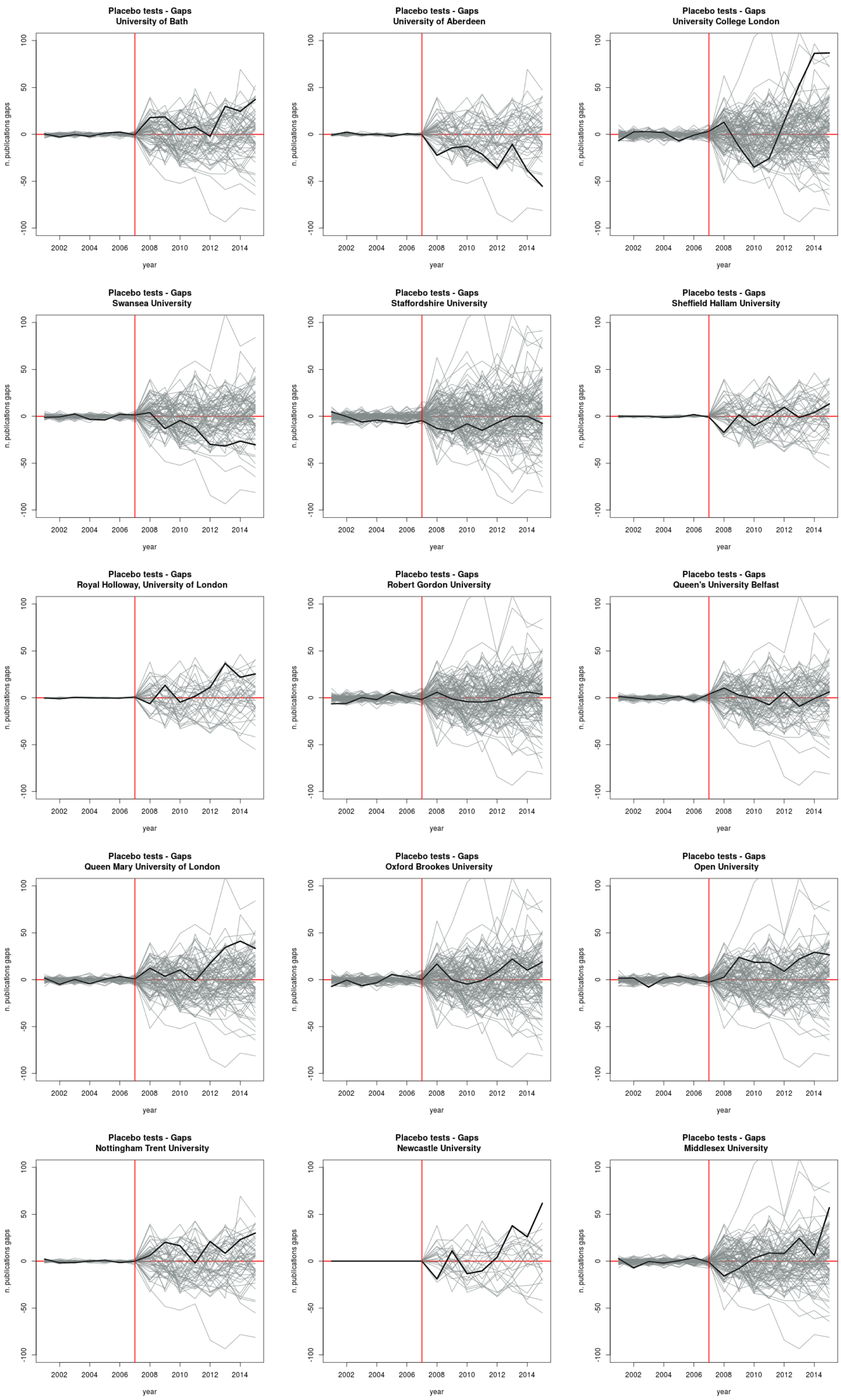
Values are marked by *, **, ***, **** if they are significant at a level of, respectively, 0.10, 0.05, 0.01 or 0.001.

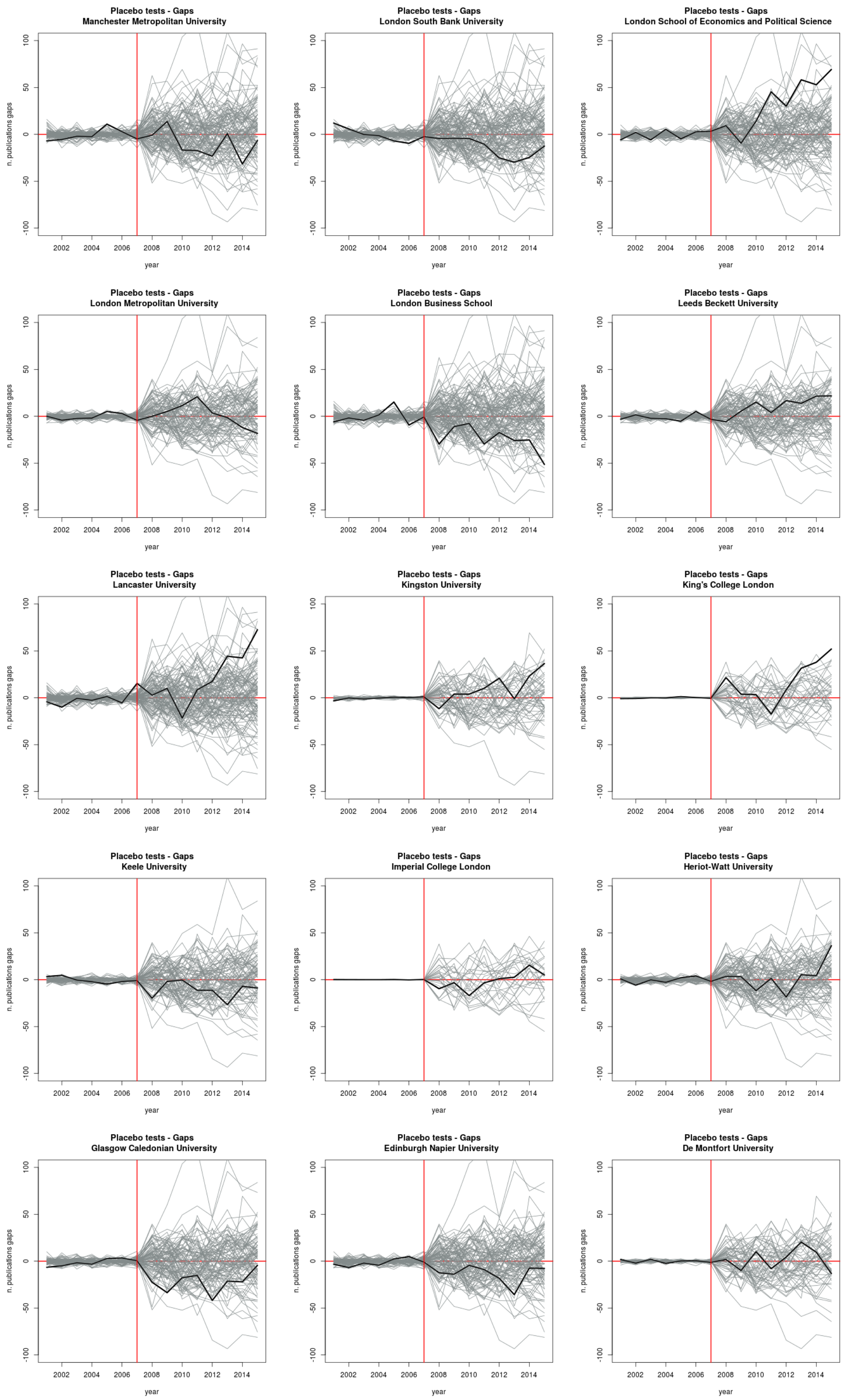
Figure A14: Graphs of placebo effects for the total number of publications.

