POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell'Informazione

Dipartimento di Matematica

Master of Science in

Mathematical Engineering

POLITECNICO

MILANO 1863

Non-negative Matrix Factorization and Compositional Clustering of National Input-Output Tables

Supervisor:

PROF. SIMONE VANTINI

Assistant Supervisor:

PROF. MARIKA ARENA

PROF. ALESSANDRA MENAFOGLIO

Master Graduation Thesis by:

ANDREA MASCARETTI

Matr. 864671

Academic Year 2017-2018

"Quer o destino que eu não creia no destino
E o meu fado é nem ter fado nenhum
Cantá-lo bem sem sequer o ter sentido
Senti-lo como ninguém, mas não ter sentido algum"
– Desfado, Ana Moura

ACKNOWLEDGMENTS

Uno spazio pubblico non è mai solamente l'insieme delle strutture e degli arredi che lo compongono. O, almeno, non è solo questo. Un'università non trova la sua caratterizzazione nel solo fatto di essere abitata. La sua essenza risulta dalla somma delle persone che la vivono e dal fatto che siano quelle persone e non altre. È la somma delle speranze, delle capacità e delle storie di una moltitudine che si trova a convivere nello stesso luogo. La mia esperienza è l'esperienza del mio Politecnico, vissuto attraverso i miei occhi, i miei umori e le mie scelte. E come i limiti dello spazio costringono i comportamenti collettivi, così i miei hanno costretto quello che ho visto. Potremmo dire che un edificio è riuscito quando i principi che lo hanno ispirato sono diventati invisibili, inglobati nel calcestruzzo e nel vetro, e intuibili solamente attraverso l'esperienza ripetuta di quel posto, con l'abitudine. Così è stato per il mio percorso di formazione. Alcune delle cose che non comprendevo le intuisco soltanto ora e, mentre la mia capacità di osservare si riscopre più fine, capisco che proprio questo era l'obiettivo di un periodo che ora volge al termine, anche se non sempre sono stato in grado di vederlo. Da solo non sarei riuscito a fare molto.

Ringrazio pertanto il Prof. Secchi, perché ha intuito che cosa mi avrebbe appassionato. Ringrazio il mio gruppo di ricerca perché mi ha supportato, guidato e ascoltato: i Proff. Vantini, Menafoglio, Arena, Dell'Agostino e Flori.

Ringrazio la mia famiglia, mia madre, mia sorella e mio padre, perché sono il mio rifugio e perché mi hanno insegnato che cosa significhi volersi bene.

Ringrazio Marti, perché vibriamo all'unisono.

Ringrazio Phelps, perché mi ha mostrato quanto sia importante trovare la propria dimensione eroica (sperando che un giorno lo capirà anche lui).

Ringrazio il compagno Piazza, perché è grazie a lui se le parole sono rimaste un fine.

Ringrazio Iuri, perché anche per me la fiamma è rimasta vive grazie a lui.

Ringrazio Natalia, perché ci siamo finalmente ritrovati e abbiamo così tanto in comune.

Ringrazio il Defe, perché la sua curiosità ha salvato anche la mia.

Ringrazio le tante e i tanti che forse si stupirebbero sapendo quanto sono stati importanti per me: il signore romano che ci ha ospitati a casa sua a Roma mentre affrontavo un provino teatrale, Tiziana e le sue regie teatrali, Mattia e la sua musica, Conte e Khal perché ci sono stati sempre, Margherita per la sua passione per i tomini, il gruppo del Mind, Mezza per queste ultime settimane di patio scandite da musica orrenda e personaggi dubbi, le tante persone con le quali ho condiviso la passione per la politica, così importante e così trascurata.

Agradeço a Lisboa, às suas luzes e cores, à Faculdade de Ciências e Tecnologia, à Sara, pelos nossos dias juntos. Vou guardar para sempre dentro de mim a música das festas, o cheiro do oceano, os deslizes quando chove (ah, a calçada portuguesa), os bitoques de Dona Teresa, as vielas de Alcântara, a alegria e a bondade de um lugar que me acolheu e ao qual agora pertenço.

Agradezco a María José, porque su pureza, su fuerza y su alegría fueron un ejemplo para mí, mientras nos convertíamos en poetas que regresaban a casa.

Ringrazio Anna, perché non si arrabbiava quando la svegliavo tornando tardi, perché mi aspettava sempre per conversare e per il modo in cui fa le cose.

Ringrazio Arianna, perché è cominciato tutto con lei e perché il Portogallo non sarebbe stata la stessa cosa.

I thank Iliana because she is so sweet.

I thank Klara, Lucka, Martin and Matti because we are a group, possibly one of the best.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACRONYMS

UN	United Nations
EU	European Union
Eurostat	European Statistical Office
WIOD	World Input Output Database
ISTAT	Istituto Nazionale di Statistica
OECD	Organisation for Economic Co-operation and Development

ABSTRACT

In this work we present a novel methodological framework aimed at clustering cross-sectional datasets of Input-Output Tables. Input-Output Tables collect monetary transactions amongst the various sectors of an economic system over a period of time, usually a country considered for a year. To this aim, we apply Non-negative Matrix Factorisation (NMF) to reduce the dimensionality of the dataset. A small number of fundamental "Archetypes" is identified and countries are characterised as the expression of such Archetypes. NMF allows us to decouple the clustering problem into two different parts. The interpretation of the clusters is based on the interpretation of the Archetypes. We consider each Archetype as an economic network and interpret it on the basis of a Random Walk Centrality measure inducing a ranking of its nodes. Such centrality is defined based on the Mean First Passage Time (MFPT) computed considering the propagation of a small and indivisible supply-side shock hitting another sector. Clustering is computed considering the economies as the expression of the different Archetypes. We implement a k-medoid algorithm and compare two different metrics: the well-known Euclidean metric and the Aitchison metric. The Aitchison metric belongs to the framework of compositional data analysis, a branch of statistics whose focus is the analysis of the parts constituting the datapoint notwithstanding the total. This allows for a characterisation of countries considering the mix that originates them. The analysis is conducted on the OECD Input-Output Table database.

SOMMARIO

Il presente lavoro di tesi definisce una metodologia innovativa finalizzata al clustering di dataset cross-sectional di tavole Input-Output. Una tavola Input-Output offre una rappresentazione schematica di un sistema economico, registrando le transazioni tra vari settori di un'economia lungo un arco di tempo, mediante flussi monetari. Il problema che ci proponiamo di risolvere presenta svariati fattori di complessità. Anzitutto, per rendere le tavole raffrontabili, gli istituti che si occupano della loro pubblicazione armonizzano i dati a scapito di una significativa perdita di dettaglio rispetto alle misurazioni nazionali. Una seconda problematica è costituita dalla difficoltà di interpretazione di un cluster, una volta che questo sia stato individuato. Ogni nazione, infatti, è il frutto di un processo storico unico e particolare, del quale le tavole Input-Output provano a catturare l'aspetto economico. Questo rende l'interpretazione del clustering particolarmente difficile, dal momento che non ci è possibile validare i nostri risultati con una misura della loro bontà diversa dall'omogeneità interna e dal rapporto tra questi e le nostre aspettative.

La nostra analisi propone una soluzione alle problematiche evidenziate mediante una decomposizione del problema in due parti. Attraverso una tecnica di riduzione dimensionale nota come Non-negative Matrix Factorisation, ipotizziamo che sia possibile individuare un numero ridotto e significativo di "archetipi". I vari paesi che analizziamo vengono quindi considerati come espressione di diversi livelli di attivazione di questi archetipi.

L'interpretazione degli archetipi avviene considerando ciascuno di questi come un network economico. Ogni network viene caratterizzato in relazione alle sue modalità di reazione a uno shock economico lato domanda, considerando come centrali quei settori raggiunti più velocemente.

Il clustering avviene quindi considerando i paesi in questo nuovo sistema di coordinate. A questo fine, vengono considerate due metriche diverse. La prima metrica è la ben nota distanza euclidea: in questo caso ci aspettiamo che due paesi appartenenti allo stesso gruppo mostrino attivazioni di intensità simile per tutte le componenti. La seconda metrica che utilizziamo è la metrica di Aitchison. Tale metrica appartiene all'insieme di tecniche statistiche afferenti al campo della statistica composizionale. In questo caso, la distanza viene calcolata considerando i rapporti tra le varie attivazioni e prescindendo dai valori complessivi di queste ultime. L'ipotesi che sottende questo metodo è che l'informazione non sia contenuta nei dati in quanto tali, ma nelle proporzioni tra le parti che li compongono. Il clustering viene effettuato mediante un algoritmo k-medoids, in modo da trovare, per

ogni cluster, un prototipo che ne rappresenti i membri.

L'analisi dei dati è stata condotta sul dataset di tavole Input-Output elaborato dall'OCSE (Organizzazione per la cooperazione e lo sviluppo economico), nella sua terza revisione.

1 INTRODUCTION

The structural analysis of the differences between countries plays a fundamental role in economic analysis. In this work, we aim at finding groups of countries that share similarities amongst each other. This particular problem is known in statistics and machine learning with the name of *clustering*.

The main idea behind clustering is that of finding a particular representation of the data at hand, in order to extract some meaning that can be useful in the domain under analysis.

We will consider datasets containing monetary transactions between different economic sectors to describe the system under scrutiny, collected yearly and for each country. Such data carry the name of *Input-Output Tables*. In particular, we will consider the data published by the OECD.

The problem is intrinsically complex for three different orders of reasons. First of all, the data we will analyse are the result of the aggregation and harmonisation of different datasets prepared by national statistical offices. This means that we will be dealing with data that underwent significant processing and are the result of statistical procedures that originally might have differed for their precision, design priorities and methodologies. Even though the existence of an international and homogenous dataset, comprising a vast number of countries, allows for direct analyses of different economic realities, a major drawback is given by the fact that such uniformity comes at the expense of detail. Input-output tables are as significant as the amount of detail they convey, as different countries can be better characterised when considering the subtle nuances that distinguish them. This is especially true when the analysis is led on groups of states that share common characteristics, such as, for instance, the members of the EU. A second issue is given by the fact that there exists no intuitive or straightforward way to interpret a cluster of countries, as each member of the cluster is described by an entire Input-Output Table, whose dimensions render direct comparisons unfeasible. The information contained in such tables cannot be easily summarised by means of some synthetic indicator, *f.i.* the column-wise mean of a table would be inadequate because it would imply loosing all the details on the specific relationships between different sectors. Moreover, as the aim of this work is to find groups of countries with common characteristics based on pure quantitative information, it can be rather difficult to assess the quality of the groupings found without considering the particular and unique historical, social and economic peculiarities that each country presents. In this sense, the work hereby presented should be considered as a novel instrument of research, whose aim is that of providing a tool to corroborate existing knowledge or to defy it. The problem we are considering is complex and there exists no easy nor straightforward solution to it. The third issue is then a direct consequence: we need to develop a framework, however simple, to extract information from the data in order to validate the results of the analysis and to grasp its limits. We need to develop a narrative to understand what makes a given country different from another one, balancing the trade-off between a too narrow approach, which would result in an outcome of little or no interest, and the temptation to be as general as possible, risking of losing any interpretative key.

For all of the above problems, a solution is here presented. To deal with the complexity of the data, we will limit the scope of the analysis to cross-sectional datasets of geographical areas of increasing complexity. We will abstain from making any sort of consideration regarding the dynamics of the system. We will consider the datasets, cut according to a specified year, as closed and isolated systems.

To assess the differences between countries we will interpret them as economic networks, in which each sector will be considered as a node and each connection between such sectors as an edge. Such systems will be described by analysing the different reaction they yield when hit by a supply-side shock. We will try to identify whether there exist some sectors that are central to the economic system under scrutiny by measuring the rapidity at which they are reached by a shock emerging from any other given sector.

What is more, given the high complexity of the dataset, we will perform a dimensionality reduction by means of Non-negative Matrix Factorisation (NMF). The main assumption we will make is that there exists a fixed and small number of fundamental economic structures, which we shall name “archetypes". We expect each country to be an expression of those archetypes. To put it in other words, we will try to assess if it is possible to individuate a fixed number of basic ingredients on one hand and a list of different recipes on the other. Should this approach succeed, it would provide us with an extremely flexible methodological framework, as it allows to decouple the problem into two parts and therefore customise the analysis according to one’s interests. In this work, as previously stated, such archetypes will be characterised in view of their response to small supply-side shocks, and their central sectors will be ranked accordingly. Clustering will then be conducted considering how different countries vary in this new set of coordinates and looking for groups of countries that are somehow “close by". As for the definition of closeness, two main approaches will be discussed. The first one will employ the standard Euclidean distance: two points will be regarded as close to each other if the square difference of their components is not too big. This means considering the activation of various archetypes taking into account their magnitudes. The second approach, the so-called compositional approach, will instead consider a different definition of distance. In this second case, we will consider two countries to be close if they are the result of two mixes that are not too different in their relative ratios. In other words, we will assume that it is in the particular composition of the ingredients that lies the key to assess diversity.

To the best of our knowledge, the approach here presented is original, despite elaborating on both well-established and modern tools of economic analysis. Input-output tables are considered a classical tool of economic analysis and have been widely employed since their introduction in the late 1930s ([50]). However, the contamination between such classical framework and more recent developments in statistics is recent. Non-negative Matrix Factorisation has been applied to Input-Output tables in ([75]) and ([44]), considering China and Japan respectively. To the best of our knowledge, the application of NMF to a datasets comprising more countries is novel. The interaction between input-output tables and network analysis has been discussed and analysed in various works and notably in ([10]), [59]), ([88]) and ([72]). As far as interactions between compositional data analysis and input-output tables are concerned, in ([87]) a model to extend time series forecasting to input-output tables is developed and some assumptions on the entries of the tables are made comparing them to the hypotheses underlying compositional data analysis. Yet the tentative direct application of the compositional framework to a problem of input-output tables appears to be original.

This work is organised as follows: in Chapter 2 we introduce input-output tables, together with the underlying microeconomic foundations. A theoretical model to justify the renewed interest in input-output tables is also presented, followed by the definition of a random centrality measure to assess the impact of shocks. In Chapter 3, we introduce NMF and the main aspects and limitations of this technique. In Chapter 4, the elements of compositional data analysis are presented, as well as the clustering algorithm that will be employed in the analysis. In Chapter 5 we present the datasets and the results of our analysis. Finally, in Chapter 6, we discuss possible further developments for this work.

2 INPUT/OUTPUT TABLES

2.0.1 *Profit, technology and firms*

This section aims at briefly introducing the basics of microeconomic theory. The material in this section is based on ([83]). Another exhaustive reference is ([58]).

Firms are egoistic and rational agents that possess the technological know-how required to transform some combinations of inputs into outputs under certain conditions of cost, time, location and availability of resources. Suppose, for instance, that we are in an economy with n possible goods to serve as input and output. We denote with x_j^i the input of good j, with j that spans from 1 to n, and with x_j^o the output of the same good. The **net output** for the j-th good is the difference between the two quantities, *i.e.* $x_j^o - x_j^i$. When the net output is positive, we have a net producer, otherwise a net consumer. It is straightforward to define a **production plan** as a vector $x \in \mathbb{R}^n$ of feasible net outputs. The **production possibilities** can be expressed through $X \subseteq \mathbb{R}^n$, the set of all such vectors. It might be that due to certain reasons some production plans are not "immediately" available, but will eventually be. To take this possibility into account, we define another vector $z \in \mathbb{R}^n$ of costraints on the usage of some goods and therefore have $X(z) \subseteq X$.

From now on, we will simplify this framework by hypothesising that firms only produce as their output a single good. This allows us to rewrite a production plan as the juxtaposition of outputs and inputs, *i.e.* $(y, -x)$ where we have that $x \in \mathbb{R}^{n-1}$ (we change the sign notation, denoting the input vector with a negative sign and the output with a positive one).

Definition 2.1 (Input Requirement Set)**.** The **input requirement set** is defined as

$$V(y) = \left\{x \in \mathbb{R}_+^{n-1} \text{ s.t. } (y, -x) \in X\right\}.$$

We will assume that $V(y)$ is a *closed, nonempty set* for all $y \geqslant 0$.

For the sake of completeness, we also define an **isoquant**.

Definition 2.2 (Isoquant.)**.** An **isoquant** is defined as

$$Q(y) = \left\{x \in \mathbb{R}_+^{n-1} \text{ s.t. } x \in V(y) \text{ and } \left(\forall y' > y,\ x \notin V(y')\right)\right\}.$$

Notice that Definition 2.1 gives an account of all the possible bundles that lead to a production of *at least* y, whereas from Definition 2.2 we obtain the set of bundles required to produce *exactly* y.

Now, suppose we fix x to some value. We assume that there is no **technological waste**, that is firms produce the maximum quantity of output given some input. This leads to the following definition.

Definition 2.3 (Production function.)**.** We define as **production function** the function that associates to every level of input the maximum output.

$$f : x \mapsto \max_{y}\{(y, -x) \text{ s.t. } (y, -x) \in X\}$$

The extension to the short-run case is immediate. We now introduce the two production functions we will use throughout this work. For the sake of simplicity, in the rest of the section we will assume such functions to only take two inputs. Notice that this reduces z to a vector $z = (z_1, z_2)$.

Example 2.1. *Cobb-Douglas technology.*
Let α be a parameter such that $0 \leqslant \alpha \leqslant 1$, then we have:

- $X = \left\{(y, -x_1, -x_2) \in \mathbb{R}^3 \text{ s.t } y \leqslant x_1^\alpha x_2^{1-\alpha}\right\}$
- $X(z) = \left\{(y, -x_1, -x_2) \in \mathbb{R}^3 \text{ s.t. } \left(y \leqslant x_1^\alpha x_2^{1-\alpha};\ x_i \leqslant z_i,\ i = 1, 2\right)\right\}$
- $f(x_1, x_2) = x_1^\alpha x_2^{1-\alpha}$
- $V(y) = \left\{(x_1, x_2) \in \mathbb{R}_+^2 \text{ s.t. } y \leqslant x_1^\alpha x_2^{1-\alpha}\right\}$
- $Q(y) = \left\{(x_1, x_2) \in \mathbb{R}_+^2 \text{ s.t. } y = x_1^\alpha x_2^{1-\alpha}\right\}$

Example 2.2. *Leontief technology. Let $\alpha > 0$ and $\beta > 0$ be two parameters. The Leontief technology is defined as*

- $X = \left\{(y, -x_1, -x_2) \in \mathbb{R}^3 \text{ s.t. } y \leqslant \min\{\alpha x_1,\ \beta x_2\}\right\}$
- $X(z) = \left\{(x_1, x_2) \in \mathbb{R}_+^2 \text{ s.t. } (y \leqslant \min\{\alpha x_1,\ \beta x_2\};\ x_i \leqslant z_i,\ i = 1, 2)\right\}$
- $f(x_1, x_2) = \min\{\alpha x_1,\ \beta x_2\}$
- $V(y) = \left\{(x_1, x_2) \in \mathbb{R}_+^2 \text{ s.t. } y \leqslant \min\{\alpha x_1,\ \beta x_2\}\right\}$
- $Q(y) = \left\{(x_1, x_2) \in \mathbb{R}_+^2 \text{ s.t. } y = \min\{\alpha x_1,\ \beta x_2\}\right\}$

We will only consider "smooth" functions (and usually with two input factors) in dealing with the problem, that is functions that are at least twice differentiable with continuous derivatives.

This leads to another important definition.

Definition 2.4 (Marginal product of factors)**.** Let $f : \Omega \subseteq \mathbb{R}^2 \to \mathbb{R}$ be a production function such that $f \in C^2(\Omega)$. The **marginal product** of the factor x_i, $i = 1, 2$, is then defined as

$$MR_i = \frac{\partial f(x_1, x_2)}{\partial x_i}.$$

We assume that, for $i = 1, 2$,

$$MR_i = \frac{\partial f(x_1, x_2)}{\partial x_i} \geqslant 0$$

and

$$\frac{\partial MR_i}{\partial x_i} = \frac{\partial^2 f(x_1, x_2)}{\partial x_i^2} \leqslant 0.$$

This means that augmenting the exploitation of a factor always yields to some increase in output, but that this increase slowly diminishes. This is called in economics the *law of diminishing marginal returns*.

2.0.1.1 *The technical rate of substition*

Let f be a production function of the kind considered above and y^* be some feasible output level such that we have $y^* = f(x_1^*, x_2^*) = f(x^*)$. We want to understand if it is possible to increase the amount of input 1 and decrease the amount of input 2 without variations in the output, i.e. we want to determine the **technical rate of substitution** between the factors. We can apply the implicit function theorem. We have that, at least in some neighbourhood $\mathcal{U}(x_1^*)$,

$$f(x_1, x_2(x_1)) \equiv y$$

so that, by differentiating,

$$\frac{\partial f(x^*)}{\partial x_1} + \frac{\partial f(x^*)}{\partial x_2}\frac{dx_2}{dx_1} = 0$$

and therefore

$$\frac{dx_2}{dx_1} = -\frac{\frac{\partial f(x^*)}{\partial x_1}}{\frac{\partial f(x^*)}{\partial x_2}}$$

We can, in a similar fashion, find the other rate of substitution.

2.0.1.2 *Returns to scale*

Suppose we produce some output given some fixed input. We are interested in finding out what happens by using t times the initial output. We have three possible cases. If $f(tx) = tf(x)$, *i.e.* if the function is *homogeneous of degree 1*, we have **constant return to scale.** Of course, we have **increasing return to scale** whenever $f(tx) > tf(x)$ and **decreasing return to scale** when $f(tx) < tf(x)$.

It might be that a technology exhibits increasing returns of scale for a while and then start presenting decreasing return to scales: we neglect those cases and just limit ourselves to production functions that fall into one of those three categories.

2.0.1.3 *Profit Maximisation*

Economic *profit* is defined as the difference between revenues and costs. All costs should be considered. Firms, being rational and egoistic agents, have to decide which policy will allow them to maximise their profit. In doing so, they face two order of constraints:

- *Technological constraints* are those constraints due to the technological feasibility of plans.
- *Market constraints* concern the effect of the behaviour of other agents on that of the firm.

In our discussion, we shall consider firms as *price-takers*. Firms will consider prices to be an exogenous, given variable. This is a direct consequence of the assumption of perfect competition we now make.

Definition 2.5 (Profit function)**.** Let $p = (p_1, p_2, \ldots, p_n)$ be a price vector of non-negative elements. The *profit function* is defined as

$$\pi(p) = \max_{x \in X} p^t x, \tag{2.1}$$

In our case, firms only produce single-good outputs. We can therefore reformulate (2.1) as

$$\pi(p, w) = \max pf(x) - w^t x, \tag{2.2}$$

where $x = (x_1, \ldots, x_n)$ is the vector of input goods, $w = (w_1, \ldots, w_n)$ is the vector of costs of factors and p reduces to the price of the output good.

The optimum of this function is found at x^* satisfying

$$p \nabla f(x^*) = w. \tag{2.3}$$

From (2.3) we see that first-order conditions imply that at the optimum the price of factors has to be equal to its marginal product value.

2.1 CONSUMER BEHAVIOUR

We consider consumers facing possible different choices of bundles, belonging to some set X. We usually take X to lie in the positive orthant of $\mathbb{R}^k$. Consumers are endowed with a preference relation $(X, \succeq)$, where $x \succeq y$ means that bundle x is at least as good as y in the eyes of the consumer.

We assume four properties ([2]) regarding the preference relation.

- *Completeness.* For all $x \in X$ and $y \in X$, either $x \succeq y$, $y \succeq x$ or both.

- *Reflexivity.* For all $x \in X$, it holds $x \succeq x$.
- *Transitivity.* For all $x, y, z \in X$, if $x \succeq y$ and $y \succeq z$, then $x \succeq z$.
- *Continuity.* For all $y \in X$, the sets $\{x : x \succeq y\}$ and $\{x : x \preceq y\}$ are closed sets. Moreoever, $\{x : x \succ y\}$ and $\{x : x \prec y\}$ are open sets.

Such properties allow to summarise the preference relation by means of a *utility function*: a function $u : X \to \mathbb{R}$ with the property that $x \succeq y$ if and only if $u(x) \geqslant u(y)$. The properties also imply that the utility function is continuous ([84]).

It is possible to define a *marginal rate of substitution* following what has already been done regarding production function.

2.1.0.1 *The Consumer Problem*

In this work, we shall assume that consumers are rational agents endowed with a utility function over a set of bundles. It is a natural consequence, thus, that they will find the bundle maximising their utility, *i.e.* their optimal bundle.

Let m be the monetary budget of the consumer and let $p = (p_1, \dots, p_k)$ be a vector of prices. The problem can be formulated as

$$\max u(x), \tag{2.4}$$

such that $p^t x \leqslant m$ and $x \in X$.

The problem can be solved applying the well-known theory of Langrage multipliers and namely by solving

$$\mathcal{L} = u(x) - \lambda(p^t x - m). \tag{2.5}$$

It is of interest to notice that at the optimum we have

$$\frac{\frac{\partial u(x^*)}{\partial x_i}}{\frac{\partial u(x^*)}{\partial x_j}} = \frac{p_i}{p_j}.$$

In this simple framework, the problem was of easy solution because of the various hypotheses we have made. For a thorough treatment of utility functions and consumer behaviour, we refer the reader to ([83]).

2.2 INPUT-OUTPUT ANALYSIS

We now briefly introduce input-output analysis, relying heavily on ([61]), ([74]) and ([3]). Two other important works are those by ([63]) and ([41]). Input-output analysis (or *interindustry analysis*, as it is also called) is an analytical framework developed by the economist W.

Leontief ([50, 51]) in the late 1930s and has since played a fundamental role in economic analysis. The model is widely employed today: the United Nations (UN) has published a handbook on input-output analysis[1] and the European Statistical Office (Eurostat) is working on new methodologies[2] to collect and use data. The idea of a fine-grained accounting of interindustry activities, however, dates back way before Leontief. Input-output tables can be considered as a formalisation of a concept developed in 1758 by the French economist François Quesnay: the tableau économique ([71]). Quesnay's work received mixed reviews, ranging from the enthusiasm of Marx ([57]) to the disdain of Gray ([35]). A fundamental contribution to the development of input-output analysis came from the work of Léon Walras, who applied ideas stemming from Newtonian physics to develop a theory of economic equilibrium. In ([85]), Walras introduced the idea of a set of coefficients to relate input factors and output: an idea not too dissimilar from the technical coefficients of input-output analysis we shall introduce.

The fundamental characteristic of industrial economies is the rapid and continuous flux of goods and means of payments (money) amongst the various actors constituting the system. As opposed to economic systems of self-consumption, in which markets only serve the scope of allocating surpluses, in industrial economies production is not decided on the basis of a direct need, but rather upon the expectation of a future demand. Goods, or commodities, are produced to be sold, so that specialised sectors do not usually consume the entire output of their own production. A consequence of this organisation of production has been the separation of those deciding what to produce from those offering their workforce to produce it. An industrial economy can thus be represented as a circular system, at least at this level of approximation: on one side we have workers, willing to offer their labour, on the other one we have entrepreneurs, who need workforce to produce the various commodities as it is depicted in Figure 2.1.

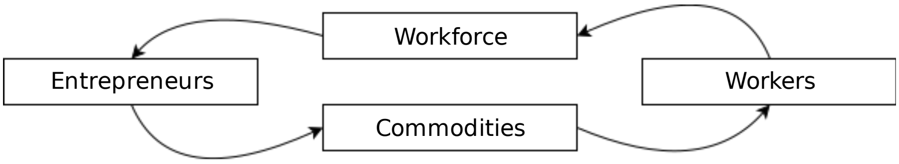

Figure 2.1: Physical flow between workers and entrepreneurs.

Parallel to this physical flow, there exists a monetary one. Workers buy commodities by paying them and receive wages in exchange for their workforce. This second flow is reported in Figure 2.2.

This basic scheme can be further sophisticated, for instance by considering that a part of the flow of commodities is absorbed by

[1] https://unstats.un.org/unsd/nationalaccount/docs/SUT_IOT_HB_wc.pdf
[2] https://ec.europa.eu/eurostat/web/experimental-statistics/figaro

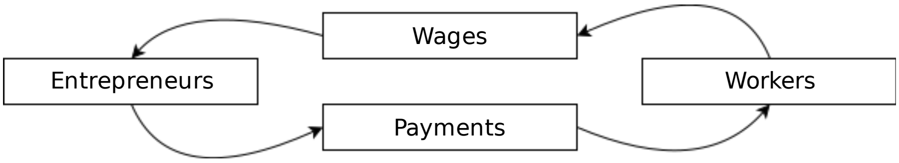

Figure 2.2: Financial flow between workers and entrepreneurs.

entrepreneurs, in order to satisfy their own demand and to invest in new means of production. The purchasing power of entrepreneurs is given by the fact that commodities are sold at a price which is higher than that of production: they profit from the difference between revenues and costs. Another way to sophisticate the model is to consider different sectors operating within the system, grouping them according to some classification scheme. This implies considering pairwise connections between the sectors, namely the *interindustry flows*. It is important to notice that such flows require the identification of an economic area (usually a country) and a given period of time (usually a year) to be well-defined.

2.2.1 *An example*

To introduce input-output tables, we consider an economy with only two sectors: agriculture and manufacturing. We can create a table of the requirements for the output of the agricultural sector as we did in Table 2.1. On the rows of Table 2.1 we have the sectors of the economy: reading them across the columns we know the destination of the output of that particular sector (whether it is agriculture, manufacturing or labour) to a another sector of the economy.

	Agriculture	Manufacturing
Agriculture (kg)	q_{11}	q_{21}
Manufacturing (kg)	q_{21}	q_{22}
Labour (hrs)	L_1	L_2

Table 2.1: Requirements table for a two sector economy in physical terms.

Of course, for a system to reproduce over time, we need to have that the total amount of commodities produced by each sector should not be inferior to that required by the various sectors. We can therefore consider Table 2.2.

As we pointed out before, we can build the financial version of Table 2.2, considering the monetary exchanges occurred between the various agents of the economic system. We report this in Table 2.3. We see that the column-wise and the row-wise sums, once a sector has been fixed, give the same results: this is because we have added a profit row. If entrepreneurs sell the total amount of good they

	Agriculture	Manufacturing	Final Demand	Production
Agriculture	q_{11}	q_{12}	y_1	q_1
Manufacturing	q_{21}	q_{22}	y_2	q_2
Labour	L_1	L_2		
Final Production	q_1	q_2		

Table 2.2: Input-Output transaction table in physical terms.

produce to intermediate sectors and to the final market, they profit from an amount equal to the difference between what they have paid to produce and the total revenue.

	Agriculture	Manufacturing	Final Demand	Production
Agriculture	z_{11}	z_{12}	f_1	x_1
Manufacturing	z_{21}	z_{22}	f_2	x_2
Wages	l_1	l_2		
Profit	n_1	n_2		
Final Production	x_1	x_2		

Table 2.3: Input-Output transaction table in financial terms.

2.2.2 *Input-Output Table*

We are now ready to give a more general and formal definition of input-output tables. To do so, we will rely heavily on both ([74]) and ([60]). An input-output model is constructed starting from a well-defined geographic area and given period of time. The usual choice is a pair nation-year. The economy under analysis has to be divided into a number of sectors. It is possible to set the level of detail: we can have industries in the usual sense, such as “wood", at a finer level of detail (“surfboards") or at a lager one (“manufacturing"). The model is constructed registering the flow in monetary terms between different sectors during the fixed amount of time. Monetary records are usually kept because of the difficulties arising when keeping physical ones, due to the single-output approximation we make for every sector. Let $I = \{1, 2, \dots, n\}$ be the set of indices associated with each sector. We define z_{ij} as the monetary value of the transactions occurred between sector i and sector j. If we let f_i be the final demand for sector i and x_i the total quantity of commodity i we have produced, we obtain that for every $i \in I$,

$$x_i = \sum_{j=1}^{n} z_{ij} + f_i. \tag{2.6}$$

If we set $x = (x_1, \ldots, x_n)$, $f = (f_1, \ldots, f_n)$, $\mathbf{Z} = \left(z_{ij}\right)_{i,j \in I}$ and we indicate by $\mathbf{1}$ the vector of length n composed by 1s, we can rewrite in a more compact form the previous equation as

$$x = \mathbf{Z1} + f \tag{2.7}$$

The fundamental assumption we make in input-output analysis is that, for every $j \in I$,

$$z_{ij} = \phi(x_j), \ \forall i \in I. \tag{2.8}$$

This means that the total flow from sector i to sector j is a function that takes as an argument the output of the destination sector. As for the explicit form of this relationship, we define the following.

Definition 2.6 (Technical Coefficient)**.** We define as technical coefficient the ratio

$$a_{ij} = \frac{z_{ij}}{x_j}, \ i, j \in I. \tag{2.9}$$

We therefore have

$$z_{ij} = a_{ij} x_j, \tag{2.10}$$

Notice that because of (2.10) we see that we have *constant return to scale*: the proportion between output and amount of input is fixed.

Another hypothesis we make is that inputs are absorbed with constant proportions. Let $I(j) = I \setminus \left\{i : a_{ij} = 0\right\}$. Mathematically, this means

$$x_j = \min_{i \in I(j)} \left\{ \frac{z_{ij}}{a_{ij}} \right\}, \tag{2.11}$$

that is to say we have a Leontief production function.

Definition 2.7 (Matrix of Technical Coefficient)**.** Let $\mathbf{A} = (a_{ij})_{i,j \in I}$. We call $\mathbf{A}$ the $(n \times n)$ matrix of technical coefficients.

We are now able to rewrite (2.7) as

$$x = \mathbf{A}x + f. \tag{2.12}$$

Rewriting, we obtain

$$(\mathbf{I} - \mathbf{A})\, x = f$$

and if $(\mathbf{I} - \mathbf{A})$ is invertible,

$$x = (\mathbf{I} - \mathbf{A})^{-1} f. \tag{2.13}$$

Definition 2.8 (Leontief inverse)**.** Let $\mathbf{A}$ be the technical coefficients matrix. If $(\mathbf{I}-\mathbf{A})$ is invertible, we define $\mathbf{L} = (\mathbf{I}-\mathbf{A})^{-1}$ to be the Leontief Inverse.

From (2.13), we see that the Leontief inverse is the matrix relating the input required by the economic system in order to produce some given final demand vector. Regarding the existence of a Leontief inverse, in ([39]) conditions are derived regarding the positivity of all principal minors of $\mathbf{I}-\mathbf{A}$. Weaker conditions have also been derived in ([30]). Mathematically, the problem is known as the inversion of M-matrices ([62]).

2.3 THEORETICAL MOTIVATION

Input-output tables have also recently found new applications in the theory of production networks. This is due to two main recent developments. On one hand, the field of network and complex network analysis has led to the introduction of new conceptual frameworks in economic analysis, especially as far as shock propagation is concerned ([16, 64]). Moreover, the increasing accuracy of databases has greatly broadened the ability of economists and statisticians to verify their theories and to validate models at a greater level of accuracy.

2.3.1 *Modelling production networks*

To provide the analysis with some theoretical motivation, we present here a simple production network model from ([17]) and ([18]). We consider an economy of n industries, denoted by $\{1, 2, \dots, n\}$. Each industry is assumed to produce a different output. Commodities produced by industries, or sectors, can be either consumed by households or can be used as an intermediate good by some other sector. Firms employ a Cobb-Douglas production function (which we will be using for its tractability), with constant return to scale, to transform both the intermediate good and labour into their output product. In particular, considering industry i, we obtain

$$y_i = z_i \zeta_i l_i^{\alpha_i} \prod_{j=1}^{n} x_{ij}^{a_{ij}}, \tag{2.14}$$

where l_i is the labour requirement, x_{ij} is the amount of commodity j used to produce i, α_i is the share of labour employed by industry i , z_i is the Hicks-neutral productivity shock and $\zeta_i = \alpha_i^{-\alpha_i} \prod_{j=1}^{n} a_{ij}^{-a_{ij}}$ is a normalisation constant, depending only on the parameters of the model.

Notice that, for each $i \in \{1, 2, \dots, n\}$, the exponents a_{ij} in (2.14) quantify if and how much sector i requires from sector j. The exponent is zero if good j is not used for the production of good i. In general, it needs not hold $a_{ij} = a_{ji}$. We also notice that a_{ii} measures the importance of good i for the production of itself. The hypothesis of constant returns to scale implies that $\alpha_i + \sum_{j=1}^{n} a_{ij} = 1$.

In addition to the firms, we populate our model with households and select a representative household. We imagine that this representative household will supply inelastically one unit of labour and will be endowed with a utility function

$$u(c_1, c_2, \dots, c_n) = \sum_{i=1}^{n} \beta_i \log\left(\frac{c_i}{\beta_i}\right), \tag{2.15}$$

where c_i is the amount of good i consumed and $\beta_i \geqslant 0$, $i = 1, \dots, n$, measure the shares of the various good within the utility function of the household, constrained such that $\sum_{i=1}^{n} \beta_i = 1$.

Equations (2.14) and (2.15) fully specify the environment.

The competitive equilibrium of the economy is defined as a collection of prices and quantities such that

1. the representative household maximises her utility
2. the representative firm for each sector maximises its profit
3. market clears (*i.e.* no goods are left unsold)

First of all, we need to define some key concepts that will play a role in our subsequent analysis. Given the fact that we have decided to adopt Cobb-Douglas technologies and preferences, we can consider matrix $\mathbf{A} = \{a_{ij}\}$ as a summary of all the reciprocal linkages between industries. This means that coupling $\mathbf{A}$ with a fully specified vector of shocks $z = (z_1, \dots, z_n)$ we obtain a sufficient statistic regarding the production side of the economy insofar described.

We also notice that $\mathbf{A}$ can be naturally considered as an adjacency matrix, inducing a graph that can be considered as a production network.

We also define *Domar weights* as the fraction of sales of a industry i as a fraction of the GDP.

Definition 2.9. Domar weights.
$\lambda_i = \frac{p_i y_i}{\text{GDP}}$, where p_i is the price of output i and y_i is the output of industry i.

We can now proceed and compute the equilibrium. Firms will maximise their profits. That is to say, every firm i will seek to maximise the function $\pi_i = p_i y_i - w l_i - \sum_{j=1}^{n} p_j x_{ij}$, while taking the prices $p = (p_1, \dots, p_n)$ and wages w as given.

First order conditions for firm i are given by $x_{ij} = \frac{a_{ij} p_i y_i}{p_j}$ and $l_i = \frac{\alpha_i p_i y_i}{w}$. We can insert these results into Equation (2.14) and take the logarithm to obtain

$$\log\left(\frac{p_i}{w}\right) = \sum_{j=1}^{n} a_{ij} \log\left(\frac{p_j}{w}\right) - \epsilon_i,$$

where $\epsilon_i = \log(z_i)$ is the log-productivity shock to firms in industry i. We notice that the relationship has to hold for every sector i. This means that if we define $\hat{p}_i = \log \frac{p_i}{w}$ as the relative price of good i and consider $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_n)$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$, we can write the whole system of equations as

$$\hat{p} = \mathbf{A}\hat{p} - \epsilon,$$

where $\mathbf{A}$, as it was already states, is the matrix of technological coefficients.

As a consequence, we see that, at the equilibrium, the vector of relative prices is given by

$$\hat{p} = -\left(\mathbf{I} - \mathbf{A}\right)^{-1} \epsilon. \tag{2.16}$$

In $(\mathbf{I} - \mathbf{A})^{-1}$ we immediately recognise the Leontief inverse $\mathbf{L} = \{l_{ij}\}$and thus that $\mathbf{L} = \sum_{k=0}^{+\infty} \mathbf{A}^k$.

Such decomposition shows that entry l_{ij} accounts not only for the relevance of sector j for sector i as a *direct* supplier, but also as an *indirect* one. From the property expressed above, we see that

$$l_{ij} = a_{ij} + \sum_{r=1}^{n} a_{ir} a_{rj} + \ldots$$

The first term considers the role of j as a direct supplier of i, the second one consider the role of j as a supplier of i's suppliers and so on. This means that over the production network, l_{ij} accounts for every directed walk from industry j to industry i over the network. This is what we will have in mind in deriving a random-walk centrality measure for economic networks later.

First-order conditions for the representative household show the she demands $c_j = \beta_j \frac{w}{p_j}$ units of j. The market clearing condition impose $y_j = c_j + \sum_{i=1}^{n} x_{ij}$. We can therefore write

$$p_j y_j = \beta_j w + \sum_{i=1}^{n} a_{ij} p_i y_i.$$

If we divide both sides of the equation by w and recall that the value added is equal to the income, we obtain

$$\lambda_j = \beta_j + \sum_{i=1}^{n} a_{ij}\lambda_i,$$

where λ_i is the Domar weight of sector i.
Again, if we let $\lambda = (\lambda_1, \ldots, \lambda_n)$, we obtain that

$$\lambda = \left(\mathbf{I} - \mathbf{A}^t\right)^{-1} + \beta,$$

or, equivalently, $\lambda_i = \frac{p_i y_i}{\text{GDP}} = \sum_{j=1}^{n} \beta_j l_{ji}$.
From Equation (2.16), we notice that

$$\log\left(\frac{p_i}{\text{GDP}}\right) = -\sum_{j=1}^{n} l_{ij}\epsilon_j.$$

This leads to an important result.

Theorem 2.3. The log-output of industry i is given by

$$\log y_i = \sum_{j=1}^{n} l_{ij}\epsilon_j + \delta, \qquad (2.17)$$

where δ is a constant independent of the shocks.

Theorem 2.3 has some noteworthy implications. Firstly, we notice that the output of industry i depends on more than itself and that input-output linkages are a mechanism of propagation of shocks from an industry to another. Moreover, it shows that propagation patterns are captured by Leontief inverse $\mathbf{L}$ and not by the matrix of technological coefficients $\mathbf{A}$. Propagation can therefore happen directly or indirectly. Finally, we see that the shocks propagate from sectors acting as suppliers to sectors acting as customers and not vice-versa. This is clear when considering the network interpretation of the Leontief inverse.

It is relatively straightforward to grasp the intuition underlying Theorem 2.3. If we imagine that a sector, say j, is hit by a negative shock, then its production level is affected. This implies a reduction in the total output and thus an increase in the price p_j. As a first direct consequence, customers of j are also affected, for they have to pay more to obtain the same amount of input. This shock will then reverberate to the customers of j's clients and, as it should be clear, a cascade of down propagations will occur. It is the complex effect of this initial shock that is captured by the coefficients in Leontief inverse matrix.

One may wonder why shock propagation only happens downstream and not the other way round. This is a consequence of the hypotheses of the production model, namely Cobb-Douglas technologies, a single factor of production (labour) and constant returns to scale. The price p_i of industry i equals its marginal costs, in turn depending only on its suppliers and its productivity coefficients. To put it in another way, we see that Domar weights do not depend on shocks.

We are now interested in analysing the macroeconomic effects of shocks in production networks (at least in this simple model). Considering Equation (2.16), we multiply every $\hat{p}_i$ by β_i so that summing we obtain

$$\log(\mathrm{GDP}) = \sum_{i,j=1}^{n} \beta_i l_{ij} \epsilon_j + \sum_{i=1}^{n} \beta_i \log p_i.$$

We also notice that we can choose a consumption good bundle whose price is given by

$$P_c = \prod_{i=1}^{n} p_i^{\beta_i},$$

because the numéraire implies that $\sum_{i=1}^{n} \beta_i \log p_i = 0$.

Theorem 2.4. The log-real value of the economy is given by

$$\log(\mathrm{GDP}) = \sum_{i=1}^{n} \lambda_i \epsilon_i, \tag{2.18}$$

where

$$\lambda_i = \frac{p_i y_i}{\mathrm{GDP}} = \sum_{j=1}^{n} \beta_j l_{ji}. \tag{2.19}$$

Theorem 2.3 is important for two main reasons. First of all, we see from Equation (2.17) that the log-output is a linear combination of the productivity shocks, where the coefficients are Domar weights. This means that Domar weights are a sufficient statistics to assess how a shock in a given industry will impact the overall production. Moreover, we see that in our model Domar weights take a simple form: they depend on the preferences and on the relative coefficients of L. That is to say, λ_i depends on the production network. In this network, the downstream propagation of a shock hitting a sector is related to the importance of the customers of that sectors and, proceeding downstream, to the importance of all the indirect customers.

2.3.1.1 *Demand-side Shocks*

Following ([1]) and ([17]), we now wish to assess what happens whenever a demand-side shock hits the economy.

We incorporate government purchases in our model, considering g_i as the exogenous purchase of good i. As a consequence, market clearing conditions become

$$y_i = c_i + g_i + \sum_{j=1}^{n} x_{ij}.$$

To simplify the derivation of the result, we assume that $z_i = 1$, for all $i \in \{1, 2, \dots, n\}$.