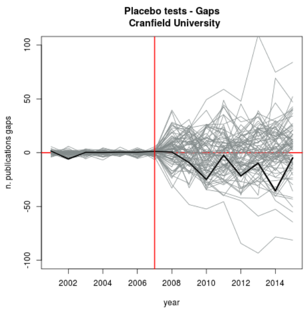


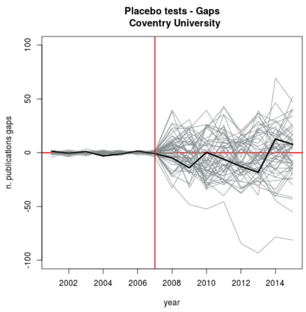

Placebo tests - Gaps
University of Salford
Placebo tests - Gaps
University of Reading
Placebo tests - Gaps
University of Portsmouth
Placebo tests - Gaps
University of Plymouth
Placebo tests - Gaps
University of Oxford
Placebo tests - Gaps
University of Nottingham
Placebo tests - Gaps
University of Northumbria at Newcastle
Placebo tests - Gaps
University of Manchester
Placebo tests - Gaps
University of Liverpool
Placebo tests - Gaps
University of Leicester
Placebo tests - Gaps
University of Leeds
Placebo tests - Gaps
University of Kent
Placebo tests - Gaps
University of Hull
Placebo tests - Gaps
University of Hertfordshire
Placebo tests - Gaps
University of Greenwich
n. publications gaps
year

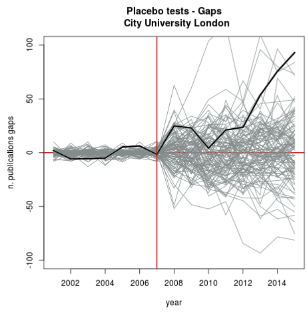

Placebo tests - Gaps
University of Glasgow
Placebo tests - Gaps
University of Exeter
Placebo tests - Gaps
University of Essex
Placebo tests - Gaps
University of Edinburgh
Placebo tests - Gaps
University of East London
Placebo tests - Gaps
University of East Anglia
Placebo tests - Gaps
University of Durham
Placebo tests - Gaps
University of Dundee
Placebo tests - Gaps
University of Central Lancashire
Placebo tests - Gaps
University of Cambridge
Placebo tests - Gaps
University of Bristol
Placebo tests - Gaps
University of Brighton
Placebo tests - Gaps
University of Bradford
Placebo tests - Gaps
University of Birmingham
Placebo tests - Gaps
University of Bedfordshire
n. publications gaps
year

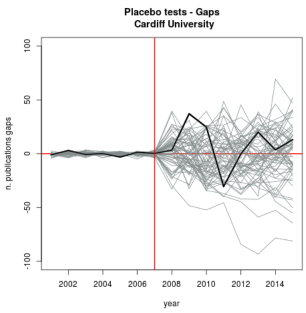

Placebo tests - Gaps
University of Bath
Placebo tests - Gaps
University of Aberdeen
Placebo tests - Gaps
University College London
Placebo tests - Gaps
Swansea University
Placebo tests - Gaps
Staffordshire University
Placebo tests - Gaps
Sheffield Hallam University
Placebo tests - Gaps
Royal Holloway, University of London
Placebo tests - Gaps
Robert Gordon University
Placebo tests - Gaps
Queen's University Belfast
Placebo tests - Gaps
Queen Mary University of London
Placebo tests - Gaps
Oxford Brookes University
Placebo tests - Gaps
Open University
Placebo tests - Gaps
Nottingham Trent University
Placebo tests - Gaps
Newcastle University
Placebo tests - Gaps
Middlesex University
n. publications gaps
year
100
50
0
-50
-100
2002
2004
2006
2008
2010
2012
2014

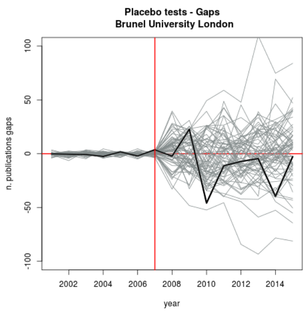

Placebo tests - Gaps
Manchester Metropolitan University
Placebo tests - Gaps
London South Bank University
Placebo tests - Gaps
London School of Economics and Political Science
Placebo tests - Gaps
London Metropolitan University
Placebo tests - Gaps
London Business School
Placebo tests - Gaps
Leeds Beckett University
Placebo tests - Gaps
Lancaster University
Placebo tests - Gaps
Kingston University
Placebo tests - Gaps
King's College London
Placebo tests - Gaps
Keele University
Placebo tests - Gaps
Imperial College London
Placebo tests - Gaps
Heriot-Watt University
Placebo tests - Gaps
Glasgow Caledonian University
Placebo tests - Gaps
Edinburgh Napier University
Placebo tests - Gaps
De Montfort University
n. publications gaps
-100
-50
0
50
100
2002
2004
2006
2008
2010
2012
2014
year

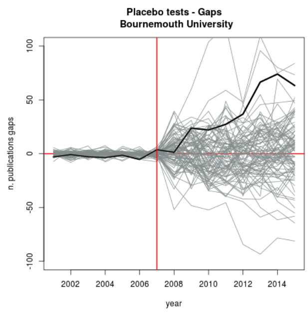

Placebo tests - Gaps
Cranfield University
n. publications gaps
year

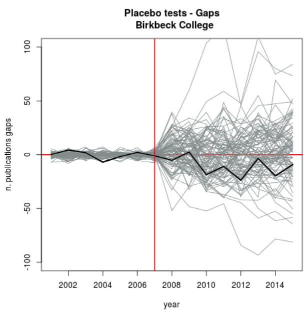

Placebo tests - Gaps
Coventry University
n. publications gaps
year

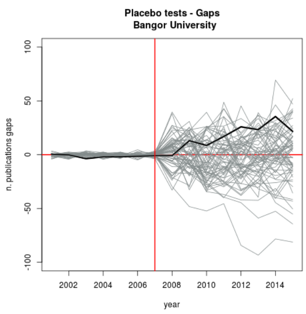

Placebo tests - Gaps
City University London
n. publications gaps
year

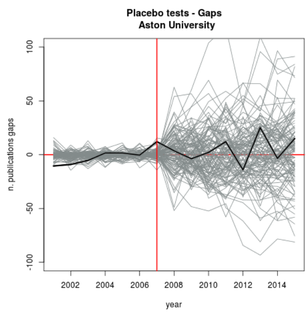

Placebo tests - Gaps
Cardiff University
n. publications gaps
year

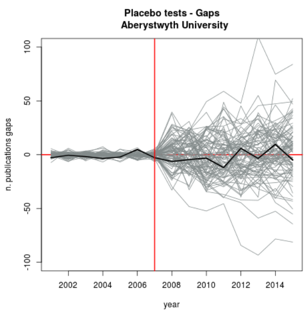

Placebo tests - Gaps
Brunel University London
n. publications gaps
year


Placebo tests - Gaps
Bournemouth University
n. publications gaps
year


Placebo tests - Gaps
Birkbeck College
n. publications gaps
year


Placebo tests - Gaps
Bangor University
n. publications gaps
year


Placebo tests - Gaps
Aston University
n. publications gaps
year


Placebo tests - Gaps
Aberystwyth University
n. publications gaps
year

Appendix B

Derivation of Eq. 5.14.