We now need to solve the equation. First-order conditions are given by

$$x_{ij} = a_{ij} p_i \frac{y_i}{p_j}$$

and

$$l_i = \alpha_i p_i \frac{y_i}{p_j}.$$

We plug those results into Equation (2.14) and solve. We obtain that for all $i \in \{1, 2, \dots, n\}$, $p_i = w$. The first difference we notice is that demand-side shocks have no impact whatsoever on relative prices. On the other hand, we see that the budget constraint of a representative household is given by $\sum_{i=1}^{n} p_i c_i = w - T$, where $T = \sum_{i=1}^{n} p_i g_i$. This is the government budget, financed by a lump tax on households. We can rewrite market clearing conditions as

$$y_i = \beta_i \left(1 - \sum_{j=1}^{n} g_j\right) + g_i + \sum_{j=1}^{n} a_{ji} y_j.$$

If we let $\mathbf{1} = (1, \dots, 1)$ be the n-dimension vector of components equal to 1, we can write in matrix form

$$y = \left(\mathbf{1} - g^t \mathbf{1}\right) \beta + g + \mathbf{A}^t y,$$

where $g = (g_1, \dots, g_n)$ is the vector of government demands.

This system of equations leads to the following.

Theorem 2.5. The output of sector i is

$$y_i = \sum_{j=1}^{n} l_{ji} g_j + \left(1 - \sum_{k=1}^{n} g_k\right) \left(\sum_{j=1}^{n} l_{ji} \beta_j\right). \tag{2.20}$$

We immediately notice a difference between Theorem 2.5 and Theorem 2.3: demand-side shocks are propagated considering l_{ji}, whereas supply-side shocks are propagated through l_{ij}. In other words, demand-side shocks propagate upwards: to direct and indirect suppliers.

We can see this noticing that whenever a positive demand shock hits a sector, the sector in question increases the demands of input goods. This, in turn, increase the demands of its suppliers and the demand of theirs suppliers and so on.

2.3.1.2 *The Network Origins of Aggregate Fluctuations*

We have tried to show how production networks contribute to the propagation of shocks because of their specific configuration. The question of whether localised shocks can cause fluctuations at an aggregate level has long been debated in economics. Dating back to ([55]), the idea that macroeconomic shocks could have an aggregate impacts was always deemed unlikely. The main argument was that of *diversification*: if n sectors are hit by different shocks, the standard deviation of the aggregate fluctuations would be proportional to $\frac{1}{\sqrt{n}}$. However, this argument ignores the role played by linkages. We have tried to shed some light on the effect such connections play, especially in economies with particular configurations.

2.4 VERTEX CENTRALITY IN INPUT-OUTPUT NETWORK

We are now interested in considering production networks as graphs. This will be useful mainly to derive a *centrality measure*, that is to say a measure allowing us to understand which sectors of an economy play a role whenever a shock hits an industry. To this aim, we will follow ([10].)

A vast number of centrality measures has been developed ([64]). However, there are three characteristics about graphs induced by production networks that makes it hard to apply what already exists in the literature. First of all, standardised data are usually completely connected. This renders measure based on shortest paths of little help. Moreover, they are directed because of the supplier-customer nature. Finally, self-loops play a fundamental role in some specific sectors.

We define a graph $G = (V, E)$, where $V = \{1, \dots, n\}$ is the set of nodes, each one corresponding to a sector of the economy, and $E \subset (V \times V)$ is a linkage between sectors. To each $(i, j) \in E$, we assign a weight a_{ij} corresponding to respective entry of the matrix $\mathbf{A}$ of the technical coefficients. To express missing edges, we set whenever necessary a_{ij} to zero.

Definition 2.10 (Strength of a node)**.** Let i be a node in V. We denote its *strength* as

$$k_i = \sum_{j=1}^{n} a_{ij}.$$

Definition 2.11 (Neighbourhood of a node)**.** Let $i \in V$ be a node. We define its *neighbourhood* as

$$N(i) = \{j \mid (i, j) \in E\}.$$

To model the movements of goods between sectors of an economy, we consider a *random walk* ([12]). A random walker, in graph theory, starts out walking at a given position, to select an edge incident to the one she is currently at. Such choices are made based on a probability distribution determined by the weight of the edges. The walker proceeds until either she runs out of time or a destination is reached ([11]).

Here, we are interested in the transition probabilities of the output produced by a sector as it flows from a given sector to the others. Thus, we normalise according to the rows of our table. To this aim, we create a matrix **K** that is equal to the various k_is on the diagonal and zero otherwise. We use this matrix to compute the transition matrix

$$\mathbf{M} = \mathbf{K}^{-1}\mathbf{A}.$$

Following the ideas in ([9]), we wish to develop a centrality measure accounting for the response of sectors in case of a shock. Such shocks are considered as a change occurring in an exogenous variable that has repercussion on the endogenous variable under scrutiny. In this setting, we shall consider as exogenous prices, technologies, firms, the distribution of profits, the government and its policies and final demand. We consider to be endogenous the flow of goods between sectors, with the relative payments.

Shocks are traced from the sector where they have origin, until the end of their random journey, after which they are assimilated by final demand. Thus, the target of the random walk is here the sector after which the shock is absorbed by final demand.

As an example, we consider an extra euro of production in some sector, say the naval industry, fixing as a target the agricultural sector. The extra production could be due, for instance, to a government programme. Such production will be randomly sold to another sector, according to the patterns deduced from **A**. Thus, this supply shock becomes the input of another sector. Of course, this extra euro of revenue will increase the payments to capital, labour and the indirect business taxes. This extra euro will lead to a surplus production of the same amount. This extra output will then be sold again. This will continue until eventually the shock reaches our target sector: agriculture. We can average over all pairs of shocks origins and targets to get a sense of how central a particular sector may be.

Notice that we are going to assume that shocks are *indivisible*. There are two main arguments for that. First and foremost, input-output models requires shocks not to be too strong. This is because a strong shocks would inevitably change the very nature of the economic systems under consideration. Thus, by assuming that shocks are small we are considering the case of shocks whose strength can somehow exist within the system without disrupting it. Moreover, if we allow shocks to be *divisible*, we also have to provide the model with a specification of how such divisions occur. The first idea might be that of assuming shocks split up at each node according to the transition matrix **M**. Yet, this means the shock immediately spreads out on the densely connected graph. Alternatively, we might consider a shock starting at some sector. At each transition, a split occurs and the fractional effects accumulate over all the industries. Such quantities will sooner or later reach a steady-state, as the absolute size of the shock is conserved. Such steady-state would be independent of the initial state. This could naturally induce a centrality measure. However, the proportion of divisible shock found in a node would be the same as the likelihood of finding an indivisible shock there. We can therefore consider our measure as a proxy of a steady-state distribution. Considering indivisible shocks also provides us with a centrality measure arising from an intuition of closeness of a node with respect to the others.

2.4.0.1 *Random Walk Centrality*

The main idea behind this measure of random walk centrality is the *mean first-passage time* (MFPT) ([11]).

Definition 2.12. Let $s \in V$ and $t \in V$. We define as

$$\mathbb{P}\left(s \xrightarrow{r} t\right)$$

to be the probability that a random walker going from s to t employs exactly r steps to move from her source to the destination.

Definition 2.13. Mean First-Passage Time. Let $s \in V$ and $t \in V$. The mean first-passage time from node s to node t is defined as

$$H(s,t) = \sum_{r=1}^{+\infty} r\,\mathbb{P}\left(s \xrightarrow{r} t\right). \tag{2.21}$$

Notice that for all t, $H(t,t) = 0$ as $\mathbb{P}\left(t \xrightarrow{r} t\right) = 0$ for all $r \geqslant 1$.

To consider the first visit of a target node t, we consider an absorbing random walk. That is to say, we assume that node t can never be left once it has been reached. To this aim, we modify **M**, deleting its t-th row and t-th column. We obtain a $(n-1) \times (n-1)$ matrix which we will call $\mathbf{M}_{-t}$.

Thus, taking entry (s,j) of $(\mathbf{M}_{-t})^{r-1}$ we obtain the probability of starting in s and being in j after exactly $(r-1)$ steps, without passing from t. Suppose to have a random walk starting from s and reaching t in r steps. We have that

$$\mathbb{P}\left(s \overset{r}{\to} t\right) = \sum_{i \neq t} \left\{ (\mathbf{M}_{-t})^{r-1} \right\}_{si} m_{it}.$$

We can combine this with Equation (2.21) to obtain

$$H(s,t) = \sum_{r=1}^{+\infty} r \sum_{i \neq t} \left\{ (\mathbf{M}_{-t})^{r-1} \right\}_{si} m_{it}.$$

Recognising the geometric series for matrices, we obtain

$$\sum_{r=1}^{+\infty} r \left(\mathbf{M}_{-t}\right)^{r-1} = \left(\mathbf{I} - \mathbf{M}_{-t}\right)^{-2}, \tag{2.22}$$

where $\mathbf{I}$ is the $(n-1)$-dimensional identity matrix. We know from ([54]) that it is possible to invert $\mathbf{I} - \mathbf{M}_{-t}$ because there are no absorbing states. Thus, we obtain

$$H(s,t) = \sum_{i \neq t} \left\{ \left(\mathbf{I} - \mathbf{M}_{-t}\right)^{-2} \right\}_{si} m_{it}. \tag{2.23}$$

From a computational point of view, we vectorise (2.23) to obtain vector

$$H(\cdot,t) = \left(\mathbf{I} - \mathbf{M}_{-t}\right)^{-2} m_{-t},$$

where $H(\cdot,t)$ is the vector of MFPTs for nodes ending in t (from any other node) and $m_{-t} = (m_{1,t}, \dots, m_{t-1,t}, m_{i+1,t}, m_{n,t})$ is the t-th column of $\mathbf{M}$ without its t-th row. Notice that if we let $\mathbf{1}$ be the $(n-1)$-dimensional vector of ones, we obtain

$$m_{-t} = \left(\mathbf{I} - \mathbf{M}_{-t}\right)^{-1} \mathbf{1}.$$

Thus,

$$H(\cdot,t) = \left(\mathbf{I} - \mathbf{M}_{-t}\right)^{-1} \mathbf{1}. \tag{2.24}$$

Notice that the computation of (2.24) can be further sped up employing the Sherman-Morrison algorithm ([34]).

Definition 2.14. Random Walk Centrality. We define as *random walk centrality* of the nodes in an input-output graph

$$C_{rw}(i) = \frac{n}{\sum_{j \in V} H(j,i)} \tag{2.25}$$

The indicator of Definition 2.14 has a straightforward economic interpretation. If a shock is about to hit a node with uniform probability, a high random walk centrality of a sector implies that it will be likely be affected in a short time.

3 NON-NEGATIVE MATRIX FACTORIZATION

3.1 LOW-RANK MATRIX APPROXIMATION

The necessity of approximating a matrix by means of another one of lower rank naturally arises in many problems of statistics and machine learning. In the usual context, a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ is given. Each column of $\mathbf{X}$ represents a data point in a m-dimension space. We are interested in approximating $\mathbf{X}$ through the product of two matrices, $\mathbf{W} \in \mathbb{R}^{m \times k}$ and $\mathbf{H} \in \mathbb{R}^{k \times n}$, so that

$$\mathbf{X} \approx \mathbf{WH} \tag{3.1}$$

In other words, we are willing to find a k-dimensional representation for the data points of $\mathbf{X}$. The columns of $\mathbf{W}$ constitute the basis, whereas those of $\mathbf{H}$ the coordinates of each point.

Depending on the exact mathematical formulation of this problem (namely the loss function measuring the error or any constraints on $\mathbf{W}$ or $\mathbf{H}$), it is possible to obtain several well-known techniques ([32]). For instance, Principal Component Analysis (PCA) can be seen as a low-rank approximation method where we impose orthogonality to the columns of $\mathbf{W}$ and to the rows of $\mathbf{H}$, considering the Frobenius norm to estimate the error ([42]).

Another example is k-medoids, a vector quantisation technique ([38]), that forces the columns of $\mathbf{H}$ to unary vectors and those of $\mathbf{W}$ to a subset of the columns of $\mathbf{X}$.

3.2 NON-NEGATIVE MATRIX FACTORISATION

If we set $\mathcal{N} = \{1, \ldots, n\}$ and $\mathcal{M} = \{1, \ldots, m\}$, we can write $\mathbf{X} = \{x_{ij}\}_{i \in \mathcal{M}, j \in \mathcal{N}}$. In a similar fashion, we have $\mathcal{K} = \{1, \ldots, k\}$ and $\mathbf{W} = \{w_{ij}\}_{i \in \mathcal{M}, j \in \mathcal{K}}$, $\mathbf{H} = \{h_{ij}\}_{i \in \mathcal{K}, j \in \mathcal{N}}$.

Whenever the data points in $\mathbf{X}$ lie in a space whose components are non-negative, namely when $x_{ij} \geqslant 0$, $\forall i \in \mathcal{M}, j \in \mathcal{N}$, it is of interest to impose the same constraint on $\mathbf{W}$ and $\mathbf{H}$.

This particular specification of a low-rank matrix approximation is known as *non-negative matrix factorisation* ([68, 49, 48])

The non-negativity constraint on $\mathbf{W}$ allows to interpret its columns as parts (or archetype or meta-genes). Analogously, the columns of $\mathbf{H}$ can be interpreted as coefficients signalling the importance of an archetype to that particular point.

This is particularly helpful in some statistical problems arising in genomics ([29]), text mining ([27]) and sound recognition ([8])

After the seminal work by ([49]), a plethora of different non-negative matrix factorisation methods have been investigated. Different implementations vary on their different definition of a *loss function* (3.2.1), *update algorithm* (3.2.2), *stopping criterion* (3.2.3) and *initialisation method* (3.2.4).

3.2.1 *Loss functions*

Loss functions for NMF have all a general form, which we can formulate as

$$\mathcal{L}\left(\mathbf{W},\mathbf{H};\mathbf{X}\right)+\rho\left(\mathbf{W},\mathbf{H}\right), \tag{3.2}$$

subject to the already described non-negativity constraints. The first addendum is a *loss function* measuring the fitness of the approximation. The second addendum is a *regularisation term*, whose aim is that of enforcing particular conditions (*f.i.* sparsity) on the solution ([23])

In ([49, 48]), two main proposals regarding the loss functions are made:

1. $\mathcal{L}:(\mathbf{A},\mathbf{B})\mapsto\|\mathbf{A}-\mathbf{B}\|_F^2$, which is a minimisation according to the well-known Frobenius norm.
2. $\mathcal{L}:(\mathbf{A},\mathbf{B})\mapsto\sum_{ij}\left[\left(\mathbf{A}_{ij}\log\frac{\mathbf{A}_{ij}}{\mathbf{B}_{ij}}\right)-\mathbf{A}_{ij}+\mathbf{B}_{ij}\right]$, which is the Kullback-Leiber divergence.

As ([48]) points out, despite the problems induced by the loss functions being convex with respect to either **W** or **H**, convexity is not preserved whilst optimising with respect to both. Thus, we aim at reaching a *local minimum* in our endeavour.

It is possible to further generalise the loss-function to a general β-divergence, thus obtaining the so-called β-NMF ([80]).

As far as the regularisation term is concerned, ([4]), ([20]) and ([78]) provide a thorough review of different possibilities. Usually, sparsity is induced by means of a linear combination of both ℓ_2 and ℓ_1 regularisation terms on both matrices (and this is how NMF is implemented in the machine learning library Scikit-learn ([70])). An exception can be found in ([90]), where the regularisation term forces the columns of **W** to be orthogonal, as it is used to perform blind source separation.

3.2.2 *Update Algorithms*

NMF problems are usually formulated as non-linear optimisation problems, with an objective function of the kind already described measuring the goodness of the approximation. We will focus on the most common solutions when dealing with a Frobenius objective function, as it is the algorithm that was implemented throughout the analysis.

As it was pointed out, the intrinsic difficulty of the problem and the necessity of obtaining results in a reasonable time result in the search for a *local optimum* of the NMF problem.

The optimisation problem becomes that of minimising

$$||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2,$$

constraining **W** and **H** to be element-wise non-negative.

Algorithms usually targets stationary points of the problem. We can formulate the well-known Karush-Kuhn-Tucker conditions ([66]) for the problem as ([32]):

$$\nabla_{\mathbf{W}} ||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2 \geqslant 0$$

$$\mathbf{W} \circ \nabla_{\mathbf{W}} ||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2 = 0$$

$$\nabla_{\mathbf{H}} ||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2 \geqslant 0$$

$$\mathbf{H} \circ \nabla_{\mathbf{H}} ||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2 = 0,$$

where

$$\nabla_{\mathbf{W}} ||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2 = -2\left(\mathbf{X} - \mathbf{W}\mathbf{H}\right)\mathbf{H}^t$$

$$\nabla_{\mathbf{H}} ||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2 = -2\mathbf{W}^t\left(\mathbf{X} - \mathbf{W}\mathbf{H}\right).$$

Notice that by $\circ$ we refer to the Hadamard product.

Most algorithms for solving this problem are iterative, as when we fix either **W** or **H** we obtain a convex problem. The general scheme is the following:

1. Initialisation: select the initial matrices **W** and **H**.
2. Until convergence:
   a) Fix **H** and find an admissible **W** such that $||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2$ is reduced.
   b) Fix **W** and find an admissible **H** such that $||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2$ is reduced.

We notice the symmetry between the update of **W** and that of **H**, as up to a transposition they are the conceptually the same problem. Although, at least theoretically, one could solve the non-negative least squares problems exactly at each iteration, it is usually the case that approximate solutions are found. As a consequence, an inexact two-block coordinates scheme is usually obtained ([32]). To this aim, any non-linear optimisation method can be employed, such as multiplicative updates, Newton methods or block-coordinate descent ([24, 4, 46, 53]). We now briefly introduce some of the main algorithms, following ([32])

3.2.2.1 *Multiplicative Updates*

The main update rule of multiplicative updates is, given **X**, **W** and **H**, to do

$$\mathbf{W} \leftarrow \mathbf{W} \circ \frac{[\mathbf{X}\mathbf{H}^t]}{[\mathbf{W}\mathbf{H}\mathbf{H}^t]}, \tag{3.3}$$

where $\frac{[\mathbf{A}]}{[\mathbf{B}]}$ indicates the element-wise division of two matrices. The algorithm was first proposed in ([26]) and consequently modified for use with NMF in ([48]). This algorithm belongs to the class of majorisation-minimisation algorithms (a generalisation of the well-known family of expectation-maximisation methods). This is because (3.3) is the global minimisers of a function majorising a Frobenius loss function. That is to say, a function which is greater or equal than a Frobenius loss-function and equals to it at the current iterate. Minimising it thus guarantees the loss-function to decrease monotonically. Another way of see (3.3) is as a variation of a gradient descent, as

$$\mathbf{W} \circ \frac{[\mathbf{X}\mathbf{H}^t]}{[\mathbf{W}\mathbf{H}\mathbf{H}^t]} = \mathbf{W} - \left(\frac{[\mathbf{W}]}{[\mathbf{W}\mathbf{H}\mathbf{H}^t]} \circ \nabla_{\mathbf{W}} \frac{1}{2} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 \right).$$

A third, and perhaps more revealing, way of considering multiplicative updates is by noticing that

$$\frac{\{[\mathbf{X}\mathbf{H}^t]\}_{ik}}{\{[\mathbf{W}\mathbf{H}\mathbf{H}^t]\}_{ik}} \geqslant 1 \iff \left\{ \nabla_{\mathbf{W}} \frac{1}{2} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 \right\}_{ik} \leqslant 0.$$

This means that for the KKT conditions to be satisfied, we need to have that each entry of **W** increases if its partial derivative is negative, decreases whenever it is positive and remains unchanged if it is equal to zero. This poses risks to the convergence of the algorithm as an entry of **W** might be zero and its partial derivative be negative. Several solutions have been proposed to deal with the issue, by either changing the step of the upgrade or replacing zero-entries with small values ([52]).

3.2.2.2 *Alternating Least Squares*

The Alternating Least Squares method computes the optimal solution of the problem we obtain by lifting the entry-wise non-negativity constraints on **W**. We then set

$$\{\mathbf{W}\}_{ij} \leftarrow \max\left\{\left\{\underset{\mathbf{A}\in\mathbb{R}^{(n\times k)}}{\arg\min}\ \|\mathbf{X}-\mathbf{AH}\|_F^2\right\}_{ij}, 0\right\}. \tag{3.4}$$

(3.4) does not guarantee convergence, as the objective function might oscillate between consecutive updates, but the algorithm is computationally cheap with respect to other methods. It is therefore employed in preprocessing.

3.2.2.3 *Alternating Non-negative Least Squares*

Alternating non-negative least squares algorithms aims at finding an optimal solution to the iterative updates of both **W** and **H**. Namely, they seek to find a solution to

$$\mathbf{W} \leftarrow \underset{\{\mathbf{W}\}_{ij}\geqslant 0}{\arg\min}\ \|\mathbf{X}-\mathbf{WH}\|_F^2,$$

or to the respective **H** update step. There exist a number of algorithms to compute the optimal solution. In practice, *ad hoc* active set methods ([47]) have proven to perform well. Other solutions involve the use of projected gradients ([53]), Quasi-Newton methods ([21]) or fast gradient methods ([37]). Active set methods are guaranteed to find a local optimum ([36]). However, as finding the exact solution to non-linear non-negative least squares problems is computationally expensive, it is usually advisable to first perform some steps with a less expensive algorithm (*f.i.* using alternating least squares or multiplicative updates).

3.2.2.4 *Hierarchical Alternating Least Squares*

Hierarchical Alternating Least Squares (HALS) ([22]) solve the non-linear non-negative least squares problems of the update phase by decoupling the problem. In other words, each column of **W** (or **H** respectively) is considered separately. This is due to the fact that the columns of **W** do not interact with each other. We thus obtain k different problems (where k is the rank of the NMF). If we let w_j be the k-th column of **W**, with $j = 1, \ldots, k$, and h_i be the i-th row of **H** in a similar fashion, we obtain

$$w_j = \underset{w_{j_l}\geqslant 0}{\arg\min}\ \left\|\mathbf{X} - \sum_{k\neq j} w_k h_k - w_j h_j\right\|_F,$$

HALS also converges to local optima under rather mild conditions ([31]).

3.2.3 *Stopping criterion*

Stopping criteria are usually based on either a time/iteration budget constraint, on the value of the objective function or of its variation, as it is common with non-linear optimisation problems ([66]).

3.2.4 *Initialisation method*

The initialisation of matrices **W** and **H** is of crucial importance as NMF algorithms converge to local optima.

Several initialisation techniques have been proposed, although they all come with little or no theoretical guarantee.

- *Clustering techniques.* A clustering technique is employed in order to find k centroids (*f.i.* k-means methods) and initialise **W**. Matrix **H** is scaled accordingly.
- *Singular Value Decomposition.* The idea relies upon the well-known singular value decomposition technique or upon its non-negative extension ([13]).
- *Column Subset Selection.* A subset of k columns of **X** is considered as the initial point for **W**. **H** is then initialised randomly.

3.3 RANK SELECTION

Together with convergence to local optima, rank selection is one of the most controversial aspects of NMF. Various solutions have been proposed.

In ([15]), NMF is computed considering several random initial points for growing ranks. Data points are then clustered assigning them based on their biggest loading in the respective activation column in matrix **H**, assuming the number of cluster equals that of the rank k. An $(m \times m)$ connectivity matrix $C_k^{(i)}$ is then built setting entries equal to 1 whenever two points are in the same cluster and 0 otherwise. A consensus matrix $\mathcal{C}_k$ is then computed by considering the element-wise averages of the various connectivity matrices. In this way, it is possible to assert the stability of the clusters. Such stability is computed by means of cophenetic coefficients. This idea is also present in ([29]), although consensus matrices are weighted according to the residual error of the NMF inducing them. Moreover, the comparison is also carried out by comparing how well the reconstructed matrix scores against a permutation of the data. ([40]) propose to consider the variation of the residual sum of squares for different ranks (which is equal to the Frobenius loss-function) and try to investigate whether a particular point emerges. Other possibilities involve considering either the opinion of a group of experts in the field, computing a singular value decomposition of the data matrix and observe the decay of singular values ([6]) or by first performing a k-means clustering of the data at hand ([33]).

More sophisticated approached have also been proposed. Bayesian NMF allows to estimate the rank by incorporating prior information into the model ([81]). ([79]) propose a rank selection method obtained through a *minimum description length* procedure, in which rank is chosen weighing the complexity of the model and its ability to compress the data.

3.3.1 *Cross-Validation for rank estimation*

Cross-validation is a well-known technique employed in supervised learning to estimate parameters ([7]). The main idea behind cross-validation is to separate a dataset into two fractions: one of the two is employed to train the model and the other part to test its validity. Such estimation is repeated considering subsequent redefinitions of the train and test sets. Two different adaptations of this idea were considered in this work.

In the first case, column i of matrix $\mathbf{X}$ was held out. NMF was then computed for the remaining matrix, $\mathbf{X}_{-i}$. $\mathbf{W}_{-i}$ and $\mathbf{H}_{-i}$ were subsequently computed. Column i of matrix $\mathbf{X}$, x_i was then considered. Its activation was obtained by solving

$$\hat{h}_i = \underset{h_{i_j} \geqslant 0}{\arg\min} \; ||x_i - \mathbf{W}_{-i} h_i||_F^2 \,,$$

to set

$$\hat{e}_i = \big|\big|x_i - W_{-i}\hat{h}_i\big|\big|_F \,.$$

The error was computed as

$$\hat{e}^{(k)} = \sum_{j=1}^{n} \hat{e}_j^{(k)},$$

for increasing ks. We show the details of the procedure in Algorithm 3.1.

The second adaption is from ([67]). The main idea is that of censoring a random $(k \times k)$ matrix and then reconstructing the censored matrix, thus assessing the error. We show two different versions of this procedure in Algorithms 3.2 and 3.3.

Algorithm 3.1 Cross-Validation for the NMF

```
Let $\mathbf{X}$ be an entry-wise non-negative matrix of dimensions $m \times n$ and let $\mathcal{K} \subseteq \{1, \dots, \min\{m, n\}\}$ be a set of ranks. Let $\mathbf{X}_{-i}$ be the matrix obtained by removing the $i$-th column of $\mathbf{X}_i$ and let $x_i$ be the removed column.
for $k \in \mathcal{K}$ do
    $CV(k) \leftarrow 0$
end for
for $k \in \mathcal{K}$ and $i \in \{1, \dots, n\}$ do
    $\mathbf{W}_{-i} \leftarrow NMF(\mathbf{X}_{-i}, k)$
    $\hat{h}_i \leftarrow \arg\min_{h_{ij} \geqslant 0} ||x_i - \mathbf{W}_{-i} h_i||_F^2$
    $CV(k) \leftarrow CV(k) + \left|\left|x_i - \mathbf{W}_{-i} \hat{h}_i\right|\right|_F$
end for
```

Algorithm 3.2 Bi-Cross-Validation for the NMF with non-negative residuals

```
Let $\mathbf{X}$ be an entry-wise non-negative matrix of dimensions $(m \times n)$, let $\mathcal{I}_l \subset \{1, \dots, m\}$ and $\mathcal{J}_l \subset \{1, \dots, n\}$ be, respectively, a row and column holdout subsets for $l \in \{1, \dots, L\}$, where $L$ is a positive integer. Let $\mathcal{K} \subseteq \{1, \dots, \min\{m, n\}\}$ be a set of ranks.
for $k \in \mathcal{K}$ do
    $BCV(k) \leftarrow 0$
end for
for $l \in \{1, \dots, L\}$ and $k \in \mathcal{K}$ do
    $\mathcal{I} \leftarrow \mathcal{I}_l$ and $\mathcal{J} \leftarrow \mathcal{J}_l$
    $\mathbf{H}^{(k)}_{-\mathcal{I},-\mathcal{J}}, \mathbf{W}^{(k)}_{-\mathcal{I},-\mathcal{J}} \leftarrow NMF(\mathbf{X}_{-\mathcal{I},-\mathcal{J}}, k)$
    $\mathbf{W}^{(k)}_{\mathcal{I},\mathcal{J}} \leftarrow \arg\min_{W_{ij} \geqslant 0} \left|\left|\mathbf{X}_{\mathcal{I},-\mathcal{J}} - \mathbf{W}\mathbf{H}^{(k)}_{-\mathcal{I},-\mathcal{J}}\right|\right|_F^2$
    $\mathbf{H}^{(k)}_{\mathcal{I},\mathcal{J}} \leftarrow \arg\min_{H_{ij} \geqslant 0} \left|\left|\mathbf{X}_{-\mathcal{I},\mathcal{J}} - \mathbf{W}^{(k)}_{-\mathcal{I},-\mathcal{J}}\mathbf{H}\right|\right|_F^2$
    $\hat{\mathbf{X}}^{(k)}_{\mathcal{I},\mathcal{J}} \leftarrow \mathbf{W}^{(k)}_{\mathcal{I},\mathcal{J}} \mathbf{H}^{(k)}_{\mathcal{I},\mathcal{J}}$
    $BCV(k) \leftarrow BCV(k) + \left|\left|X_{\mathcal{I},\mathcal{J}} - \hat{\mathbf{X}}^{(k)}_{\mathcal{I},\mathcal{J}}\right|\right|_F^2$
end for
```

Algorithm 3.3 Bi-Cross-Validation for the NMF with simple residuals

```
Let $\mathbf{X}$ be an entry-wise non-negative matrix of dimensions $(m \times n)$, let $\mathcal{I}_l \subset \{1, \ldots, m\}$ and $\mathcal{J}_l \subset \{1, \ldots, n\}$ be, respectively, a row and column holdout subsets for $l \in \{1, \ldots, L\}$, where $L$ is a positive integer. Let $\mathcal{K} \subseteq \{1, \ldots, \min\{m, n\}\}$ be a set of ranks. Let $\mathbf{A}^{+}$ denote the Moore-Penrose inverse of $\mathbf{A}$.
for $k \in \mathcal{K}$ do
    $\mathrm{BCV}(k) = 0$
end for
for $l \in \{1, \ldots, L\}$ and $k \in \mathcal{K}$ do
    $\mathcal{I} = \mathcal{I}_l$ and $\mathcal{J} = \mathcal{J}_l$
    $\mathbf{H}^{(k)}_{-\mathcal{I},-\mathcal{J}}, \mathbf{W}^{(k)}_{-\mathcal{I},-\mathcal{J}} = \mathrm{NMF}(\mathbf{X}_{-\mathcal{I},-\mathcal{J}}, k)$
    $\hat{\mathbf{X}}^{(k)}_{\mathcal{I},\mathcal{J}} = \mathbf{X}_{\mathcal{I},-\mathcal{J}} \left(\mathbf{H}^{(k)}_{-\mathcal{I},-\mathcal{J}}\right)^{+} \left(\mathbf{W}^{(k)}_{-\mathcal{I},-\mathcal{J}}\right)^{+} \mathbf{X}_{-\mathcal{I},\mathcal{J}}$
    $\mathrm{BCV}(k) = \mathrm{BCV}(k) + \left|\left|\mathbf{X}_{\mathcal{I},\mathcal{J}} - \hat{\mathbf{X}}^{(k)}_{\mathcal{I},\mathcal{J}}\right|\right|^{2}_{F}$
end for
```

4 COMPOSITIONAL DATA ANALYSIS

We now wish to introduce the main elements of *compositional data analysis*. To this endeavour, we will rely on ([69]).

Definition 4.1 (D-part composition)**.** A vector $x = (x_1, x_2, \dots, x_D)$ is a D-part composition whenever all its components are strictly positive real numbers and only carry relative information.

By *relative information* we mean to say that information is contained in the ratios between components of the D-composition and thus the actual numerical values are irrelevant by themselves. This can be due to the fact that the data are *closed data* ([19]), *i.e.* bound to sum to some constant κ, or because there is a straightforward transformation allowing the data to form proportions.

Definition 4.2 (Compositions as equivalence classes)**.** Let x and y be two D-compositions. They are *compositionally equivalent* if there exists $\lambda \in \mathbb{R}_+$ such that $x = \lambda y$.

It is therefore possible to express any D-composition by selecting any other D-composition lying in the same equivalence class.

Definition 4.3 (Closure)**.** Let z be a D-composition. We define the *closure* of z to $\kappa \in \mathbb{R}_+$ as

$$\mathcal{C}(z) = \left(\frac{\kappa z_1}{\sum_{j=1}^{D} z_j}, \dots, \frac{\kappa z_D}{\sum_{j=1}^{D} z_j} \right). \tag{4.1}$$

Remark 4.4. *Two* D*-compositions* x *and* y *are compositionally equivalent if and only* $\mathcal{C}(x) = \mathcal{C}(y)$*, for all* $\kappa \in \mathbb{R}_+$*.*

For the reasons reported above, we will henceforth regard compositional data as vectors summing to some $\kappa \in \mathbb{R}_+$.

Definition 4.5 (Sample space)**.** The sample space of compositional data is given by the simplex

$$\mathcal{S}^D = \left\{ x = (x_1, \dots, x_D) \text{ s.t. } x_i \geqslant 0, i = 1, \dots, D; \sum_{j=1}^{D} x_j = \kappa \right\} \tag{4.2}$$

With an abuse of notation, despite compositions being equivalence classes, we will use the term to refer to members of such classes.

To reduce the dimensionality of compositions, two procedures are usually employed.

Definition 4.6 (Subcompositions)**.** Let x be a D-compositions and let $S = i_1, \dots, i_s$ be a subset of its indices. A subcomposition x_S, with S parts, is defined as

$$x_S = \mathcal{C}\left((x_{i_1}, \dots, x_{i_s})\right).$$

Definition 4.7 (Amalgamation)**.** Let $x \in \mathcal{S}^D$ and let $A = \{i_1, \dots, i_a\}$ be a subset of its indices such that $D - a \geqslant 1$. Let $\bar{A}$ be the set of remaining indices. Then the value

$$x_A = \sum_{j \in A} x_i$$

is the *amalgamated part* or *amalgamated component*. The resulting vector $x^* = (x_{\bar{A}}, x_A)$ is called *amalgamated composition* and lies in $\mathcal{S}^{(D-a+1)}$.

4.0.1 *Principles of compositions analysis*

There are three main principles that must hold whenever compositional analysis is employed: scale invariance, permutational invariance and subcompositional coherence. Scale invariance implies that data only carry *relative information*. In other words, we do not deem totals to be meaningful *per se*. Analyses should yield the same output notwithstanding totals.

Definition 4.8 (Scale invariant functions.)**.** Let f be a function defined on $\mathbb{R}^D_+$. We say that f is scale invariant if it is a 0-degree homogenous function of the parts of $x \in \mathcal{S}^D$. In other words, if we let $\lambda \in \mathbb{R}_+$, it holds that for any $x \in \mathcal{S}^D$, $f(\lambda x) = f(x)$.

The second principle is permutational invariance. This means that results should not change whenever the ordering of the parts of the compositions is changed. The third one is subcompositional coherence. In loose words, subcompositions should behave like orthogonal projections do in real analysis. The size of a projected segment can never be greater than that of the segment itself.

4.0.2 *The Aitchison geometry*

We aim at endowing compositions with the algebraic structure of a vector space ([77]).

Definition 4.9 (Perturbation)**.** Let $x \in \mathcal{S}^D$ and $y \in \mathcal{S}^D$. We define as *perturbation* the operation

$$x \oplus y = \mathcal{C}\left((x_1 y_1, \dots, x_D y_D)\right) \in \mathcal{S}^D. \tag{4.3}$$

We also consider the field $\mathbb{R}$ to define the external operation.

Definition 4.10 (Powering)**.** Let $x \in \mathcal{S}^D$ and $\alpha \in \mathbb{R}$. We define as a *powering* the operation

$$\alpha \odot x = \mathcal{C}\left(x_1^\alpha, \dots, x_D^\alpha\right). \tag{4.4}$$

Lemma 4.1 (Abelian Group structure of Perturbation)**.** *Let $\mathcal{S}^D$ be a compositional space endowed with $\oplus$, as defined in* (4.9)*. Then, for all x, y, $z \in \mathcal{S}^D$ the following holds:*

1. *Commutative property:* $x \oplus y = y \oplus x$.
2. *Associative property:* $(x \oplus y) \oplus z = x \oplus (y \oplus z)$.
3. *Neutral element:*

$$n = \mathcal{C}\left((1, \dots, 1)\right) = \left(\frac{1}{D}, \dots, \frac{1}{D}\right).$$

   Uniqueness descends from the uniqueness of the barycentre of a simplex.

4. *Inverse element:* $y = x^{-1}$ s.t. $x \oplus y = n$.

Lemma 4.2. *Let $\mathcal{S}^D$ be a compositional space, endowed with the internal operation $\oplus$,* (4.9)*, and the external operation $\odot$,* (4.10)*, defined between itself and the field $\mathbb{R}$. Then, for all x, $y \in \mathcal{S}^D$, α , $\beta \in \mathbb{R}$, the following holds:*

1. $\left(\mathcal{S}^D, \oplus\right)$ *is an Abelian Group.*
2. *Associative property:* $\alpha \odot (\beta \odot x) = (\alpha\beta) \odot x$.
3. *Distributive properties:*
   a) $\alpha \odot (x \oplus y) = (a \odot x) \oplus (\alpha \odot y)$
   b) $(\alpha + \beta) \odot x = (\alpha \odot x) \oplus (\beta \odot x)$.
4. *Neutral element:* $1 \odot x = x$.

We obtain that $\left(\mathcal{S}^D, \mathbb{R}\right)$ is a vector space.

Remark 4.11. *We shall indicate $x \oplus y^{-1}$ as $x \ominus y$.*

4.0.2.1 *Aitchison Inner Product*

Definition 4.12 (Aitchison Inner Product)**.** Let $x \in \mathcal{S}^D$ and $y \in \mathcal{S}^D$. We define the *Aitchison inner product* as

$$\langle x, y \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}. \tag{4.5}$$

Definition 4.13 (Aitchison Norm)**.** Let $x \in \mathcal{S}^D$. The Aitchison norm, induced by the inner product is

$$\|x\|_a = \sqrt{\langle x, x \rangle_a} = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left[\ln \left(\frac{x_i}{x_j} \right) \right]^2}. \tag{4.6}$$

Proving that $\mathcal{S}^D$ is complete with respect to the Aitchison norm (noting that it is finite-dimensional) ([5]), we then obtain that $(\mathcal{S}_D, \|\cdot\|_a)$ is a Hilbert space ([14]).

Definition 4.14 (Aitchison Metric)**.** . Let $x \in \mathcal{S}^D$ and $y \in \mathcal{S}^D$. We define the Aitchison metric (or distance) as

$$\rho_a(x, y) = \|x \ominus y\|_a = \sqrt{\langle (x \ominus y), (x \ominus y) \rangle_a} \tag{4.7}$$

Remark 4.15. *A Aitchison space* $\left(\mathcal{S}^D, \langle \cdot, \cdot \rangle_a\right)$ *endowed with the Aitchison metric is a complete metric space.*

4.0.3 *Clustering in the Aitchison geometry*

Cluster analysis serves a variety of scopes, which have in common the idea of finding subsets of data points that are somewhat homogenous. This may be of interest because a meaningful representation of the data is sought or to assess if data consist of different subgroups. The mathematical framework in which data needs to be endowed thus vary profoundly depending on the model under scrutiny. We can roughly group clustering techniques into two main categories. On the one hand there are *hierarchical clustering* techniques. Such techniques begin by considering each point as a cluster. Clusters are then aggregated by iteratively merging clusters that are deemed “closest" to each other, according to a definition of closeness, until a cluster identical to the entire dataset is found (it is also possible to proceed the other way round: from a unique cluster to every point being a cluster). The second category is that considering *point assignment*: after some initial clusters have been determined, points can be assigned to a cluster to which they belong according to some rules. This second group of clustering techniques usually represent clusters by means of a *prototype* or *centroid* ([89]). A well-known example is that of the k-means algorithms. However, the definition of centroids as soon as the usual Euclidean context is abandoned is not necessarily straightforward. A solution has been the generalisation of the definition of centroid to complete metric spaces. If we let (Ω, ρ) be a complete metric space and $X = \{x_i\}_{i=1}^n \subseteq \Omega$ be a sequence of random points, we can define the Fréchet variance of a point y as

$$\Psi_X(y) = \sum_{i=1}^{n} (\rho(y, x_i))^2 .$$

The Karcher means are then the points μ with the following property:

$$\mu \in \arg\min_{y \in M} \Psi_X\left(y\right).$$

Whenever Ψ_X admits a global minimiser, *f.i.* if it is convex, then we have a unique point μ which we call the Fréchet mean ([65]).

Another possibility is that of considering as prototype of a cluster an element belonging to that cluster. The *k-medoids*, also known as PAM (*partitioning around medoids*), is an extensions of the k-means algorithm. In this work, the solution proposed in ([45]) was employed and the general form of k-medoids as found in ([38]) is reported as Algorithm (4.1).

Algorithm 4.1 K-medoids

```
repeat
    For a given clustering $C$, find the points in the cluster that minimise the total distance to other points in that cluster:
        $i_k^* = \arg\min_{i:C(i)=k} \sum_{C(i')=k} \rho\left(x_i, x_{i'}\right).$
    Then $m_k = x_{i_k^*}$, for $k = 1, \dots, K$, are the current estimate of the medoids.
    Given the current set of cluster centers $\{m_1, \dots, m_K\}$, minimise the total error by assigning each observation to its closest medoid:
        $C\left(i\right) = \arg\min_{1 \leqslant k \leqslant K} \rho\left(x_i, m_k\right).$
until assignments do not change.
```

We notice that the extension to the compositional case is immediate. A compositional space is a Hilbert space and it is therefore a complete metric space with the metric induced by the scalar product. Considering Algorithm (4.1), we only need to exploit the Aitchison distance whenever the analysis is to be carried out within a compositional data framework or a Euclidean distance if it is of interest to work in the usual Euclidean space.

5 ANALYSIS

5.1 INPUT-OUTPUT ANALYSIS

The vast majority of national statistical offices compute input-output tables. As far as Italy is concerned, input-output tables are published by Istituto Nazionale di Statistica (ISTAT), the national statistical office[1]. National tables usually allow for a good level of detail, considering large number of sectors. However, inter-country analysis is complicated by the fact that different nations employ different classification schemes and techniques to elaborate data. International databases aim at providing inter-country homogenous data. An example, in this sense, is the experimental work Eurostat is conducting, preparing a homogenous database for countries in the EU. At present, there are two major international databases: the World Input Output Database (WIOD) and the database published by the OECD. In this work, the OECD dataset will be analysed.

The OECD input-output dataset[2] is a relevant part of the Structural Analysis (STAN) exercise, part of the Economic Analysis and Statistic Division, under the OECD Directorate for Science, Technology and Industry. In this work we will consider the Input-Output dataset in its third revision, spanning from the year 1995 to 2011. Domestic input-output tables are computed for the thirty-six members of the OECD, twenty-seven more countries and a fictitious *Rest of the World* aggregate. The number of sectors in which economies are divided is thirty-four. However, the thirty-fourth sector *"Private households with employed persons"* will not be considered in the analysis, for it is only traced by a handful of countries. Input-output tables also contain information regarding the payment sector, profits, wages and trade. However, we shall limit the scope of the analysis to the domestic inter-industrial tables. The dataset will be therefore composed by sixty-four (33×33) matrices per year. Data were normalised considering the total output of each sector, as described in Chapter 2, yielding a dataset of matrices of technical coefficients. The intrinsic difficulty of the analysis is given by the poor interpretability of clusters. Once a group of countries has been defined, there exists no immediate way of analysing its members. This is due the complex nature of the input-output matrices: a matrix of technical coefficients describes the structural nature of an economy and the nature of that economic is an emerging property of the coefficients. As an example, we see

[1] https://www.istat.it/it/archivio/225665

[2] http://oe.cd/i-o

in Figure 5.1 and Figure 5.2 that despite the differences between countries or within countries over time, there is no immediate way of understanding which feature should be relevant. The scope of the work was therefore twofold: on one hand it was necessary to find a way to reduce the dimensionality of data and on the other one to define a method to interpret results.

To this aim, a combination of the techniques insofar presented was employed. First of all, technical coefficients matrices were vectorised in a row-wise fashion: subsequent rows were juxtaposed so that data referred to a country and a specific year became a vector. Datasets were created considering both combinations of years and geographical zones, combining the relevant vectors. In particular, the analysis was conducted following a cross-sectional approach and focusing on either the countries in the EU, members of the OECD or considering the whole World dataset. It is important to notice that, as a result, our databases comprise country on the columns and inter-sector interactions on the rows.

Dimensionality reduction was performed by means of the non-negative matrix factorisation technique as implemented in ([70]). According to the notation introduced in Chapter 3, it corresponds to a Frobenius-like NMF. Updates were computed with a HALS algorithm. The stopping criterion employed considered two conditions: an error tolerance of 0.0001 and a computational budget of 200 iterations. Given the dimensionality of the dataset and the convergence to local optima of NMF, initialisation was set taken into accounts the opinion of a pool of experts. The first iteration, therefore, started from a point of economic interest. As far as rank estimation is concerned, different methodologies were considered. A singular value decomposition was performed for the dataset under scrutiny and the decay of its singular values analysed to establish a range of plausible values. This was corroborated by performing a k-means clustering on the matrices, trying to establish the optimum number of clusters of inter-sectoral relationship considering the well-known Silhouette coefficient, first presented in ([76]). As the output of k-means depends on the initialisation point, the algorithm was run 500 times and the best result was stored. To conclude, both the cross-validation and the bi-cross-validation were implemented in Python, relying on the NumPy matrix library ([86]) and the SciPy library ([43]).