Since $\mathbf{I}^k = \mathbf{I}$ and $\mathbf{U}^{m-k} = n^{m-k-1}\mathbf{U}$,

$$
\begin{aligned}
\mathbf{\Phi}_1^m = (\Delta\lambda\mathbf{I} + \lambda_2\mathbf{U})^m &= \sum_{k=0}^{m} \binom{m}{k} \Delta\lambda^k \mathbf{I} \cdot \lambda_2^{m-k} n^{m-k-1}\mathbf{U} = \\
&= \mathbf{U} \sum_{k=0}^{m-1} \binom{m}{k} \Delta\lambda^k \lambda_2^{m-k} n^{m-k-1} + \Delta^m \mathbf{I} = \\
&= \frac{\mathbf{U}}{n} [\sum_{k=0}^{m-1} \binom{m}{k} \Delta\lambda^k (\lambda_2 n)^{m-k}] + \Delta\lambda^m \mathbf{I} = \\
&= \frac{\mathbf{U}}{n} [\sum_{k=0}^{m} \binom{m}{k} \Delta\lambda^k (\lambda_2 n)^{m-k} - \Delta\lambda^m] + \Delta\lambda^m \mathbf{I} = \\
&= \frac{\mathbf{U}}{n} [(\Delta\lambda + n\lambda_2)^m - \Delta\lambda^m] + \Delta\lambda^m \mathbf{I} = \mathbf{A}_m
\end{aligned}
$$

Derivation of the numerator of Eq. 5.15

1)

$$
\begin{aligned}
\mathbf{e}_i^T \mathbf{A}_h \mathbf{\Sigma} \mathbf{e}_j &= \mathbf{e}_i^T \frac{\mathbf{U}}{n} [(\Delta\lambda + n\lambda_2)^h - \Delta\lambda^h] \mathbf{\Sigma} \mathbf{e}_j + \Delta\lambda^h \mathbf{e}_i^T \mathbf{\Sigma} \mathbf{e}_j = \\
&= [(\Delta\lambda + n\lambda_2)^h - \Delta\lambda^h] \frac{1}{n} \mathbf{e}_i^T \mathbf{U} \mathbf{\Sigma} \mathbf{e}_j + \Delta\lambda^h \mathbf{e}_i^T \mathbf{\Sigma} \mathbf{e}_j = \\
&= [(\Delta\lambda + n\lambda_2)^h - \Delta\lambda^h] (\frac{1}{n} \sum_{i=1}^{n} \sigma_{ij}) + \Delta\lambda^h \sigma ij = \\
&= [(\Delta\lambda + n\lambda_2)^h - \Delta\lambda^h] \bar{\sigma}_j + \Delta\lambda^h \sigma_{ij}
\end{aligned}
\tag{18}
$$

Derivation of the denominator of Eq. 5.15

2)

$$
\begin{aligned}
\mathbf{e}_i^T \mathbf{A}_h \mathbf{\Sigma} \mathbf{A}_h^T \mathbf{e}_i &= \mathbf{e}_i^T \Big[\frac{\mathbf{U}}{n} [(\Delta\lambda + n\lambda_2)^h - \Delta\lambda^h] + \Delta\lambda^h \mathbf{I} \Big] \mathbf{\Sigma} \Big[\frac{\mathbf{U}}{n} [(\Delta\lambda + n\lambda_2)^h - \Delta\lambda^h] + \Delta\lambda^h \mathbf{I} \Big] \mathbf{e}_i = \\
&= \mathbf{e}_i^T \mathbf{U} \mathbf{\Sigma} \mathbf{U} \mathbf{e}_i \frac{1}{n^2} [(\Delta\lambda + n\lambda_2)^h - \Delta\lambda^h]^2 + \mathbf{e}_i^T \mathbf{\Sigma} \mathbf{e}_i \Delta\lambda^{2h} + \\
&+ \frac{1}{n} \mathbf{e}_i^T (\mathbf{U}\mathbf{\Sigma} + \mathbf{\Sigma}\mathbf{U}) \mathbf{e}_i \Delta\lambda^h [(\Delta\lambda + n\lambda_2)^h - \Delta\lambda^h] = \\
&= \frac{S_T}{n} [(\Delta\lambda + n\lambda_2)^h - \Delta\lambda^h]^2 + \sigma_{ii}^2 \Delta\lambda^{2h} + \Delta\lambda^h [(\Delta\lambda + n\lambda_2)^h - \Delta\lambda^h] - 2\bar{\sigma}_i
\end{aligned}
\tag{19}
$$

Substituting Eqs. 18 and 19 to Eq. 5.15, the 1-step-ahead generalized FEVD $(H = 1)$ is:

$$\theta_{ij}^{g}(H=1) = \frac{\sigma_{jj}^{-2}(\mathbf{e}_i^T \mathbf{A}_0 \mathbf{\Sigma} \mathbf{e}_j)^2}{\mathbf{e}_i^T \mathbf{A}_0 \mathbf{\Sigma} \mathbf{A}_0^T \mathbf{e}_i} = \frac{\sigma_{ij}^2}{\sigma_{ii}^2 \sigma_{jj}^2} = R_{ij}^2 \tag{20}$$

Instead, the 2-step-ahead generalized FEVD is

$$\begin{aligned}\theta_{ij}^{g}(H=2) &= \frac{1}{\sigma_{jj}^2}\left\{\frac{(\mathbf{e}_i^T \mathbf{A}_0 \mathbf{\Sigma} \mathbf{e}_j)^2 + (\mathbf{e}_i^T \mathbf{A}_1 \mathbf{\Sigma} \mathbf{e}_j)^2}{(\mathbf{e}_i^T \mathbf{A}_0 \mathbf{\Sigma} \mathbf{A}_0^T \mathbf{e}_i) + (\mathbf{e}_i^T \mathbf{A}_1 \mathbf{\Sigma} \mathbf{A}_1^T \mathbf{e}_i)}\right\} = \\ &= R_{ij}^2 \left\{\frac{1 + (\frac{n\lambda_2 \bar{\sigma}_j}{\sigma_{ij}} + \Delta\lambda)^2}{1 + \Delta\lambda^2 + \frac{S_T}{\sigma_{ii}^2}\lambda_2^2 + 2n\lambda_2 \Delta\lambda \frac{\bar{\sigma}_i}{\sigma_{ii}^2}}\right\}\end{aligned} \tag{21}$$

Table B1: Financial companies. AS=Asia; EU=European Union; ME=Middle East; OA= Other Asian; OC=Oceania; OE= European not in EU; RU=Russia; UE= Europe with own currency; US=US.

Geopolitical Area	*label*
ME	ABU DHABI COMR BK
EU	AEGON NV
AS	AGRI BANK OF CHINA
ME	AKBANK TURK ANONIM
EU	ALLIANZ SE
EU	ALLIED IRISH BANKS
EU	ALPHA BANK SA
AS	AOZORA BANK LTD
EU	ASSIC GENI-SO PER
OC	AU & NZ BANKING GP
UE	AVIVA PLC
EU	AXA
EU	BANCA MONTE PASCHI
EU	BANCO COM PORTUGUES
EU	BANCO DE SABADELL
EU	BANCO POP ESPANOL
EU	BANCO POPOLARE SOCO
EU	BANCO SANTANDER
US	BANK OF AMERICA
AS	BANK OF CHINA LTD
OA	BANK OF INDIA LTD
EU	BANK OF IRELAND
UE	BANK OF SCOTLAND
EU	BANKINTER SA
UE	BARCLAYS BANK PLC
EU	BAWAG PSK
EU	BAYERISCHE LANDESBK
EU	BBV ARGENTARIA
EU	BCA NAZ DEL LAVORO
EU	BCA PPO MILANO
US	BK NY MELLON CORP
RU	BK OF MOSCOW (OJSC)
EU	BNP PARIBAS SA
US	CAP 1 BK USA NA
AS	CHINA DEVELOPMENT BK
OC	CMWL BK OF AUSTRALIA
EU	COMMERZBANK AG
EU	COOP RABOBANK UA
EU	CREDIT AGRICOLE SA
EU	CREDIT LYONNAIS
OE	CREDIT SUISSE GROUP
UE	DANSKE BANK A/S
OA	DBS BANK LTD
EU	DE VOLKSBANK NV
EU	DEUTSCHE BANK AG
EU	DEXIA
EU	ERSTE GROUP BANK AG
EU	EUROBANK ERGASIAS
OA	EXP-IMP BK OF INDIA
UE	FCE BANK PLC
US	GOLDMAN SACHS GROUP
EU	HAMBURG COML BANK
EU	HANNOVER RUECK SE
UE	HBOS PLC
UE	HSBC BANK PLC

Geopolitical Area	*Label*
OA	ICICI BANK LIMITED
EU	IKB DT INDSTRBK AG
AS	IND & COM BK OF CHIN
AS	INDL BK OF KOREA
EU	ING BANK NV
EU	INTESA SANPAOLO SPA
OA	JSC BK CENTERCREDIT
RU	JSC VTB BANK
EU	KBC BANK
AS	KOOKMIN BANK
EU	LB BADENWUERTTEMBERG
EU	LB HESSTHRGN GIRO
UE	LLOYDS BANK
OC	MACQUARIE BANK LTD
OA	MALAYAN BKG BERHAD
EU	MEDIOBANCA SPA
AS	MIZUHO BANK LTD
US	MORGAN STANLEY
AS	MUFG BANK, LTD
EU	MUNICH REINSURANCE
EU	NAT BK OF GREECE SA
EU	NATIXIS SNR
UE	NATWEST MARKETS PLC
EU	NORDDEUTSCHE LB
UE	NORDEA BANK AB
AS	NORINCHUKIN BANK LTD
AS	OVERSEA-CHINESE BKC
EU	PORTIGON AG
EU	RAIF ZENTRALBANK
RU	SBERBANK OF RUSSIA
AS	SHINHAN BANK
UE	SKANDINAVISKA ENSK BNKN
EU	SOCIETE GENERALE
UE	STD CHARTERED BK
AS	SUMITOMO BK
UE	SVENSKA HB
UE	SWEDBANK AB
OE	SWISS REINSURANCE
AS	THE EXPT-IMPT BK OF CH
AS	THE EXPT-IMPT BK OF KO
AS	THE KOREA DEV BANK
US	THE PNC FIN SVS GP
ME	TURKIYE IS BANKASI
OE	UBS AG
EU	UNICREDIT BANK AG
EU	UNIONE DI BANCHE
US	UNITED OS BK LTD
EU	VAN LANSCHOT NV
OC	WESTPAC BANKING CORP
AS	WOORI BANK
OE	ZURICH INSURANCE

Table B2: Sovereigns. AS=Asia; EU=European Union; ME=Middle East; OA= Other Asian; OC=Oceania; OE= European not in EU; RU=Russia; UE= Europe with own currency; US=US.

Geopolitical Area	*Label*
OC	COMMONWEALTH OF AUSTRALIA
UE	CZECH REPUBLIC
EU	GERMANY
RU	RUSSIA
EU	HELLENIC REPUBLIC
UE	HUNGARY
AS	JAPAN
EU	IRELAND
ME	KINGDOM OF BAHRAIN
EU	KINGDOM OF BELGIUM
UE	KINGDOM OF DENMARK
EU	KINGDOM OF NETH
OE	KINGDOM OF NORWAY
EU	KINGDOM OF SPAIN
UE	KINGDOM OF SWEDEN
OA	KINGDOM OF THAILAND
OA	MALAYSIA
AS	REP OF CHINA
OA	REPUBLIC OF INDONESIA
OA	REPUBLIC OF KAZAKHSTAN
EU	REPUBLIC OF LITHUANIA
OA	REPUBLIC OF PHILIPINES
EU	REPUBLIC OF AUSTRIA
UE	REPUBLIC OF BULGARIA
UE	REPUBLIC OF CROATIA
EU	REPUBLIC OF CYPRUS
EU	REPUBLIC OF ESTONIA
EU	REPUBLIC OF FINLAND
EU	REPUBLIC OF ITALY
EU	REPUBLIC OF FRANCE
AS	REPUBLIC OF KOREA
EU	REPUBLIC OF LATVIA
UE	REPUBLIC OF POLAND
EU	REPUBLIC OF PORTUGAL
EU	REPUBLIC OF SLOVENIA
ME	REPUBLIC OF TURKEY
UE	ROMANIA
EU	SLOVAK REPUBLIC
ME	STATE OF QATAR
UE	UK AND NI
US	USA

References

[1] Aase,K. and Persson,S.A. (1997) *Valuation of the minimum guaranteed return*, JRI, 599–617, 64, n.4.

[2] Abadie, A. and Gardeazabal, J. (2003) *The Economic Costs of Conflict: A Case Study of the Basque Country*, The American Economic Review, pag.113–132, vol.93, n.1, American Economic Association.

[3] Abadie, A., Diamond, A., and Hainmueller, J. (2010) *Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program*, Journal of the American Statistical Association, pag.493-505, vol.105, n.490, American Economic Association.

[4] Abadie A, Diamond A, Hainmueller, J (2015), **Comparative politics and the synthetic control method** , Am J Pol Sci, (59) 495–510.

[5] D. Acemoglu, V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi (2012) *The network origins of aggregate fluctuations*, Econometrica, 80:1977–2016.

[6] D. Acemoglu, A. Ozdaglar, and A. Tahbaz-Salehi (2015). *Systemic risk and stability in financial networks*. American Economic Review, 105:564–608.

[7] Acemoglu, D. and Johnson, S. and Kermani, A. and Kwak, J., and Mitton, T. (2016) *The value of connections in turbulent times: Evidence from the united states*, Journal of Financial Economics, pag.368–391, vol.121, American Economic Association.

[8] V. Acharya, I. Drechsler, and P. Schnabl (2014). A pyrrhic victory? bank bailouts and sovereign credit risk. The Journal of Finance, 69:2689–2739.

[9] Albizzati, M. O. and Geman, H. (1994) *Interest rate risk management and valuation of surrender option in life insurance policies*, JRI, 616-637, 61, n.4.

[10] Anderberg, M. R (1973) *Cluster Analysis for Applications*, Academic Press, New York.

[11] Arbenz, P. and Hummel, C. and Mainik, G. (2012) *Copula based hierarchical risk aggregation through sample reordering*, Insurance: Mathematics and Economics, 122-133, 51, n.1.

[12] Babbel, D.F. (2001) *Asset/liability management for insurers in the new era: Focus on value*, JRF, 9–17.

[13] Babbel, D.F. and Merril, C. (1998) *Economic valuation models for insurers*, North American Actuarial Journal, 1–17, 2, n.3.

[14] Bacinello, A.R. (2005) *Endogenous model of surrender conditions in equity-linked life insurance*, IME, 270–296, 37.

[15] Bacinello, A.R. (2003) *Fair valuation of a guaranteed life insurance participating contract embedding a surrender option*, JRI, 461–487, 70, n.3.

[16] Bacinello, A.R. (2001) *Fair pricing of life insurance participating policies with a minimum guarantee*, Astin Bulletin, 275–297, 31, n.2.

[17] Bak, P., Tang, C. and Wiesenfeld, K. (1987) *Self-organized criticality: an explanation of 1/f noise.*, 59 (4): 381–384, Physical Review Letters.

[18] Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286: 509–512.

[19] Barabási AL, Jeong H, Neda Z, Ravasz E, Schubert A, et al. (2002) Evolution of the social network of scientific collaborations. Physica A 311: 590–614.

[20] Bayraktar, E. and Young, V. (2008) *Pricing Options in Incomplete Equity Markets via the Instantaneous Sharpe Ratio*, Annals of Finance, 4(4), pp 399–429.

[21] Beck, U. (2006) *Cosmopolitan Vision*, Polity; 1st edition.

[22] Belhadji, E. and Dionne, G. and Tarkhani, F. (2000) *A Model for the Detection of Insurance Fraud*, The Geneva Papers on Risk and Insurance - Issues and Practice, 517–538, 25, n.4.

[23] A. Belloni and V. Chernozhukov (2011). *L1-penalized quantile regression in high-dimensional sparse models.* The Annals of Statistics, 39(1):82–130.

[24] Bence, V. and Oppenheim, C. (2005) *The evolution of the UK's research assessment exercise: publications, performance and perceptions*, Journal of Educational Administration and History, 137–155, 37, n.2.

[25] Berketi, A.K. and MacDonald, A.S. (1999) *The effect of the nature of the liabilities on the solvency and maturity of a U.K. life office fund: A stochastic evaluation*, IME, 117–138, 24.

[26] Bermúdez, L. and Pérez, J. and Ayuso, M. and Gómez, E. and Vázquez, F. (2008) *A Bayesian dichotomous model with asymmetric link for fraud in insurance*, Insurance: Mathematics and Economics, 779–786, 42, n.2.

[27] Blondel, Vincent D; Guillaume, Jean-Loup; Lambiotte, Renaud; Lefebvre, Etienne (2008) Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 10, P10008.

[28] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU, (2006) Complex networks: Structure and dynamics, Physics Reports 424, 175 – 308.

[29] Bouttell J, Craig P, Lewsey J, et al. (2018), *Synthetic control methodology as a tool for evaluating population-level health interventions*,, J Epidemiol Community Health (72) 673–678.