NMF, as already described in Chapter 3, yields a low-rank decomposition of a matrix into two parts, which can be read as a matrix of archetypes and as a matrix of activations. This means that we are able to move from a dataset that comprises a large number of countries to a small dictionary of reduced number of fundamental economic systems. Those archetypes can be rewritten in matrix forms, yielding actual technical coefficients matrices because of the non-negativity constraints NMF imposes. The factorisation allows therefore for a decoupling of the problem into two parts, separating the clustering problem from the interpretation. The basis, formed of a certain number of input-output tables (the archetypes) will give meaning. The activation matrix will be considered to compute differences between countries, considering how different archetypes participate to their definition.

In a way, we might say that instead of analysing a set of buildings we decide to separate them into bricks and compositions. We wish to find some fundamental units such as “fired bricks" or “rock bricks" and then to see if we can distinguish the Pyramids from the Colosseum by seeing what bricks are present in each structure and how they were used.

Yet, as it was already pointed out, the mere decomposition is not sufficient for the interpretation of the results. For this reason, each archetype was considered as an adjacency matrix of a graph, as already described in Section 2.4.0.1. Different sectors were ranked according to the mean velocity at which they are reached by an economic shock spreading from any other node. This allowed to establish a ranking of the sectors with a straightforward and plausible economic interpretation. Furthermore, it greatly simplifies the interpretation of archetypes and their input-output tables, as we move from considering inter-sector relations to analysing a more limited number of actual sectors. Considering economies as the superposition of different types of meta-countries, with particular characteristics which can be analysed with all the techniques developed for input-output tables can be thought of as a way of decoupling an otherwise very difficult problem. In fact, once we know of which “bricks" a building is made of, we can perform any type of test on such “bricks", depending on how we wish to characterise them. The advantage of the non-negativity constraints assumed by NMF are therefore twofold: on the one hand, an economic interpretation is possible, on the other one, the fact that different archetypes are unable to cancel out leads to a factorisation into parts that can unveil the structure of a complex dataset.

As far as the activation matrix is concerned, clustering was conducted considering two different frameworks. In both cases, the k-medoids algorithm described as Algorithm 4.1 was considered. The analysis was carried out in R ([73]) and the implementation of the k-medoid algorithm is that found in ([56]). The first approach was compositional, following Chapter 4. We know a country is represented by a limited number of components, each representing the importance of a particular archetype in its definition. In a compositional approach, we are not very much interested in the magnitude of such components *per se*, but rather on their relative weights with respective to each other. There are some implication deriving from this decision. For instance, in a compositional approach, a country with equally spread out and yet low activations is substantially identical to another one with equally

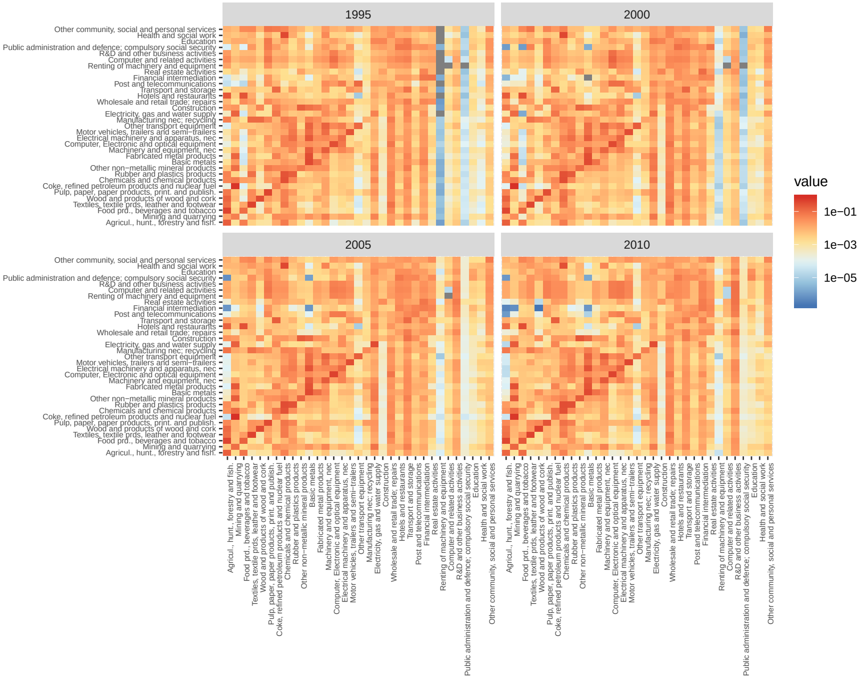

Figure 5.1: Heatmap of the log-technical coefficients for China in 1995, 2000, 2005, 2010.

spread out but higher activations. To give a geometric intuition to this assumption, we can imagine to only have two archetypes participating to the definition of an economy. In this imaginary framework, countries will be spread around the positive orthant of $\mathbb{R}^2$. We can see that the distance between countries lying on the same line starting from the origin will be null, whereas it will be maximised when countries belong respectively to the abscissae and to the ordinates. A motivation for adopting this approach is represented by the fact that we may not be particularly interested in the particular magnitude of the components, but rather on the resulting mix. To think of it with a metaphor, we can imagine various recipes to fill a glass: the taste of the outcome is not related to the capacity of the glass being filled. From a computational point of view, the dissimilarity matrix based on the Aitchison distance (4.14) was computed once again in R, using ([82]). The other approach is, instead, a standard Euclidean approach. In this case, not only are different level of activations relevant, but also their activation across various components plays a role. It is therefore not necessary that two points lying on the same line be closest than two lying on the axes of the positive orthant of the real numbers. The advantage of this approach is therefore a focus on the specific and absolute magnitude of the activations, at the expense of losing the focus on the relative values.

Clustering was attempted considering cross-sectional datasets. In other words, data were sliced according to a specific year. Moreover,

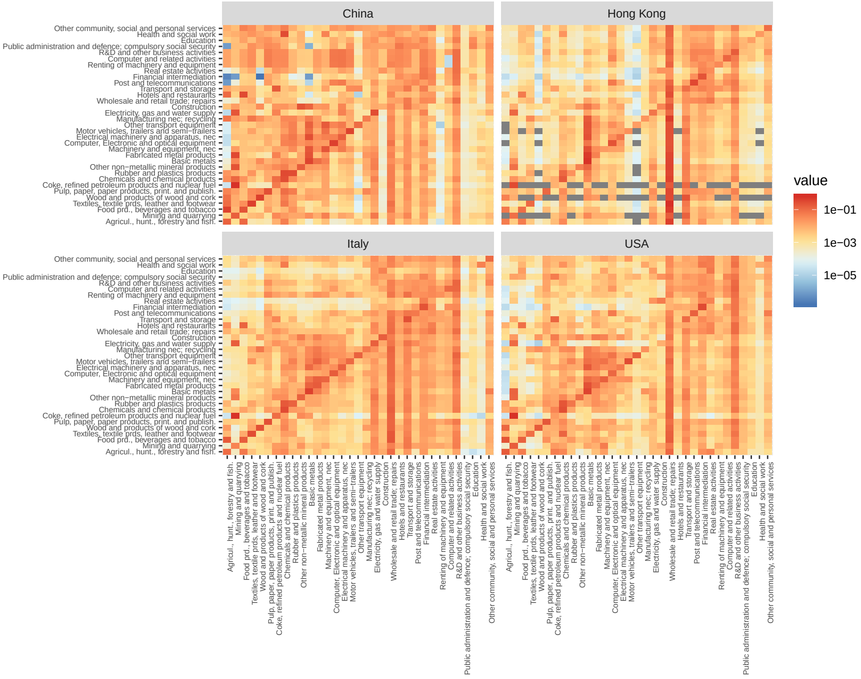

Figure 5.2: Heatmap of the log-technical coefficients for China, Hong Kong, Italy and the USA in 2011.

the analysis was conducted considering three different geographic zones: the EU, the OECD and the entire dataset. The reason behind the choice of considering increasingly bigger subsets of the original dataset was to maintain complexity lower in a first moment. It is important to notice that, by considering cross-sectional datasets, we are abstracting from any kind of time dependency between different years and refraining from making any consideration whatsoever regarding the dynamics of the phenomena we will analyse.

We remark that cross-sectional datasets were built by vectorising technical coefficients matrix and juxtaposing them. Fixing a year, the dataset $\mathbf{X}$ lied therefore in a space of dimension $(33^2 \times n)$, where n depends on the specific zone under scrutiny, namely $n = 28$ when we considered the EU, $n = 34$ when we considered countries member of the OECD and $n = 64$ when we considered the whole World.

In this work, we decided to focus on two years in particular, namely 1995 and 2011, to assess the validity of the method proposed and its limitations, considering three different groups of countries: the EU, OECD members and the entire World.

5.2 analysis of the eu cross-sectional datasets

We now present the results of the analysis for the EU cross-sectional datasets for the years 1995 and 2011. At first, we will present and comment the archetypes our analysis yielded. Afterwards, the analysis of the subsequent clusterings will be performed.

The EU datasets comprises twenty-eight countries: Austria, Belgium, Bulgaria, Croatia, Cyprus[3], Czech Republic, Denmark, Estonia, Finland, Germany, Greece, Hungary, Ireland, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden and the United Kingdom[4].

5.2.1 *Archetypes for the EU cross-sectional datasets*

The first step in this analysis consisted in the estimation of the rank k of the factorisation. We eventually resolved to set $k = 3$.

The results yielded from the singular value decomposition for the year 1995 and 2011 are displayed in Figures 5.3 and 5.4 respectively. Both plots show a significant reduction of the singular values as soon as we consider those succeeding the first one. This indicate an initial rank estimation in-between $k = 2$ and $k = 5$. We wish to balance two tendencies: the smaller the number of components we consider, the simpler the model, the greater the number, the better its capacity to reconstruct the original matrix. For this reason, we are looking for an “elbow" in the plot: we want to justify complexity with performance, but keep the model as simple as possible.

The second technique employed consisted in computing the Silhouette score to diagnose the efficacy of k-means clusterings with k growing from 2 to 5. Again, we are willing to find the optimum number of clusters able to summarise sectors well enough to be of interest and yet only considering the fundamental aggregations. We report the full Silhouette plots for $k = 3$ in Figures 5.5 and 5.6.

[3] As stated at https://www.oecd.org/sti/ind/input-outputtables.htm, there exist different positions regarding Cyprus. The European Union Member states of the OECD and the EU note that the Republic of Cyprus is recognised by all members of the UN with the exception of Turkey. The information contained in the dataset therefore relates to the area under the effective control of the Government of the Republic of Cyprus. Turkey claims that there exists no single authority representing both Turkish and Greek Cypriot people living on the island. Turkey recognises the Turkish Republic of Northern Cyprus. Turkey claims that until an equitable and lasting solution will be found in the context of the UN, it shall preserve its position regarding the “Cyprus issue".

[4] On 23/07/2016 citizens of the United Kingdom voted to leave the EU. On 29/03/2017 the United Kingdom officially notified the European Council of its intention of leaving the EU by triggering Article 50 of the Lisbon Treaty (https://europa.eu/european-union/about-eu/countries/member-countries/unitedkingdom_en). At the time of this writing, however, the UK remains an effective members of the EU.

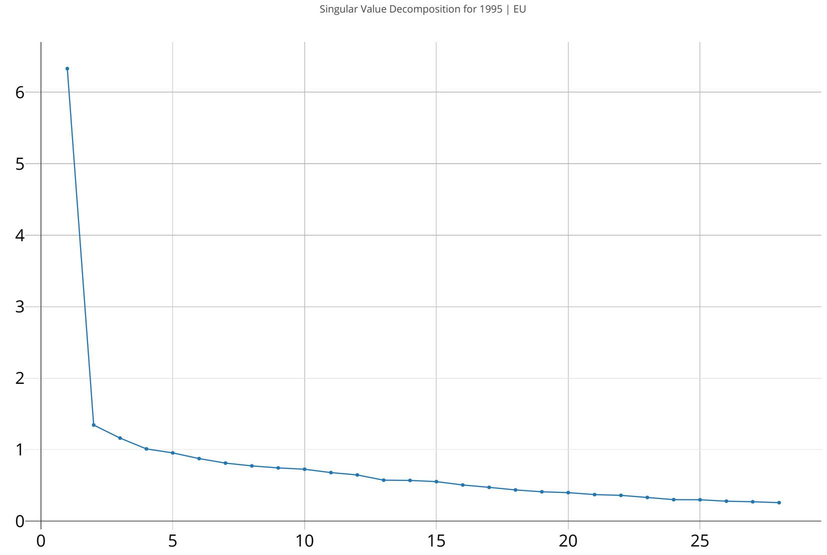

Figure 5.3: Singular Values of the SVD computed for the EU cross-sectional dataset in 1995.

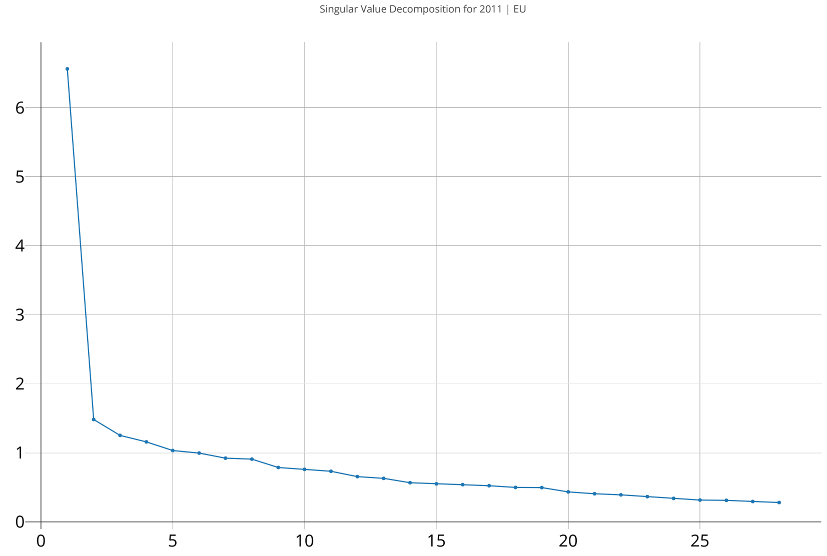

Figure 5.4: Singular Values of the SVD computed for the EU cross-sectional dataset in 2011.

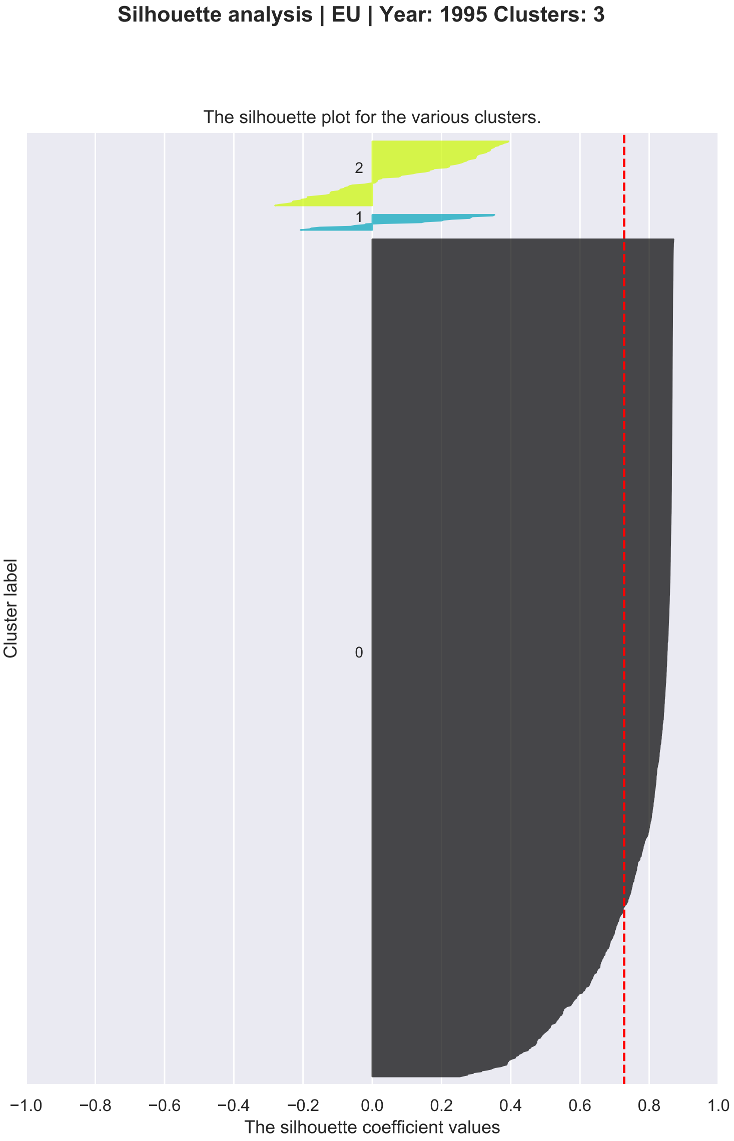

Figure 5.5: Silhouette profiles of the k-means algorithm with 3 clusters for the cross-sectional EU dataset for the year 1995.

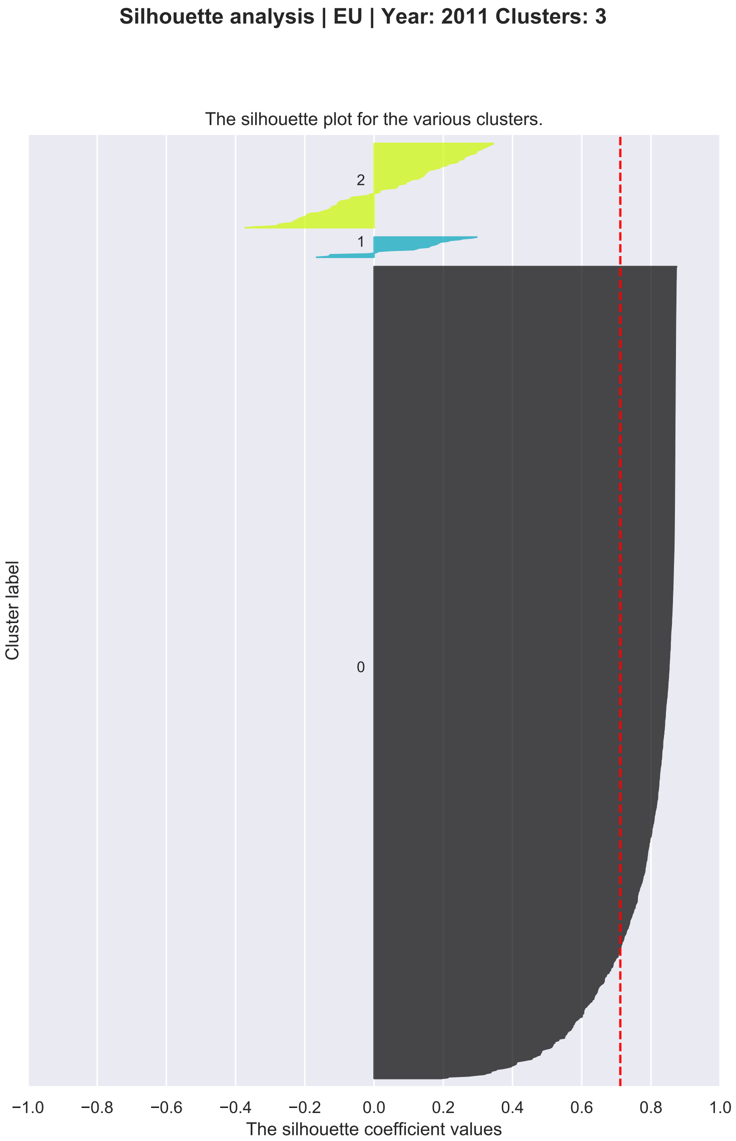

Figure 5.6: Silhouette profiles of the k-means algorithm with 3 clusters for the cross-sectional EU dataset for the year 2011.

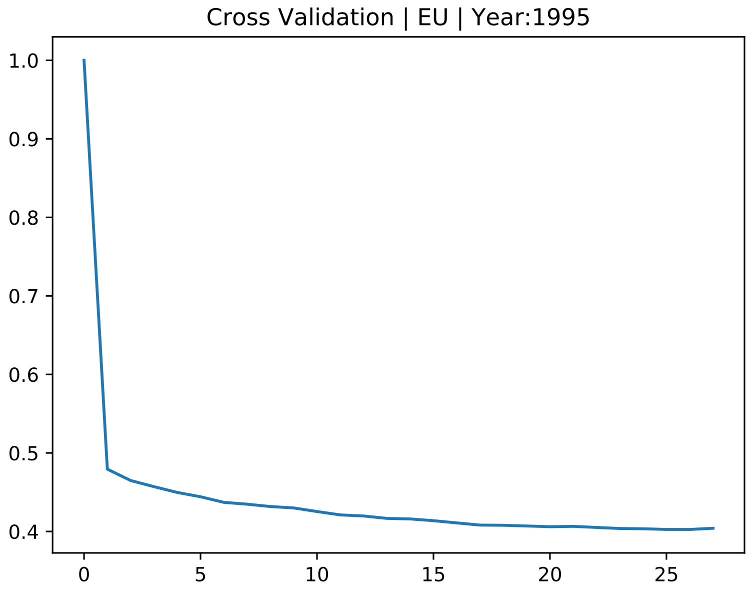

Figure 5.7: Cross-Validation for the cross-sectional EU dataset for the year 1995.

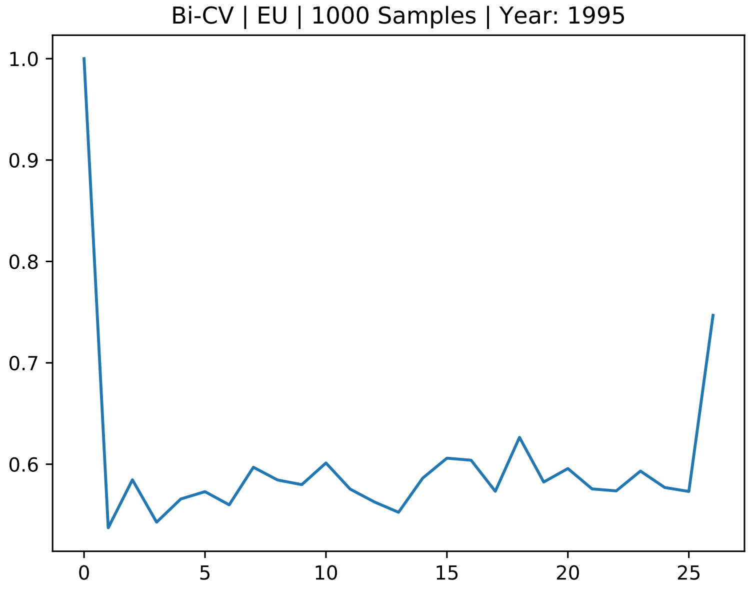

Figure 5.8: Bi-Cross-Validation with simple residuals for the cross-sectional EU dataset for the year 1995. We considered (3×3) random submatrices and, for each rank, averaged the results of 1000 random trials. Data are normalised considering the MSE of a rank 0 approximation, *i.e.* the Frobenius norm of the sampled matrices alone.