[30] Boyle, P.P and Hardy, M.R. (1997) *Reserving for maturity guarantees: Two approaches*, IME, 113–127, 21, n.2.

[31] Boyle, P.P. and Schwartz, E.S. (1977) *Equilibrium prices of guarantees under equity-linked contracts*, JRI, 639–660, 44.

[32] Brennan, M.J. and Schwartz, E.S (1976) *The pricing of equity-linked life insurance policies with an asset value guarantee*, Journal of Financial Economics, 195–213, 3.

[33] Brennan, M.J. and Schwartz, E.S. (1979) *Alternative investment strategies for the issuers of equity linked life insurance policies with an asset value guarantee*, Journal of Business, 63–93, 52.

[34] Briys, E. and de Varenne, F. (2001) *Insurance: From underwriting to derivatives: Asset/liability management in insurance companies*, John Wiley & Sons.

[35] Briys, E. and de Varenne, F. (1997) *On the risk of life insurance liabilities: Debunking some common pitfalls*, JRI, 673–694, 64, n.4.

[36] Brockett, P. and Derrig, R. and Golden, L. and Levine, A. and Alpert, M. (2002) *Fraud Classification Using Principal Component Analysis of RIDITs*, Journal of Risk and Insurance, 341-371, 69, n.3.

[37] Caudill, S. and Ayuso, M. and Guillén, M. (2005) *Fraud Detection Using a Multinomial Logit Model With Missing Information*, Journal of Risk and Insurance, 539-550, 72, n.4.

[38] Chadburn, R.G. (1997) *The use of capital, bonus policy and investment policy in the control of solvency for with-profits life insurance companies in the U.K.*, City University London.

[39] Cheng C, Sa-Ngasoongsong A, Beyca OF, Le T, Yang H, Kong Z, Bukkapatnam S (2015), *Time Series Forecasting for Nonlinear and Nonstationary Processes: A Review and Comparative Study*, IIE Transactions.

[40] Chesson J (1976) J. Appl. Probab. **13**, 795-797.

[41] Cho Y. H. and Kim J. K. (2004) Expert Sys. Appl. **26**, 233.

[42] Ciriello G, Guerra C. (2008) *A review on models and algorithms for motif discovery in protein–protein interaction networks. Brief Funct Genomics.* 7:147–56

[43] Clerides, Sofronis and Pashardes, Panos and Polycarpou, Alexandros (2011) *Peer Review vs Metric-based Assessment: Testing for Bias in the RAE Ratings of UK Economics Departments*, Economica, pag.565-583, vol.78, n.311.

[44] Coleman, T.F. and Li, Y. and Patron, M. (2006) *Hedging guarantees in variable annuities under both equity and interest rate risks*, IME, 215–228, 38.

[45] Coleman, T.F. and Kim, Y. and Li, Y. and Patron, M. (2007) *Robustly hedging of variable annuities with guarantees under jump and volatility risks*, JRI, 347–376, 74, n.2.

[46] Colizza, V *et al.* (2006) Nat. Phys **2**, 110-115.

[47] G. Covi and U. Eydam (2017) *The Bank Recovery and Resolution Directive* Academic Press. Working Papers researchgate.

[48] Dacorogna, M.M., Gencay, R., Müller, U.A., Olsen, R.B., Pictet, O.V (2001) *An Introduction to High-Frequency Finance.* Academic Press.

[49] Dahl, M. and Møller, T. (2006) *Valuation and hedging of life insurance liabilities with systematic mortality risk*, IME, 193–217, 28, n.203.

[50] Dahl, M. and Melchior, M. and Møller, T. (2008) *On systematic mortality risk and risk-minimization with survivor swaps*, Scandinavian actuarial journal, 114–146, 28, n.203.

[51] De Lange, P.E. and Fleten, S.E. and Gaivoronski, A.A. (2004) *Modeling financial reinsurance in the casualty insurance business via stochastic programming*, Journal of Economic Dynamics & Control, 991–1012, 28, n.5.

[52] Dekker, S. (2011) *Drift into Failure: From Hunting Broken Components to Understanding Complex Systems*, CRC Press.

[53] M. Demirer, F. X. Diebold, L. Liu, and K. Yilmaz (2018) *Estimating global bank network connectedness.*, Journal of Applied Econometrics, 33:1–15.

[54] Derrig, R. (2002) *Insurance Fraud*, Journal of Risk and Insurance, 271-287, 69, n.3.

[55] F. X. Diebold and K. Yilmaz (2014) *On the network topology of variance decompositions: Measuring the connectedness of financial firms*, Journal of Econometrics, 182:119–134.

[56] F. X. Diebold and K. Yilmaz (2012) *International journal of forecasting*, Journal of Econometrics, 28:57–66.

[57] Doyle, J. and Arthurs, A. and Green, R. and McAulay, L. and Pitt, M. and Bottomley, P. (1996) *The judge, the model of the judge and the model of the judged as judge: analysis of the UK 1992 RAE data for business and management studies*, Omega: International Journal of Management Science, pag.21, vol.24, n.1.

[58] Duch, J and Arenas A (2005), *Community detection in complex networks using extremal optimization*, Physical review E, 72, 027104.

[59] Duffie, D. and Richardson, H.R. (1991) *Mean-variance hedging in continuous time*, Annals of Applied Probability, 1–15, 1, n.2.

[60] D. Easley and J. Kleinberg (2010), Networks, Crowds, and Markets: Reasoning About a Highly Connected World, Cambridge University Press.

[61] El Karoui, N. and Quenez, M.C. (1995) *Dynamic programming and pricing of contingent claims in an incomplete market*, SIAM Journal on Control and Optimization, 29–66, 33.

[62] M. Elliott, B. Golub, and M. O. Jackson (2014) *Financial networks and contagion. The American Economic Review*, SIAM Journal on Control and Optimization, 104:3115–3153.

[63] Embrechts, P. (2000) *Actuarial versus financial pricing of insurance*, JRF, 17–26, 1, n.4.

[64] Erdős, P., Rényi, A. (1959) "On Random Graphs. I" (PDF). Publicationes Mathematicae. 6, 290–297,

[65] J. Fan and R. Li (2001) Variable selection via non-concave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456):1348–1360.

[66] E. Farhi and J. Tirole (2018) Deadly embrace: Sovereign and financial balance sheets doom loops. The Review of Economic Studies, 85(3):1781–1823.

[67] Fiasconaro, A *et al.* (2015) Phys. Rev. E **92**, 012811.

[68] Föllmer, H. and Sondermann, D. (1986) *Hedging of non-redundant contingent claims*, Contributions to Mathematical Economics, North-Holland, Hildenbrand,W. and Mas-Colell,A. 205–223.

[69] Föllmer, H. and Schweizer, M. (1989) *Hedging by sequential regression: An introduction to the mathematics of option trading*, Astin Bulletin, 147–160,1.

[70] M. Forni, M. Hallin, M. Lippi, and P. Zaffaroni (2015) *Dynamic factor models with infinite-dimensional factor space: Asymptotic analysis.* Journal of Econometrics, 199:74–92.

[71] M. Forni, M. Hallin, M. Lippi, and P. Zaffaroni (2015) *Dynamic factor models with infinite-dimensional factor spaces: One-sided representations.* Journal of Econometrics, 185:359–371.

[72] M. Forni and M. Lippi (2000) *The generalized dynamic factor model: Representation theory*, Econometric Theory, 17:1113–1141, 2001.

[73] M. Forni, M. Hallin, M. Lippi, and L. Reichlin (2000) *The generalized dynamic-factor model: identification and estimation*, The Review of Economics and Statistics, 82:540–554.

[74] Fortunato S (2010), *Community detection in graphs*, Physics Reports, 486, 75-174.

[75] Fraley, C.; Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation, Journal of the America Statistical Association, **97**: 611-631.

[76] Franzoni, C. and Scellato, G. and Stephan, P. (2011) *Changing Incentives to Publish*, Science, vol.33.

[77] Giraldi, C. and Susinno, G. and Berti, G. and Brunello, J. and Buttarazzi, S. and Cenciarelli, G. and Daroda, C. and Stamegna, G. (2003) *Insurance optional; Exotic options: The cutting-edge collection technical papers published in Risk 1999-2003*, ch.35, Risk Books, Lipton, A.

[78] Girvan, M. and Newman, M.E.J. (2002) *Community structure in social and biological networks*, Proceedings of the national academy of sciences, 7821-7826, 99.

[79] Glass, C.J. and McCallion, Gillian and McKillop, Donal G. and Rasaratnam, Syamarlah and Stringer, Karl S. (2006) *Implications of variant efficiency measures for policy evaluations in UK higher education*, Socio-Economic Planning Sciences, pag.119-142, vol.40, n.2.

[80] P. Glasserman and H. P. Young (2016) *Contagion in financial networks.* Journal of Economic Literature, 54:779–831.

[81] P. Glasserman and H. P. Young (2015) *How likely is contagion in financial networks?*, Journal of Banking and Finance, 50:383–399.

[82] Gleick, J. (1987) *Chaos: making a new science Chaos: making a new science*, Penguin Books New York.

[83] Gorodkin, J (2004) *Comparing two K-category assignments by a K-category correlation coefficient*, Computational Biology and Chemistry **28**(5-6), 367-274.

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

[85] Grochow JA, Kellis M. *Network motif discovery using sub-graph enumeration and symmetry-breaking.* Res Comp Mol Biol. 2007;4456:92–106.

[86] Grosen, A. and Jørgensen, P.L. (2002) *Life insurance liabilities at market value: An analysis of insolvency risk, bonus policy, and regulatory intervention rules in a barrier option framework*, JRI, 63–91, 69, n.1.

[87] Grosen, A. and Jørgensen, P.L. (2000) *Fair valuation of life insurance liabilities: The impact of interest rate guarantees, surrender options, and bonus policies*, IME, 37–57, 26.

[88] Grosen, A. and Jørgensen, P.L. (1997) *Valuation of early exercisable interest rate guarantees*, JRI, 481–503, 64, n.3.

[89] D. Gross and P. Siklos (2019) *Analyzing credit risk transmission to the non-financial sector in europe: A network approach*, Journal of Applied Econometrics.

[90] Guimera R, Uzzi B, Spiro J, Amaral LAN (2005) Team assembly mechanisms determine collaboration network structure and team performance. Science 308:697–702.

[91] M. Hallin and R. Liška (2007) *Determining the number of factors in the general dynamic factor model.* Journal of the American Statistical Association, 102:603–617.

[92] Hansen, M.S. and Hansen, M. (2001) *Portfolio choice and fair pricing in life insurance companies*, working paper, University of Southern Denmark, n.4.

[93] Hansen, M. and Miltersen, K.R. (2002) *Minimum rate of return guarantees: The danish case*, Scandinavian Actuarial Journal, 230–318, 2002, n.4.

[94] Harman, Grant (2000) *Allocating Research Infrastructure Grants in Post-binary Higher Education Systems: British and Australian approaches*, Journal of Higher Education Policy and Management, pag.111-126, vol.22, n.2.

[95] Hartigan, J. A. et al. (1979) Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), **28** (1): 100-108.

[96] T. Hastie, R. Tibshirani, and J. Friedman (2009) Journal of the Royal Statistical Society. Series C (Applied Statistics), **28** (1): 100-108. The Elements of Statistical Learning. Springer.

[97] Hatzopoulos V. *et al.* (2015) Quant. Fin. **15**, 693.

[98] N. Hautsch, J. Schaumburg, and M. Schienle (2014). *Financial network systemic risk contributions.* Review of Finance, 19:1–54.

[99] Hayri, A. (1997) *The research assessment exercise and transfer of academics among departments*, Risk Decision and Policy, pag.71-86, vol.2, n.1.

[100] HEFCE (2009) *Research Excellence Framework. Second consultation on the assessment and funding of research*,`http://www.hefce.ac.uk/media/hefce1/pubs/hefce/2009/0938/09_38.pdf`.

[101] Henkel, M (1999) *The modernisation of research evaluation: The case of UK*, Higher Education, 105–122, 38, n.1.

[102] Herlocker, J. L. , Konstan, J. A., Borchers, A. and Riedl J (1999) in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 230-237.

[103] Hicks, D. (2012) Performance-based university research funding systems, Research Policy 41, 251– 261.

[104] Hole, A.R. (2017), *Ranking Economics Journals Using Data From a National Research Evaluation Exercise*, Oxford Bulletin of Economics and Statistics, 79(5), 0305–9049.

[105] Holland, PW, Laskey, KB and Leinhardt S (1983) *Stochastic block- models: some first steps*, Social Networks, 5(2), 109137.

[106] Horn, RA and Johnson, CR (1990) *Norms for Vectors and Matrices*, in Matrix Analysis. Cambridge, England: Cambridge University Press.

[107] Høyland, K. (1998) *Ph.D. thesis: Asset liability management for a life insurance company: A stochastic programming approach*, Norwegian University of Science and Technology, Trondheim, Norway.

[108] Iori, G. *et al.* (2008) J. Econ. Dyn. Contr. **32**, 259.

[109] Iorio, F., Bernardo-Faura, M., Gobbi, A., Cokelaer, T., Jurman, G. and Saez-Rodriguez, J (2016) Efficient randomization of biological networks while preserving functional characterization of individual nodes. BMC Bioinformatics, **17**(1), p.542.

[110] Jaccard, P. (1912) New Phytologist **11**, 37-50.

[111] Johnson, N. (2009) *Simply Complexity: A Clear Guide to Complexity Theory*, Oneworld Publications.

[112] Jørgensen,P.L. (2001) *Life insurance contracts with embedded options*, Center for Analytical Finance, Working Paper, n.96.

[113] Jensen, B. and Jørgensen, P. L. and Grosen, A. (2001) *A finite difference approach to the valuation of path dependent life insurance liabilities*, The GENEVA Papers on Risk and Insurance - Theory, 57–84, 26, n.1.

[114] Johnes J, Taylor J, and Francis B (1993) *The Research Performance of UK Universities: A Statistical Analysis of the Results of the 1989 Research Selectivity Exercise*, Journal of the Royal Statistical Society. Series A (Statistics in Society), pag.271–286, vol.156, n.2.

[115] Johnson, N. F. (2009) *Simply Complexity: A Clear Guide to Complexity Theory*, Oneworld.

[116] Kashtan N, Itzkovitz S, Milo R, Alon U (2004). *Efficient sampling algorithm for estimating sub-graph concentrations and detecting network motifs*. Bioinformatics. 20 (11): 1746–1758

[117] Karrer, B, Newman, MEJ (2011), Stochastic blockmodels and community structure in networks. Phys. Rev. E 83, 016107.

[118] Kenett DY, Tumminello M, Madi A, Gur-Gershgoren G, Mantegna RN, and Ben-Jacob E (2010), *Dominating Clasp of the Financial Sector Revealed by Partial Correlation Analysis of the Stock Market*, PLOS ONE, 12, 5.

[119] Kraskov, A., Stogbauer, H., Grassberger, P. (2004) Phys.Rev.E, **69** 066138.

[120] Kuo, W and Tsai, C. and Chen,C. (2003)*An empirical study on the lapse rate: the cointegration approach*, JRI, 489-508, 70, n.3.

[121] La Manna, M. A. (2008) *Assessing the assessment or, the RAE and the optimal organization of University research*, Scottish Journal of Political Economy, pag.637–653, vol.55, n.5.

[122] Laloux, L., Cizeau, P., Bouchaud, J-P., Potters, M. (1999) *Noise Dressing of Financial Correlation Matrices*, Phys Rev Lett, 1467–1470, 83, n.1.

[123] Larremore, D. B., Clauset, A. (2014) & Jacobs, A. Z. Efficiently inferring community structure in bipartite networks. Physical Review E 90, 012805.

[124] T. Lewis, *Network Science*, Wiley 2009.

[125] Lorenz, E.N. (1963) *Deterministic Non-Periodic Flow*, Journal of the Atmospheric Sciences, 20, 130-141.

[126] Latora V, Nicosia V, Russo G (2017) *Complex networks: Principles, methods and applications*, Cambridge University Press.

[127] Lü, L., Medo, M., Yeung, C. H., Zhang, Y.C., Zhang, Z.K., Zhou, T. (2012) Phys. Rep. **519**, 1.