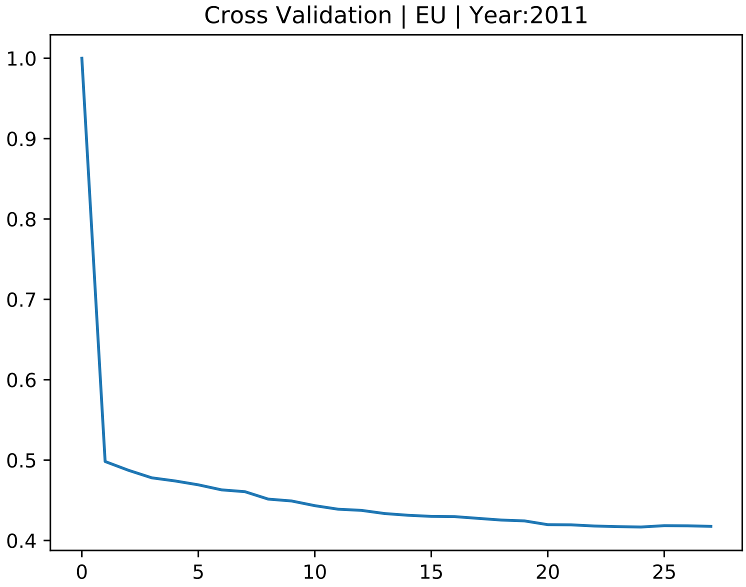

Figure 5.9: Cross-Validation for the cross-sectional EU dataset for the year 2011.

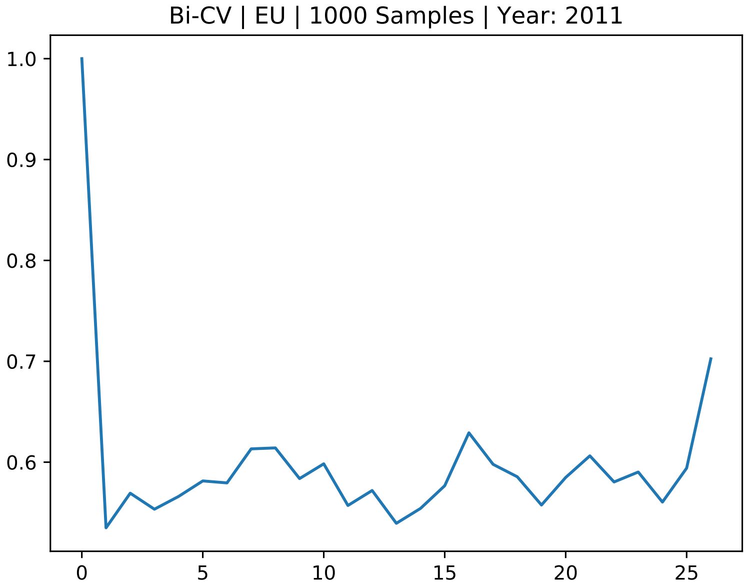

Figure 5.10: Bi-Cross-Validation with simple residuals for the cross-sectional EU dataset for the year 2011. We considered (3×3) random submatrices and, for each rank, averaged the results of 1000 random trials. Data are normalised considering the MSE of a rank 0 approximation, *i.e.* the Frobenius norm of the sampled matrices alone.

In this case, we are interested in considering the entire Silhouette profile arising from the k-means clustering because our diagnostic should also be careful as far as the homogeneity of clusters is concerned, given the pivotal importance of rank estimation. From the Silhouette profiles, we find confirmation of the results of the singular value decomposition. Moreover, it is interesting to notice a general tendency of many sectors to be grouped together in a single cluster. This suggests some structural similarities between many inter-sectoral relationship across various countries, both in 1995 and 2011. Such tendency might be due to the fact that our datasets are highly aggregated and only comprise thirty-three sectors: this might have the effect of cancelling out differences between countries and levelling data to the point where it is difficult to differentiate between countries, especially when we are considering technical coefficients matrix, which do not account for magnitudes. This is a major trade-off when dealing with different countries at the same time, as the homogenisation of data comes at the expenses of detail. Cross-validation and Bi-Cross-Validation were then computed, as described in Algorithms 3.1 and 3.3. We briefly remark that we decided to implement the algorithm for Bi-Cross-Validation that employs simple residuals for computational reasons, despite its formal inelegance.

The results for the Cross-validation and Bi-Cross-Validation are reported in Figures 5.7 and 5.8 for the year 1995 and in Figures 5.9 and 5.10 for the year 2011. The results of the procedure suggest that the rank should be in-between $k = 2$ and $k = 5$. However, we would expect at least the plot of the Bi-Cross-Validation to be less noisy and to increase more decisively as k tends to 27, the number of the countries in the EU dataset. The fact that this does not happen suggest that data are not easily divisible into clusters. The final rank estimation of $k = 3$ was made after comparing the results of the techniques insofar presented and after discussing the decision with a panel of experts in economics and industrial engineering to validate it.

We now present the results of the decomposition. As far as the year 1995 is concerned, we report the three archetypes we obtain according to their shock centrality in Figures 5.11, 5.12 and 5.13. We also report the networks induced by the archetypes in Figures 5.14, 5.15 and 5.13, where the dimension of the nodes representing the sectors is proportional to the shock-centrality.

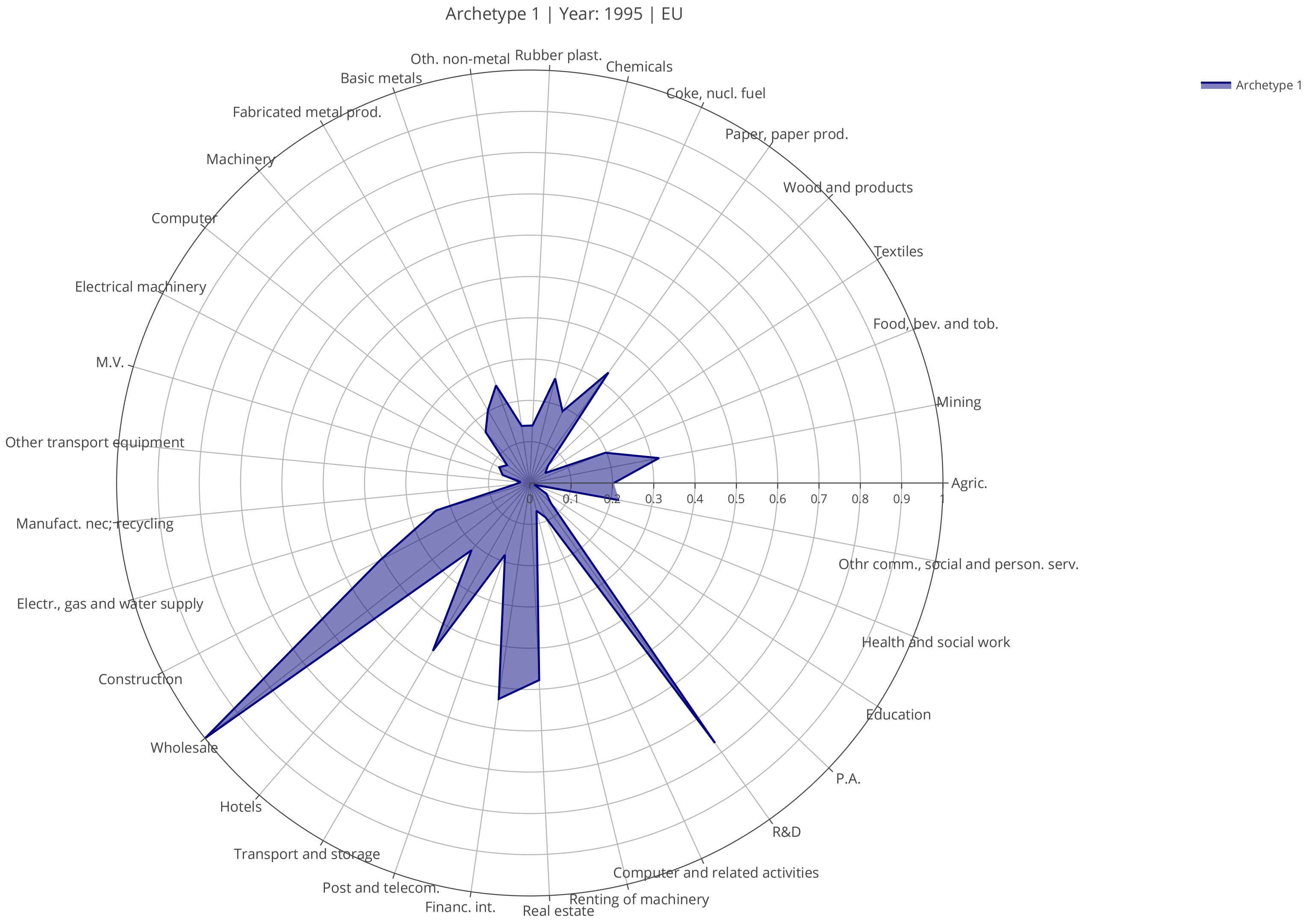

Figure 5.11: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional EU dataset for the year 1995.

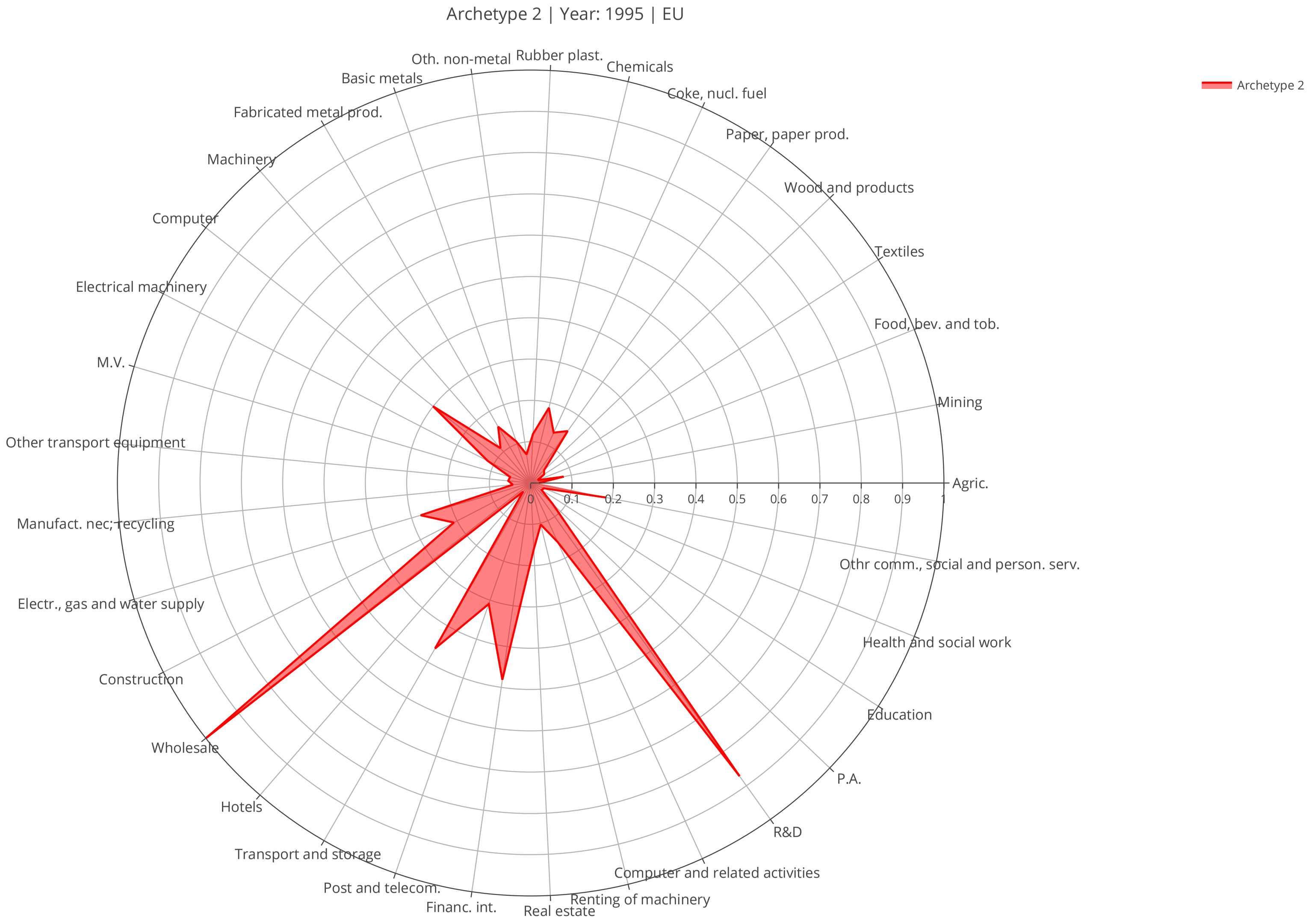

Figure 5.12: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional EU dataset for the year 1995.

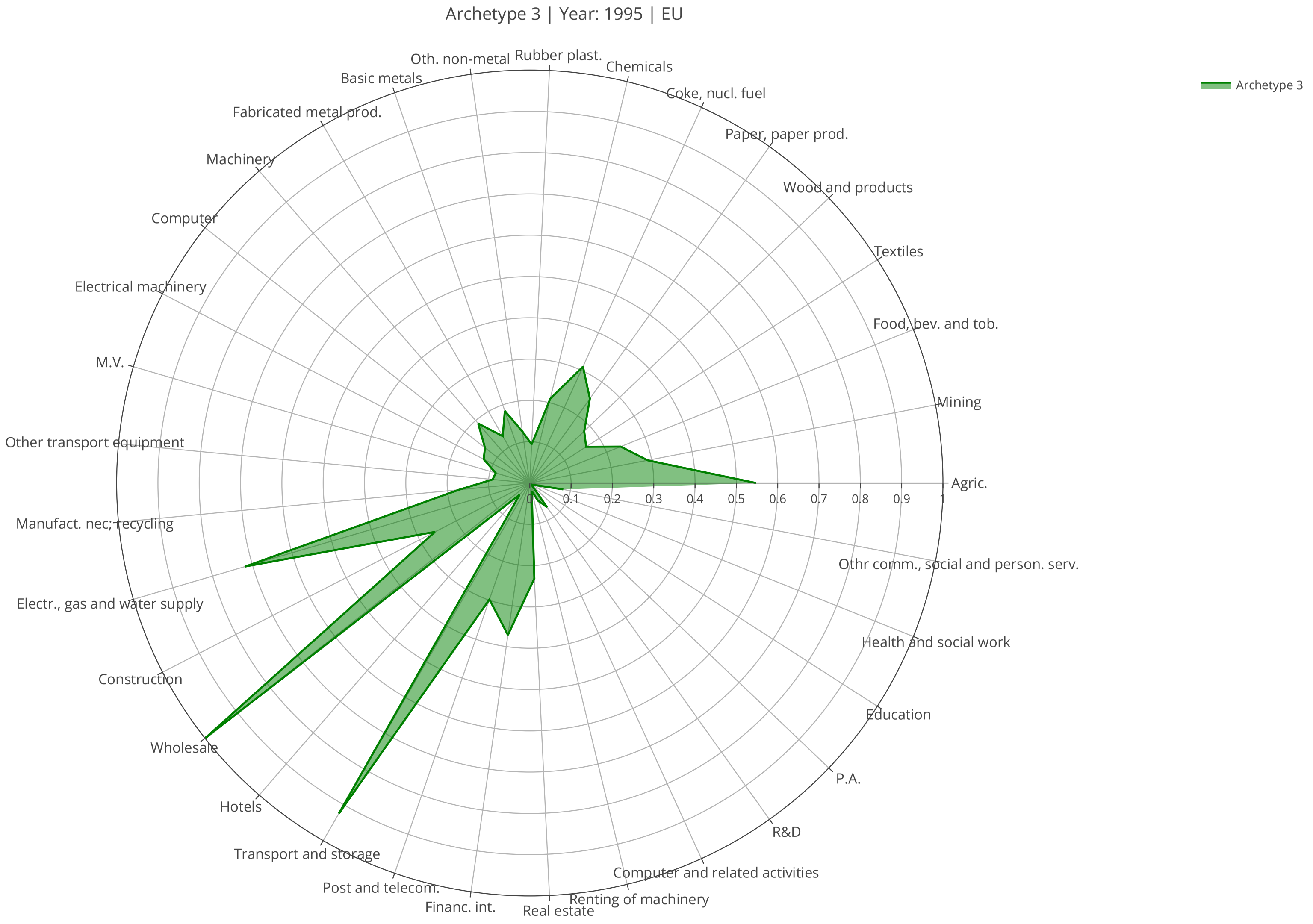

Figure 5.13: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional EU dataset for the year 1995.

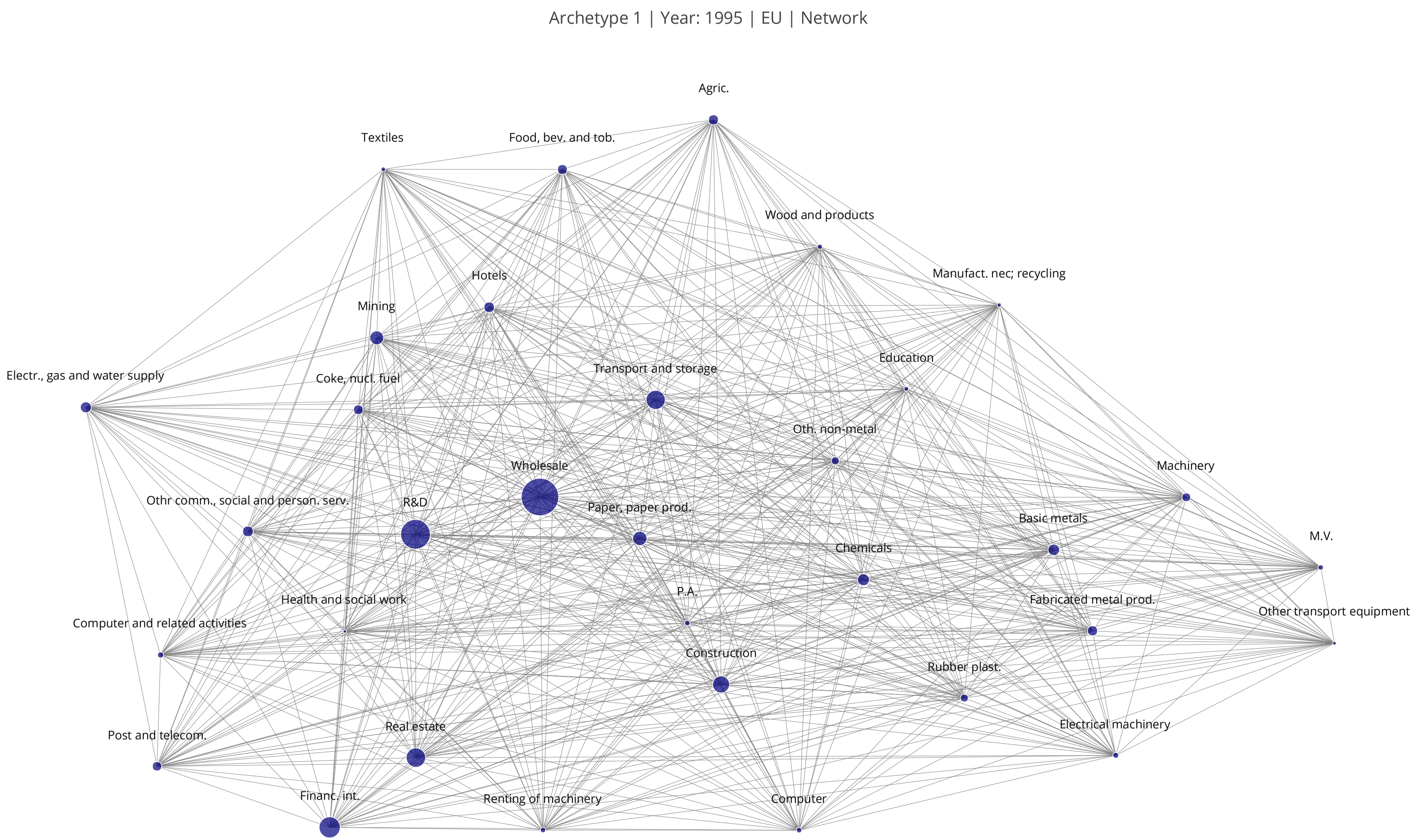

Figure 5.14: Network induced by the first archetype of the cross-sectional EU dataset for the year 1995. The dimension of the nodes is proportional to their shock-centrality.