[128] Lubell, M., Zahran, S., Vedlitz, A. (2007) *Collective Action and Citizen Responses to Global Warming*, Political Behaviour, 29 (3), 391–413.

[129] H. Lütkepohl (2007) *Introduction to multiple time series analysis*, Springer-Verlag, Berlin, 1991.

[130] MacMahon, M, Garlaschelli, D (2015) Phys. Rev. X **5**, 021006.

[131] Manly, BFJ (1974) Biometrics **30**, 281-294.

[132] Martin, B (2011) *The Research Excellence Framework and the 'impact agenda':are we creating a Frankenstein monster?*, Research Evaluation, 247–254, 20, n.3.

[133] Matthews, B. W. (1975) *Comparison of the predicted and observed secondary structure of T4 phage lysozyme*, Biochimica et Biophysica Acta (BBA) - Protein Structure., 405 442–451 n.2.

[134] McCallum A, Wang XR, Corrada-Emmanuel A (2007) Topic and role discovery in social networks with experiments on Enron and academic email. J Artif Intell Res 30: 249–272.

[135] McPherson M, Smith-Lovin L, and Cook J. (2001) Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology 27: 415–444.

[136] Milevsky, M. A. and Posner, S. (2001) *The Titanic option: valuation of the guaranteed minimum death benefit in variable annuities and mutual funds*, JRI, 55-79, 68, n.1.

[137] Milevsky, M. and Salisbury, T. (2006) *Financial valuation of guaranteed minimum withdrawal benefits*, IME, 21–38, 38.

[138] Miller, J.H., Page, S.E. (2007) *Complex Adaptive Systems. An introduction to computational models of social life.* Princeton: Princeton University Press.

[139] Milo R, Shen-Orr SS, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002). *Network motifs: simple building blocks of complex networks*. Science. 298 (5594): 824–827

[140] Miltersen, K.R. and Persson, S. (1999) *Pricing rate of return guarantees in a Heath–Jarrow–Morton framework*, IME, 307–325, 25.

[141] Moed, H. (2008) *UK Research Assessment Exercise: Informed judgement on research quality or quantity?*, Scientometrics, pag.153-161, vol.74, n.2.

[142] Møller, T. (2003) *Indifference pricing of insurance contracts in a product space model: applications*, IME, 295–315, 32, n.2.

[143] Møller, T. (2003) *Indifference pricing of insurance contracts in a product space model*, Finance and stochastics, 197–217, 7, n.2.

[144] Møller, T. (2002) *On valuation and risk management at the interface of insurance and finance*, British Actuarial Journal, 787-827, 8, n.4.

[145] Møller, T. (2001) *Hedging equity-linked life insurance contracts*, North American Actuarial Journal, 79–95, 5, n.2.

[146] Moore, K.S. and Young, V.R. (2003) *Pricing equity-linked pure endowments via the principle of equivalent utility*, IME, 497–516, 33.

[147] Mryglod, O, Kenna R, Holovatch Yu, Berche B (2015), *Predicting results of the Research Excellence Framework using departmental h-index*, Journal Scientometrics, 102(3), 2165-2180.

[148] Newman, M.E.J. (2010), *Networks: An Introduction*, Oxford University Press.

[149] Newman M.E.J. (2001) Phys.Rev. E **64**, 016131.

[150] Newman M.E.J. (2001) Phys.Rev. E **64**, 016132.

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

[152] Newman M.E.J. (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci USA 98: 404–409.

[153] Omidi S, Schreiber F, Masoudi-Nejad A (2009). "MODA: an efficient algorithm for network motif discovery in biological networks". Genes Genet Syst. 84 (5): 385–395

[154] Onnela J. P. *et al.* (2007) Proc. Nat. Aca. Sci. **104**, 7332.

[155] Onnela J. P. *et al.* (2007) New J. Phys. **9**, 179.

[156] Peixoto, T. P (2014) Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. Physical Review X 4, 011047.

[157] H.H. Pesaran and Y. Shin. (1998). Generalized impulse response analysis in linear multivariate models. Economics Letters, 58:17–29, 1998.

[158] Plerou V. *et al.* (1999) Phys. Rev. Lett. **83**, 1471.

[159] Puccio, E. et al. (2016), Phys. A **462**, 167-185.

[160] Puccio E, Vassallo P, Piilo J, Tumminello M (2019), *Covariance and Correlation estimators in bipartite complex systems with a double heterogeneity*, Journal of Statistical Mechanics:Theory and Experiments, **5**, 053404.

[161] Ravazzolo F. Rigobon R. Caporin M., Pelizzon L. (2018). *Measuring sovereign contagion in europe.* Journal of Financial Stability, 34(C):150–181.

[162] Roberts, C. (1999) *Possible bias due to panel membership in the 1996 RAE*,Journal of the Royal Statistical Society International Conference, Warwick University.

[163] Robertson, R. (1995) *Glocalization: Time-space and homogeneity-heterogeneity R Robertson* - Global modernities.

[164] Ross, M.D. (1989) *Modelling a with-profit life office*, Journal of the Institute of Actuaries, 691-716, 116.

[165] Rosvall, M. and Bergstrom, CT (2008), *Maps of random walks on complex networks reveal community structure*, Proceedings of the National Academy of Sciences, 105(4), 1118-1123

[166] Schreiber F, Schwöbbermeyer H (2005). *Frequency concepts and pattern detection for the analysis of motifs in networks.* Transactions on Computational Systems Biology III. Lecture Notes in Computer Science. 3737. pp. 89–104.

[167] Seidman S, (1983), *Network structure and minimum degree*, Social Networks, (5) 269-287.

[168] Shi,P. and Frees, W. (2010) *Long-tail longitudinal modeling of insurance company expenses*, Insurance: Mathematics and Economics, 303–314, 47, n.3.

[169] Siglienti, S. (2000) *Consequences of the reduction of interest rates on insurance*, The Geneva Papers on Risk and Insurance, 63–77, 25, n.1.

[170] Simon, A.H. (1969) *The Sciences of the Artificial*, MIT Press.

[171] Smith, S. and Ward, V. and House, A. (2011) *Impact in the proposals for the UK's Research Excellence Framework: Shifting the boundaries of academic autonomy*, Research Policy, pag.1369-1379, vol.40.

[172] Solvency and Actuarial Issues Subcommittee (2006) *Standard on Asset-Liability Management*, Available at IAIS website: http://www.iaisweb.org, 61, n.13.

[173] Song CM, Havlin S, Makse HA (2005) Self-similarity of complex networks. Nature 433: 392–395.

[174] Stoutenborough, J., Kirkpatrick, K., Field, M., Vedlitz, A. (2015) *What butterfly effect? The contextual differences in public perceptions of the health risk posed by climate change*, Climate, 3(3), 668-688.

[175] Straub, E. (1997) *Non–Life Insurance Mathematics*, Springer.

[176] Šubelj, L. and Furlan, S. and Bajec, M. (2011) *An expert system for detecting automobile insurance fraud using social network analysis*, Expert Systems with Applications, 1039–1052, 38, n.1.

[177] Taylor J (2011), *The assessment of research quality in UK universities: peer review or metrics?*, British Journal of Management 22 (2), 202-217

[178] Tatusov, R. L., Koonin E. K., and Lipman D. J. (1997) Science **278**, 631637.

[179] Tatusov, R. L. (2003) *et al.*, BMC Bioinformatics **4**, 41.

[180] The Association of British Insurer, G. (2017) *UK Insurance & Long Term Savings Key Facts*, Insurance: Mathematics and Economics, `https://www.abi.org.uk/data-and-resources/industry-data/uk-insurance-and-long-term-savings-key-facts/`.

[181] R. Tibshirani (1996) *Regression analysis and selection via the lasso.* Journal of the Royal Statistical Society, Series B, 58(1):267–288.

[182] Tumminello, M. *et al.* (2013) PLoS One **8**, e64703.

[183] Tumminello, M *et al.* (2011) PLoS One **6**, e17994.

[184] Tumminello, M. and Edling, C. and Liljeros,F. and Mantegna, R.N. and Sarnecki, J. (2013) *The Phenomenology of Specialization of Criminal Suspects*, PLOS ONE, 1-8, 8, n.5 .