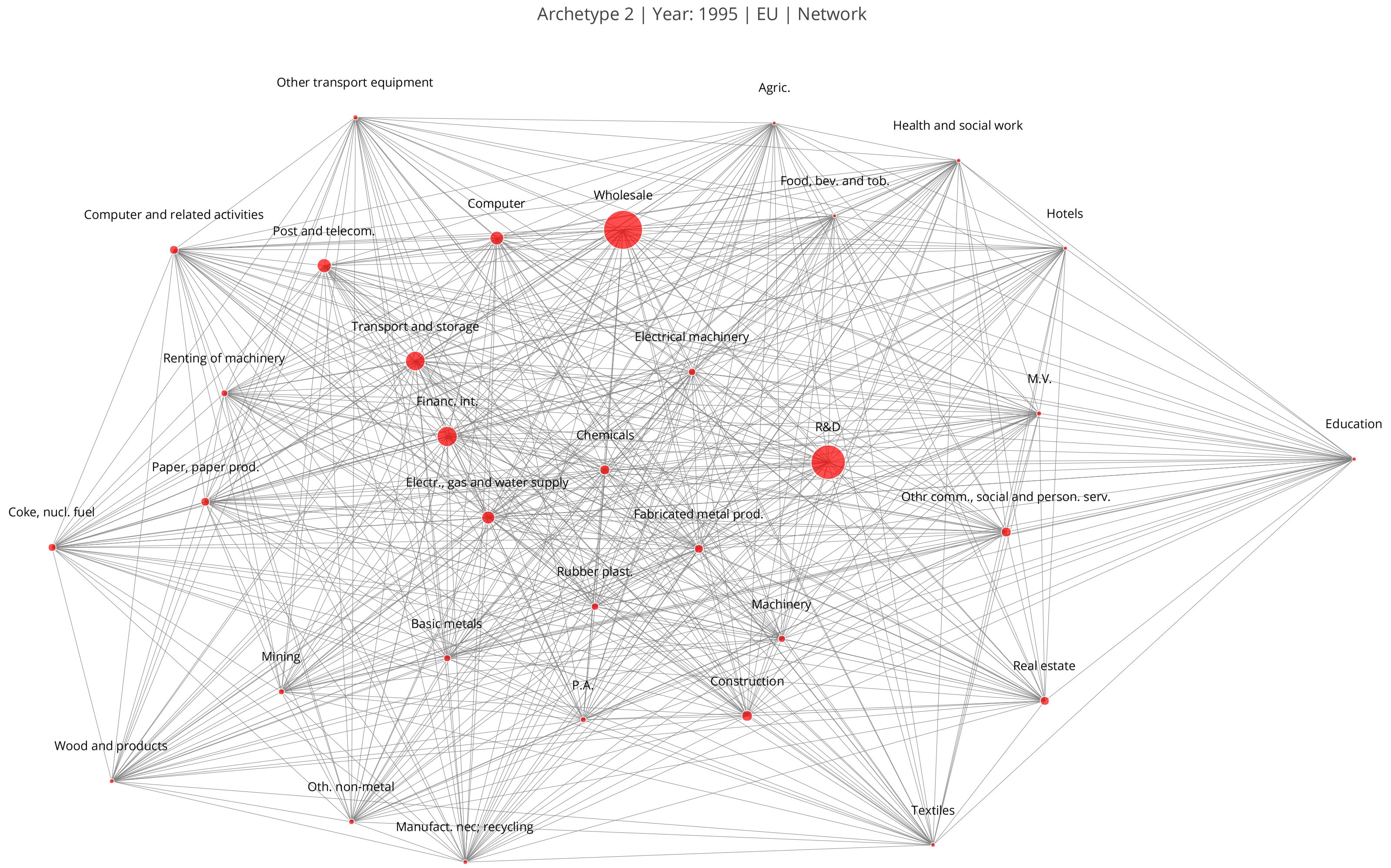

Figure 5.15: Network induced by the second archetype of the cross-sectional EU dataset for the year 1995. The dimension of the nodes is proportional to their shock-centrality.

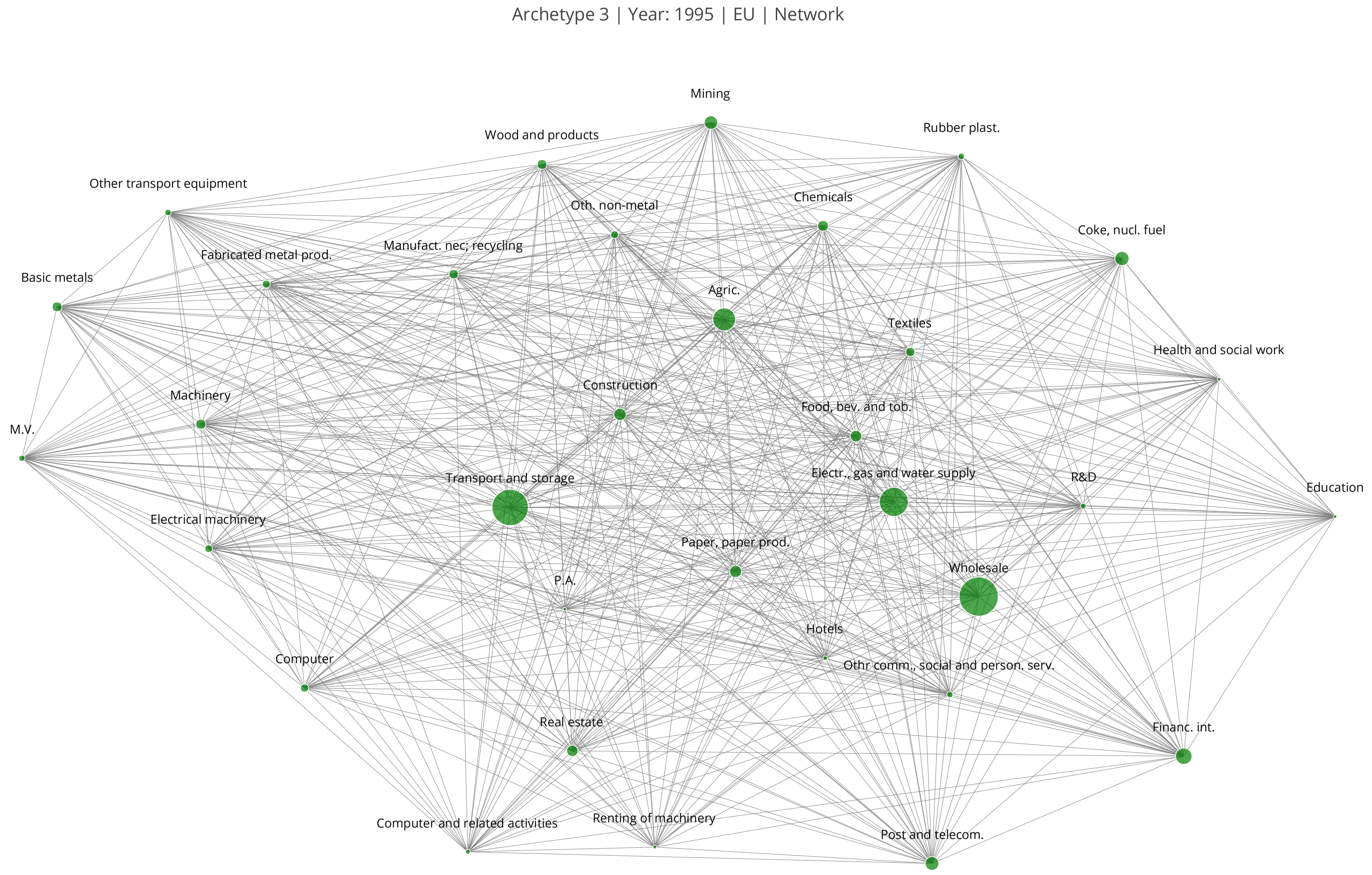

Figure 5.16: Network induced by the third archetype of the cross-sectional EU dataset for the year 1995. The dimension of the nodes is proportional to their shock-centrality.

The first thing we notice is that the “Wholesale" sector appears to be central in all the archetypes. This is reasonable: shocks are very likely to reach this sector in a short period of time. Another sector which appears to be easily reached by shocks is that of “R&D", *research and development*. This may be justified by considering that some firms may find difficulties in investing in innovation whenever a shock occurs. Two other sectors tend to present a high vertex centrality: financial intermediation and transports. This seems to suggest that the centrality measure can somewhat capture the most exposed sectors. We can seek further confirmation of this by noticing those sectors that are more peripheral. We consider, as an example, the health and the education sectors. It should not come as a surprise that such sectors tend to be stable under shocks, especially in EU countries, where the organisation of both is usually managed by states.

We notice that the construction sector plays a role in this first archetype. “Construction" is usually considered a proxy to understand how well an economy is scoring, as well as “Transports". We could therefore interpret this first pattern as those part of an economy relevant in assessing growth and the relevance of the tertiary sector to the economy, as also “Real Estate" and “Financial Intermediation" show some degree of centrality. The second archetype yields a more complex economic interpretation. We see activations in sectors connected to the industrial sectors of the economy and the energetic supply one. This pattern might then be considered as connected to the basic industrial structure of an economic system. The third archetype shows a uniform activation across the whole spectrum of the agricultural and food sectors, together with a centrality of the “Energy, gas and water supply" sector. We find a coherence in this mutual activations. This archetype might be capturing the agricultural part of the system.

We repeat the analysis for the year 2011. We report the archetypes in Figures 5.17, 5.18 and 5.19. The networks induced by the archetypes are reported in Figures 5.20, 5.21 and 5.22. The interpretation for 2011 is less straightforward, possibly because the situation it tries to describe is rather different from that of 1995. We notice that the first archetype shows the usual centrality of the “Wholesale" and “R&D" sectors. We also notice that “Financial intermediation", “Transport and Storage", “Mining" and “Coke and nuclear fuel" are relevant. The second archetype again points to the basic industrial infrastructure. The third archetype appears to be vulnerable in the usual sectors of “Wholesale" and “R&D", together with “Real estate". While we would expect some degree of centrality for all the sectors above, it is not immediate to understand the meaning of this particular repartition. The model might be capturing some structural properties of the system that are not immediately visible or it might be the case that different parts have been distributed in unexpected ways.

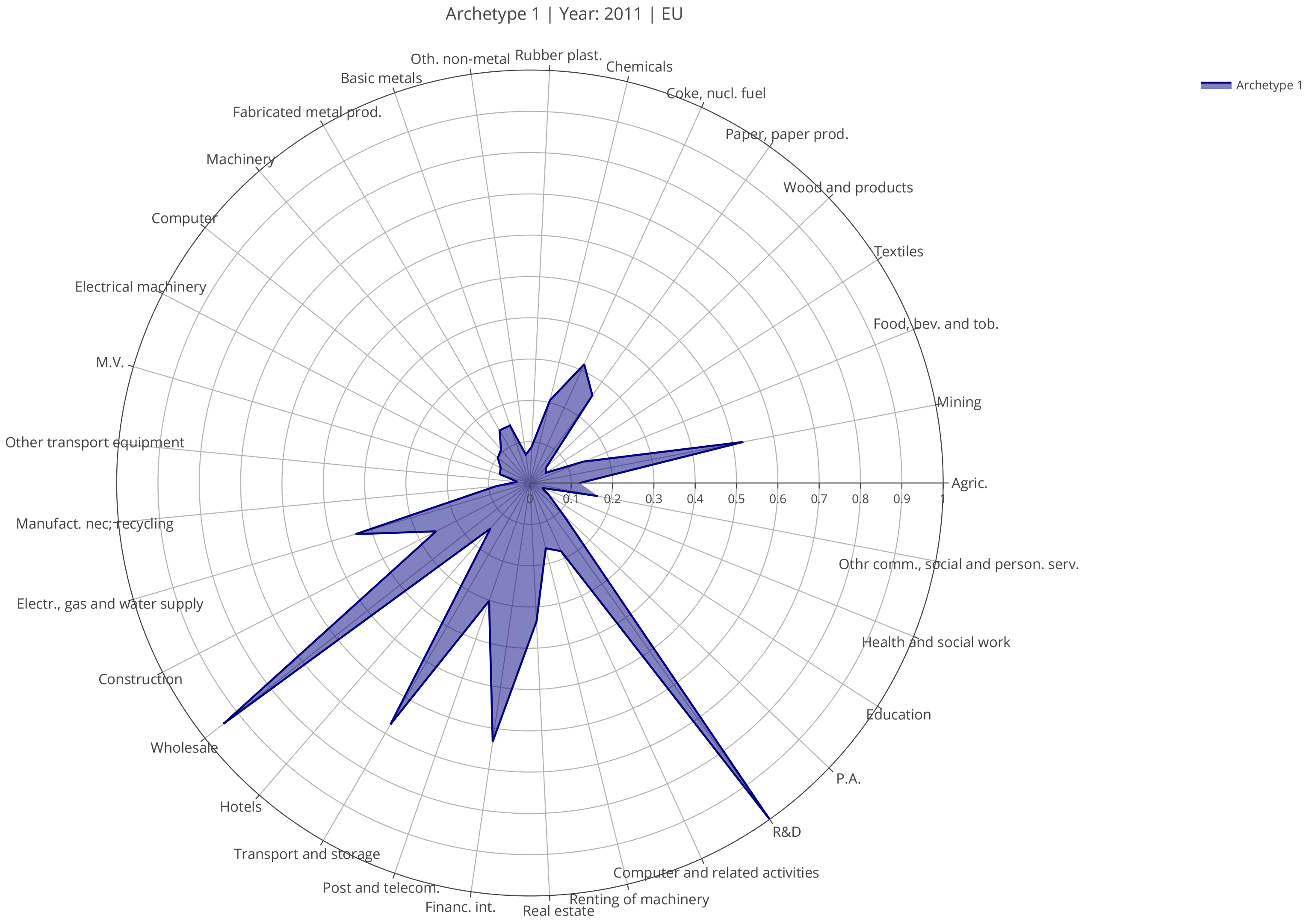

Figure 5.17: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional EU dataset for the year 2011.

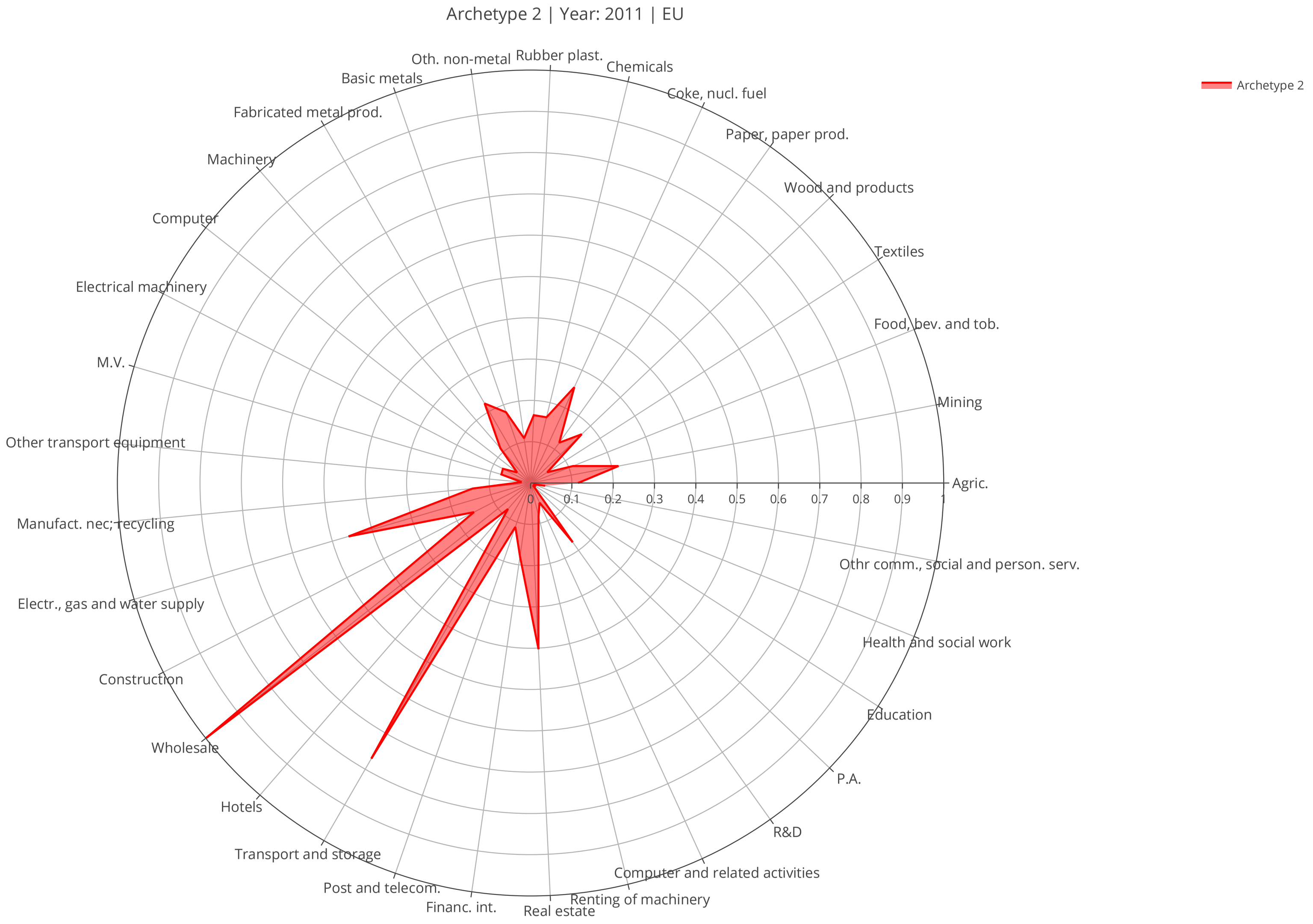

Figure 5.18: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional EU dataset for the year 2011.

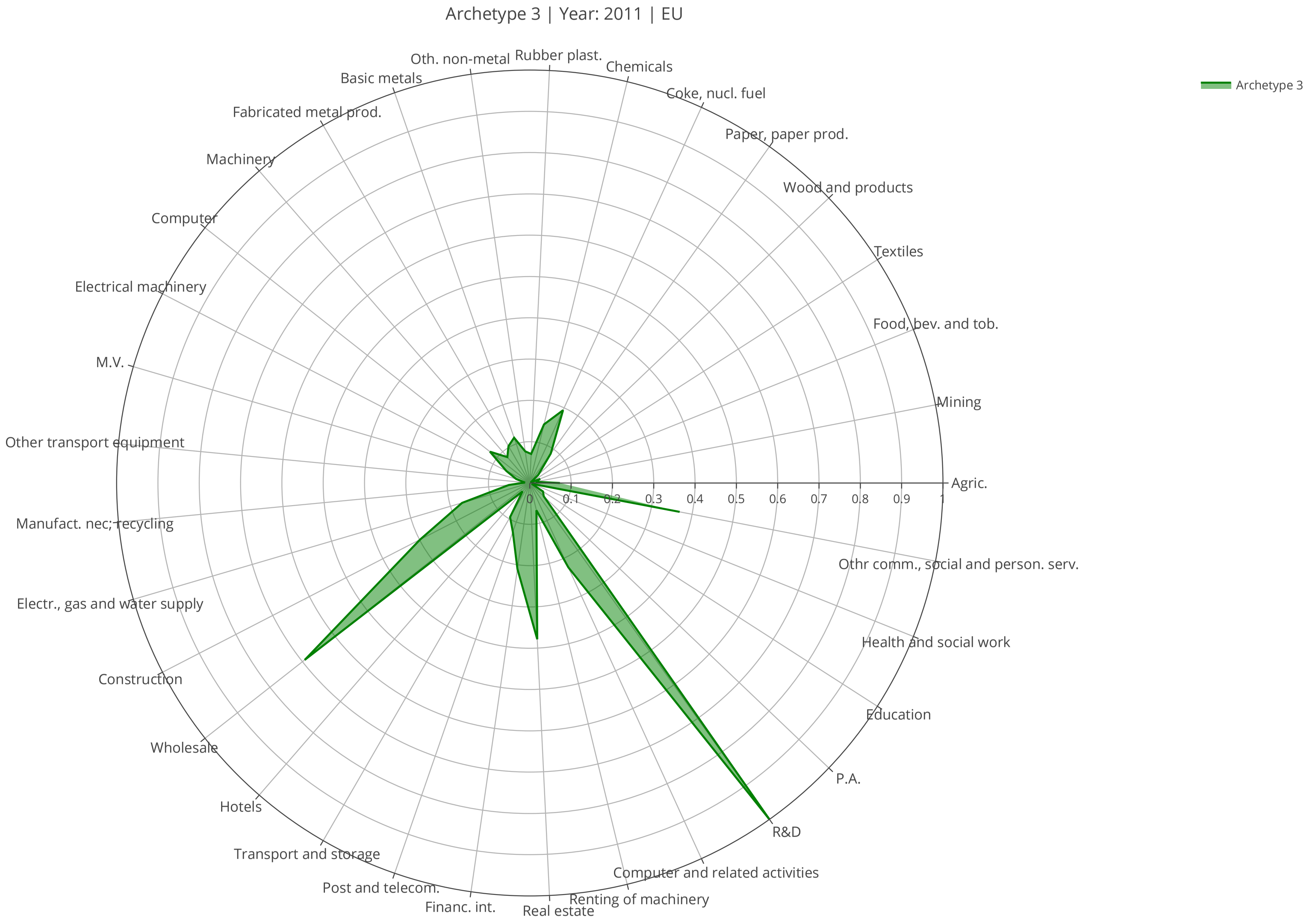

Figure 5.19: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional EU dataset for the year 2011.

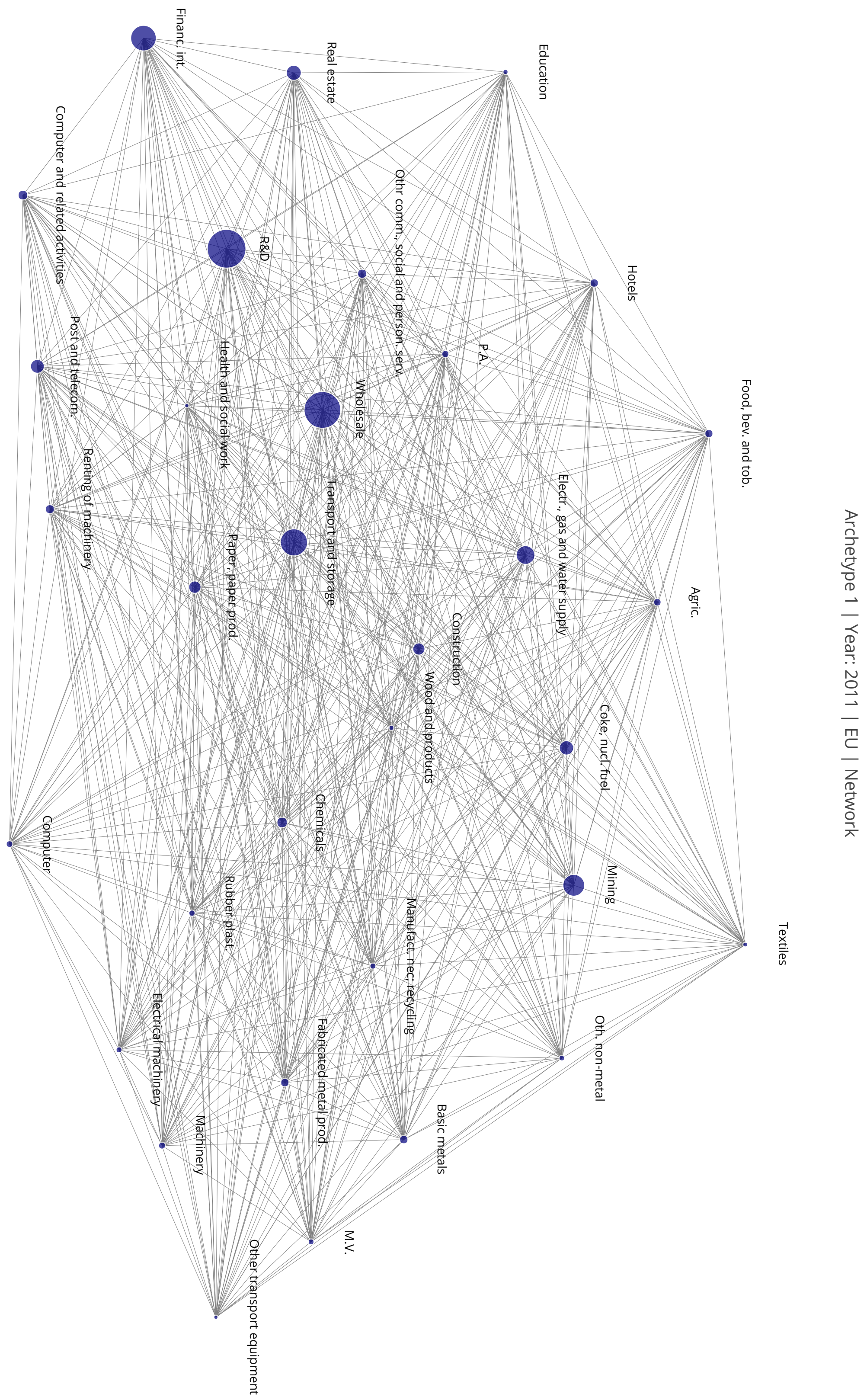

Figure 5.20: Network induced by the first archetype of the cross-sectional EU dataset for the year 2011. The dimension of the nodes is proportional to their shock-centrality.