[185] Tumminello, M. *et al.* (2014) New J. Phys. **14**, 013041.

[186] Urry, J. (2000) *Mobile sociology*, The British journal of sociology, 51 (1), 185-203.

[187] Van Vlasselaer, V. and Eliassi-Rad, T. and Akoglu, L. and Snoeck, M. and Baesens, B., (2017) GOTCHA! Network-Based Fraud Detection for Social Security Fraud, Management Science, 63(9), 3090-3110

[188] Vanderhoof, I.T. and Altman, E.I. (1998) *The fair value of insurance liabilities*, Kluwer Academic Publishers, British Actuarial Journal, Boston, 777-964, 1, n.5.

[189] Viaene, S. and Dedene, G. (2004) *Insurance Fraud: Issues and Challenges*, The Geneva Papers on Risk and Insurance - Issues and Practice, 313–333, 29, n.2.

[190] Wallenius, K.T. (1963) Biased Sampling: The Non-central Hypergeometric Probability Distribution. Ph.D. Thesis (Thesis). Stanford University, Department of Statistics.

[191] Wang, Jian and Hicks, Diana (2013) *Detecting structural change in university research systems: A case study of British research policy*, Research Evaluation, pag.258–268, vol.22.

[192] Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393: 440–442.

[193] Watts, DJ (2003) Six Degrees: The Science of a Connected Age, W. W. Norton & Company.

[194] Wernicke S (2006). *Efficient detection of network motifs*. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 3 (4): 347–359.

[195] Wilkie, A.D. (1995) *More on a stochastic asset model for actuarial use*, British Actuarial Journal, 777-964, 1, n.5.

[196] Woods, D. (2011) *Big Data Requires a Big, New Architecture*, Insurance: Mathematics and Economics, `https://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/#429139171157`.

[197] Wolfram, S. (2002) *A New Kind of Science*, Wolfram Media.

[198] Zhou T, Kuscsik Z, Liu JG, Medo M, Wakeling JR, and Zhang YC (2010) *Solving the apparent diversity-accuracy dilemma of recommender systems*, PNAS, 107 (10) 4511-4515.

Summary

In the last decades, complex networks have started to attract the interest of scientists studying complex systems from a variety of application fields. The main reason is likely that complex networks provide a natural and mathematically manageable description of many real complex systems. In particular, they represent a very useful tool to investigate emergent phenomena in complex systems, without invoking strong assumptions on the type of interactions among the elements of the system. In this thesis, we develop multivariate and network methods for the study of complex systems monitored through a detailed recording of data for many and heterogeneous variables, stored in integrated data warehouses. The thesis consists of four essayes.

The first work is a methodological contribution in which we introduce an unbiased pairwise similarity measure between the elements of a bipartite complex network with a double source of heterogeneity. The introduced weighted covariance and correlation coefficients remove the bias observed when using standard metrics, such as the binary Pearson's and Newman's correlation coefficients. The new measures are useful to perform all the tasks that exploit similarities among the elements of a bipartite system, e.g. unsupervised classification problems, recommendation systems etc.

In the second work, we propose a method to investigate the Italian car insurance system, and, in particular, we develop an investigation automatic system, based on Statistically Validated Networks (SVN), aimed at uncovering anomalous subject-accident patterns, which might represent a mark of potential frauds. The tool has been developed within the framework of a project funded by the Italian Institute for the Supervision of Insurance (IVASS) and it is currently operative, for internal use only, at the IVASS to process the integrated database AIA - the Antifraud Integrated Archive - managed by IVASS.

The third work concerns with the empirical analysis of the effects of the so-called Research Excellence Framework (REF) on the scientific productivity of universities in the UK. In this context, we have focused the attention on two Units of Assessments (UOA): Economics and Econometrics, and Business and Management studies. To evaluate the effects due to the REF on both quan-

titative (number of published papers) and qualitative (quality of the journals they published in) outcomes, we analyse the Scopus database, exploiting the information on all of the indexed papers with at least one author affiliated in a university from the UK and/or the US in the time period 2001 and 2015. Although REF2014 has increased the overall number of publications in journals and the number of publications in top-starred journals, the effect stems from an increase in the number of publications in Finance, Business and Management and a decrease in the proportion of Economics and Econometrics publications, steered mainly by universities in the Russell Group that remained in the Economics and Econometrics panel.

Finally, in the fourth work, we integrate SVN with regularized VAR model and Forecast Error Variance Decomposition (FEVD) theory to study spillover effects in finance. Specifically, we focus on the CDS market, with the aim of finding the statistically significant (lagged) interdependencies between CDS spreads of sovereigns and financial institutions from all around the world. Eventually, the application of SVNs allows one to reveal prominent patterns of contagion, where an excess of risk transmission would lead to effects that could undermine the stability of the whole financial system.

Outputs of the PhD research

During the PhD programme I produced four works: **i)** is published, **ii)**, **iii)**, **iv)** are currently under review (whose authors are listed in alphatetical order). They are listed below.

i) Publication: Puccio E, Vassallo P, Piilo J, Tumminello M (2019); *Covariance and Correlation estimators in bipartite complex systems with a double heterogeneity*, Journal of Statistical Mechanics: Theory and Experiments, 053404;

ii) Under review: Cesari R, Consiglio A, Farabullini F, Tumminello M, Vassallo P (2019); *Insurance Fraud Detection: a Statistically Validated Network Approach*;

iii) Under review: Banal-Estanol A, Iori G, Jofre-Bonet M, Maynou L, Tumminello M, Vassallo P (2019); *Research productivity and REF2014: Do REFs produce the desired effects?*

iv) Under review: Bonaccolto G, Consiglio A, Iori G, Tumminello M, Vassallo P (2019); *Regularized Networks and Doom-Loop Effects: Evidence from the CDS Market.*

Author contributions

In publication **i)**, together with the coauthors I reviewed the literature and edited the text. I also contributed to the choice of statistical models to be used. I carried out the empirical analyses, prepared the figures, and built an R package that allows for the computation of Weighted ESTimators of Covariances/Correlations (WestC).

In work **ii)**, together with the coauthors I reviewed the literature, designed the study, and edited the text. I also contributed to the conduction of the empirical analysis, creation of figures and implementation of SAS and R codes.

In work **iii)**, together with the coauthors I edited the text and designed the study. I carried out the empirical analysis and created related figures.

In work **iv)**, together with the coauthors I reviewed the literature, designed the study, carried out the empirical analysis and edited the text.

Acknowledgments

I am so glad that three years ago I decided to enroll in the PhD programme. Such an experience made me get better in the way I think, in the way I solve problems, and it made me become more curious about how things work in the world around us.

I am really grateful to everyone I have met during this challenging and amazing experience. Foremost, I want to express my deepest gratitude to Andrea and Michele for having always guided me along the path. They have been my benchmark. Without them more efforts would have been needed to become what I am now.

I wish to thank Giulia, Christian, Mireia, Laia, Albert, Saqib and all the staff members of the Department of Economics of City University London, for being so kind while I was working there as a visiting researcher. They made me feel at home.

I also thank Riccardo Cesari, Fabio Farabullini, Antonio De Pascalis, Paola Salacotti, and Piero Franchini, managers of the Italian institution for the supervision of insurance companies (IVASS), for making me understand how such an important entity operates in a national and international context. They kindly supported me while I was working at IVASS, in Rome, as an intern.

Many thanks also to Prof. Juho Kanniainen from Tampere University and Dr. Vincenzo Nicosia from Queen Mary University London for their useful comments that helped improving the quality of this thesis.

I am also grateful to Erasmo, Valerio, Massimo, and Mariangela, academics working at the department of Economics, Business and Statistics of the university of Palermo, for giving me extra support and perspectives on the challenges of academia.

I'm grateful to my colleagues and friends for the constant source of fun and relief along the way, making the journey of PhD more pleasurable, in particular, I thank my officemates Giorgio, Cinzia, and Vincenzo for all the delightful conversations we had.

Finally, I thank my family for the unconditional support. I will always be in debt to them for the education I was given, based on values such as respect of one another, commitment, hard work and constant self development.