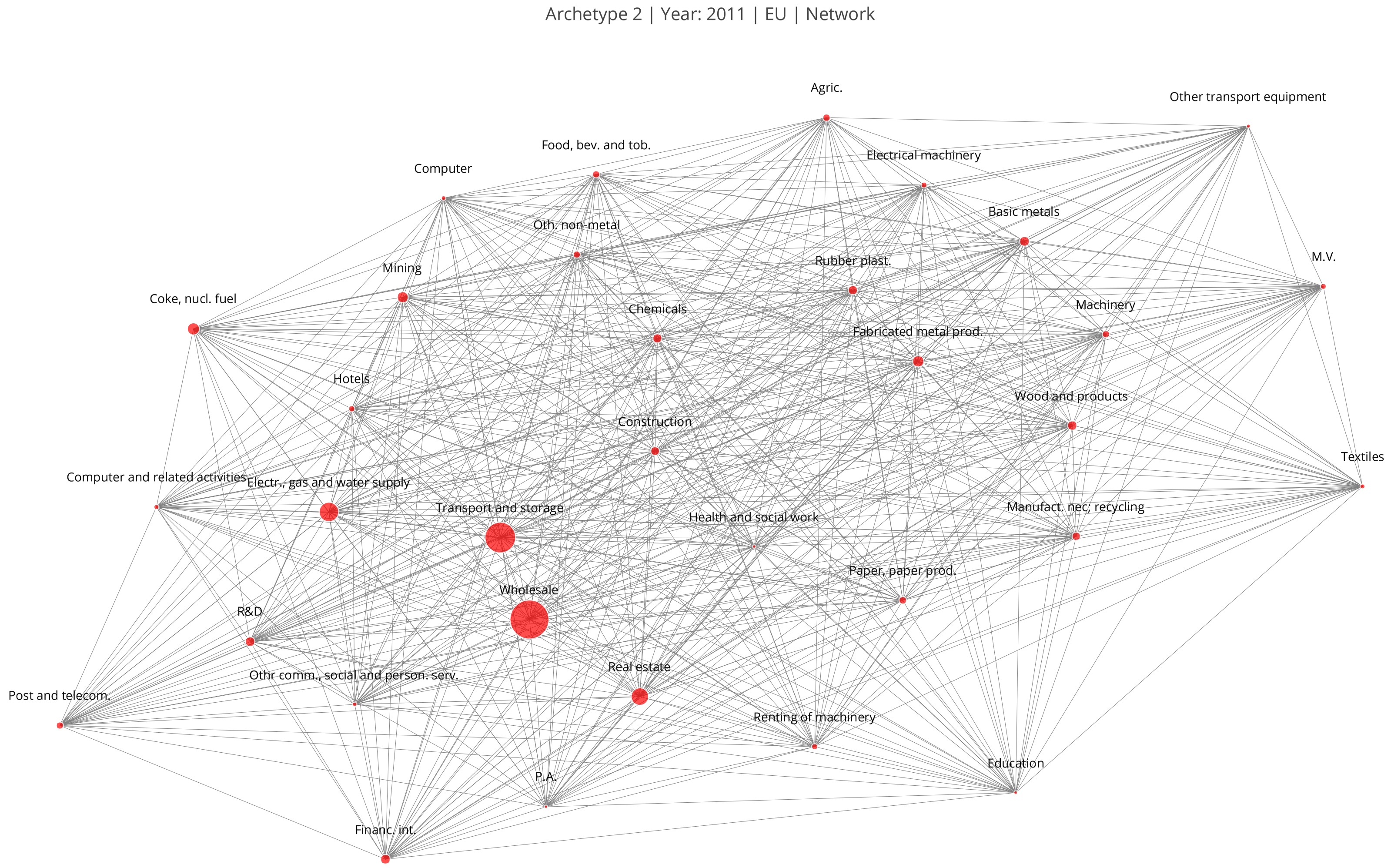

Figure 5.21: Network induced by the second archetype of the cross-sectional EU dataset for the year 2011. The dimension of the nodes is proportional to their shock-centrality.

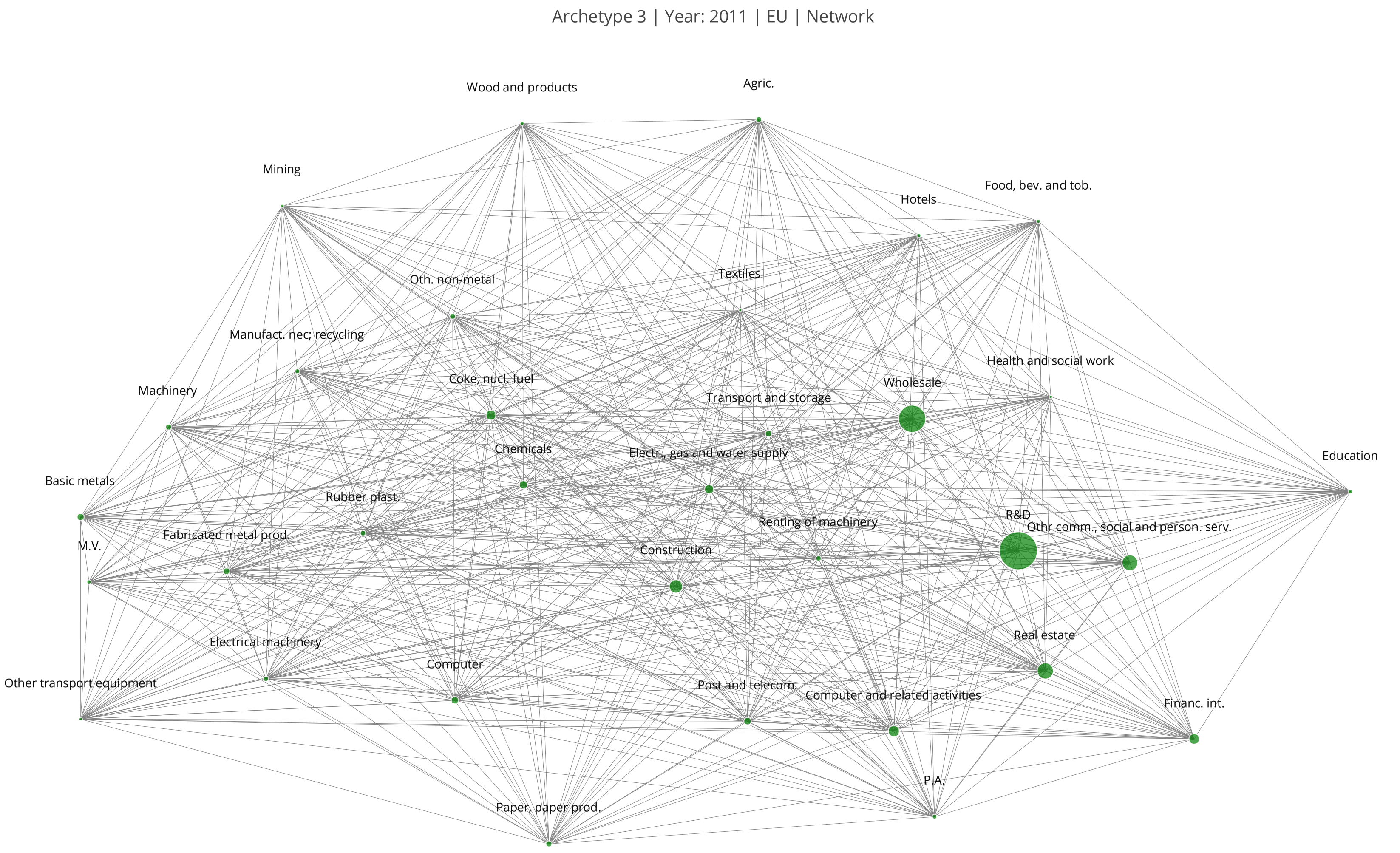

Figure 5.22: Network induced by the third archetype of the cross-sectional EU dataset for the year 2011. The dimension of the nodes is proportional to their shock-centrality.

5.2.2 *Clustering for the cross-sectional EU datasets*

We now focus on the year 1995. We report in Figure 5.23 the heatmap with the activations for the countries in the EU. Moreover, to give a geographical sense of the various clusterings, we report the maps obtained considering both the Aitchison and the Euclidean distances in Figures 5.24 and 5.25. Let us consider the closure of the activations for the centroids obtained with the Aitchison metric of the year 1995, presented in Table 5.1.

First of all, as suggested by the Silhouette coefficient, we notice that the clustering separates between four distinct types of countries. The cluster of countries represented by Spain presents a balance (although somewhat skewed towards the first two components) of the three archetypes. This suggests that countries belonging to this first cluster should be countries with an acceptable growth status, a good industrial infrastructure and a good stability. Spain, the Czech Republic, Finland, Denmark, Austria, Slovenia, Portugal, Italy, Hungary, Sweden, the United Kingdom, Bulgaria, Estonia and Ireland all belong to this cluster. As it was already stated, interpretation of clusters is not straightforward, as each country has its own particular history and characteristics and finding analogies is not intuitive. For instance, in 1995 the Czech Republic was in the middle of its transition from a state-planned economy to a free-market economy and in 1997 it would be shuttered by an economic crisis. Italy was also emerging from its major economic crisis of 1992 at the time, despite its GDP per capita being greater than that of the United Kingdom. The second cluster, represented by Greece, contains Germany, Cyprus, Belgium, Luxembourg and the Netherlands. In this case, we notice that the third archetype does not play a role. It is actually the case that the above countries all experienced similar levels of growth in 1995. The third cluster only contains Romania and Croatia. In this case, we would expect to obtain countries experiencing growth but with some stability problems. Again, it might be possible to consider a connection between this cluster and the effects Romania was suffering from the austerity of the previous years and the same can be said from Croatia. To conclude, we notice that the last and fourth cluster comprises two Baltic republics: Latvia and Lithuania, and Malta. In this case, the second and third archetypes are very relevant. For members of this cluster, we would expect a somewhat stable basic industrial structure, despite low-growth. Again, to fully understand the predictive power of such clustering, a comparison with the histories of such countries is needed. It is immediate to notice, however, that Latvia and Lithuania were experiencing the transition towards market-based economies.

If we consider the Euclidean clustering, reported in Table 5.2, we obtain three clusters. The first clusters comprises Denmark, Finland, Spain, Sweden, Slovakia, Austria, the Netherlands, the United King-

Medoid	Archetype 1	Archetype 2	Archetype 3
Spain	0.575	0.306	0.119
Greece	0.617	0.383	0.000
Romania	0.587	0.00	0.413
Latvia	0.000	0.396	0.604

Table 5.1: Closure of the activations for the centroids obtained via a k-medoids clustering in the Aitchison geometry for cross-sectional EU dataset of 1995.

Medoid	Archetype 1	Archetype 2	Archetype 3
Spain	0.462	0.245	0.095
Luxembourg	0.207	0.507	0.000
Lithuania	0.000	0.342	0.522

Table 5.2: Activations for the centroids obtained via a k-medoids clustering in the Euclidean geometry for cross-sectional EU dataset of 1995.

dom, Czech Republic, Portugal, Cyprus, Greece, Poland, Romania, Italy, Slovenia, Croatia, Germany and Bulgaria. In this case, the analysis is less intuitive because of the adopted metric. We can assume that countries belonging to this cluster all have similar values of activation. However, we should also notice that the lower Silhouette coefficient suggests a higher heterogeneity. This is confirmed by the second cluster, containing Malta, Ireland, Luxembourg, Hungary, Belgium and France: France has a negative Silhouette coefficient in this case. The third cluster contains the three Baltic Republic: Estonia, Lithuania and Latvia. This suggests that homogenous country are kept together.

Shifting our attention to 2011, we report the resulting activation heatmap in Figure 5.26. Moreover, we report the maps obtained considering both the Aitchison and the Euclidean distances. It is striking to see that in the compositional case we obtain seven clusters as opposed to the Euclidean one, resulting in three. We report in Table 5.3 the centroid and its activations for every cluster. In the compositional case, we obtain a first cluster composed by Austria, France, Ireland, the Netherlands and the United Kingdom. Those countries are spread out on the first and the third archetype. We expect them to be susceptible to shock affecting the “Wholesale", “R&D" and “Real Estate" sectors. The second cluster is composed by the Czech Republic, Belgium, Hungary, Estonia, Finland, Germany, Greece, Croatia, Italy, Latvia, Poland, Portugal, Slovakia, Spain and Sweden. In this case, we notice that the first archetype matters twice as much as the other two,

Medoid	Archetype 1	Archetype 2	Archetype 3
Austria	0.604	0.000	0.396
Czech Republic	0.490	0.231	0.279
Romania	0.568	0.432	0.000
Cyprus	0.000	0.945	0.055
Denmark	1.000	0.000	0.000
Lithuania	0.000	1.000	0.000
Luxembourg	0.000	0.000	1.000

Table 5.3: Closure of the activations for the centroids obtained via a k-medoids clustering in the Aitchison geometry for cross-sectional EU dataset of 2011.

Medoid	Archetype 1	Archetype 2	Archetype 3
Spain	0.478	0.098	0.189
Cyprus	0.000	0.927	0.054
Malta	0.000	0.056	0.672

Table 5.4: Activations for the centroids obtained via a k-medoids clustering in the Euclidean geometry for cross-sectional EU dataset of 2011.

taken individually. This suggests a predominance of the sectors highlighted in the first archetype, with some exposure on the industrial side. The third cluster comprises Bulgaria and Romania: in this case the vulnerability of industrial sector should be greater. To conclude, Cyprus is grouped with Malta, skewed towards the second archetype. The final three medoids are lone members of their respective cluster: Denmark, Lithuania and Luxembourg.

We notice that our model show an economic situation for 2011 which appears more difficult to analyse. This may be due to the actual economic moment at which data were registered. We also notice a tendency of grouping many countries in a larger cluster, whose components are more spread out. In Table 5.4, we report the medoids and their coordinates for the Euclidean case. The first cluster contains Austria, Belgium, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, the Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain and Sweden. The second cluster comprises Cyprus, Latvia and Lithuania. The third and last one is composed by Malta and Luxembourg. Such a remarkable difference is due to the metric and the different focus each one adopts. The Aitchison metric seems more revealing of the underlying structure of the clustering with respect to the Euclidean one.

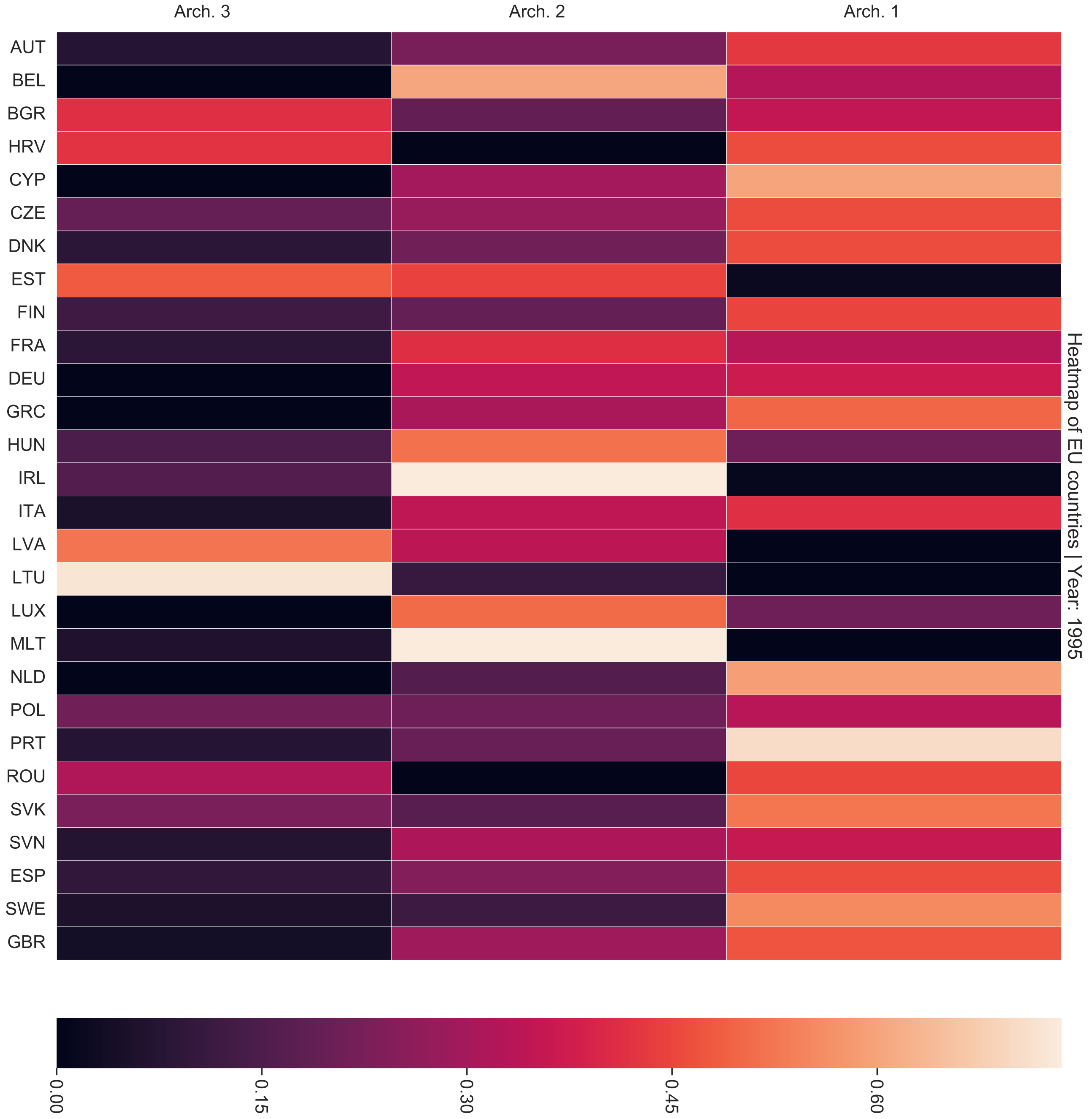

Figure 5.23: Heatmap of the activations of the countries in the cross-sectional EU dataset for the year 1995.

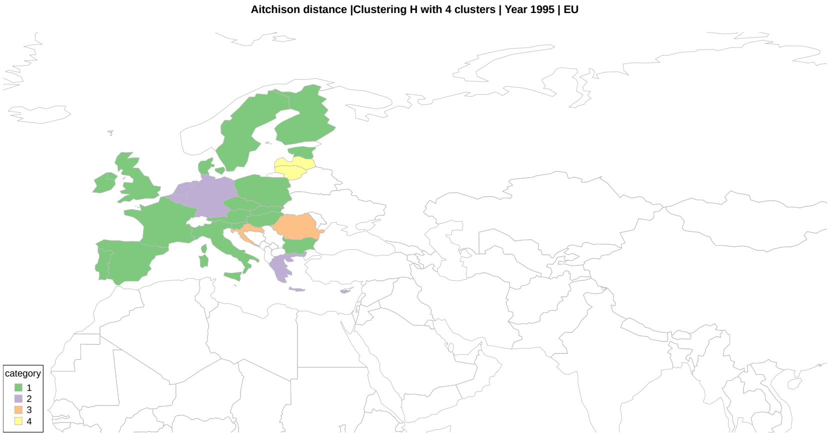
Aitchison distance |Clustering H with 4 clusters | Year 1995 | EU
category
1
2
3
4

Figure 5.24: Clustering obtained considering the cross-sectional EU dataset for the year 1995, via k-medoids and the Aitchison metric, with the number of clusters induced by the best Silhouette score.

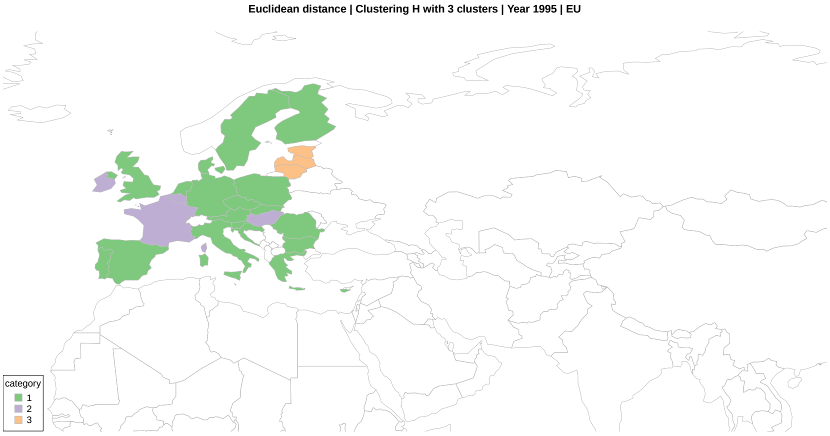
Euclidean distance | Clustering H with 3 clusters | Year 1995 | EU
category
1
2
3

Figure 5.25: Clustering obtained considering the cross-sectional EU dataset for the year 1995, via k-medoids and Euclidean distance, with the number of clusters induced by the best Silhouette score.

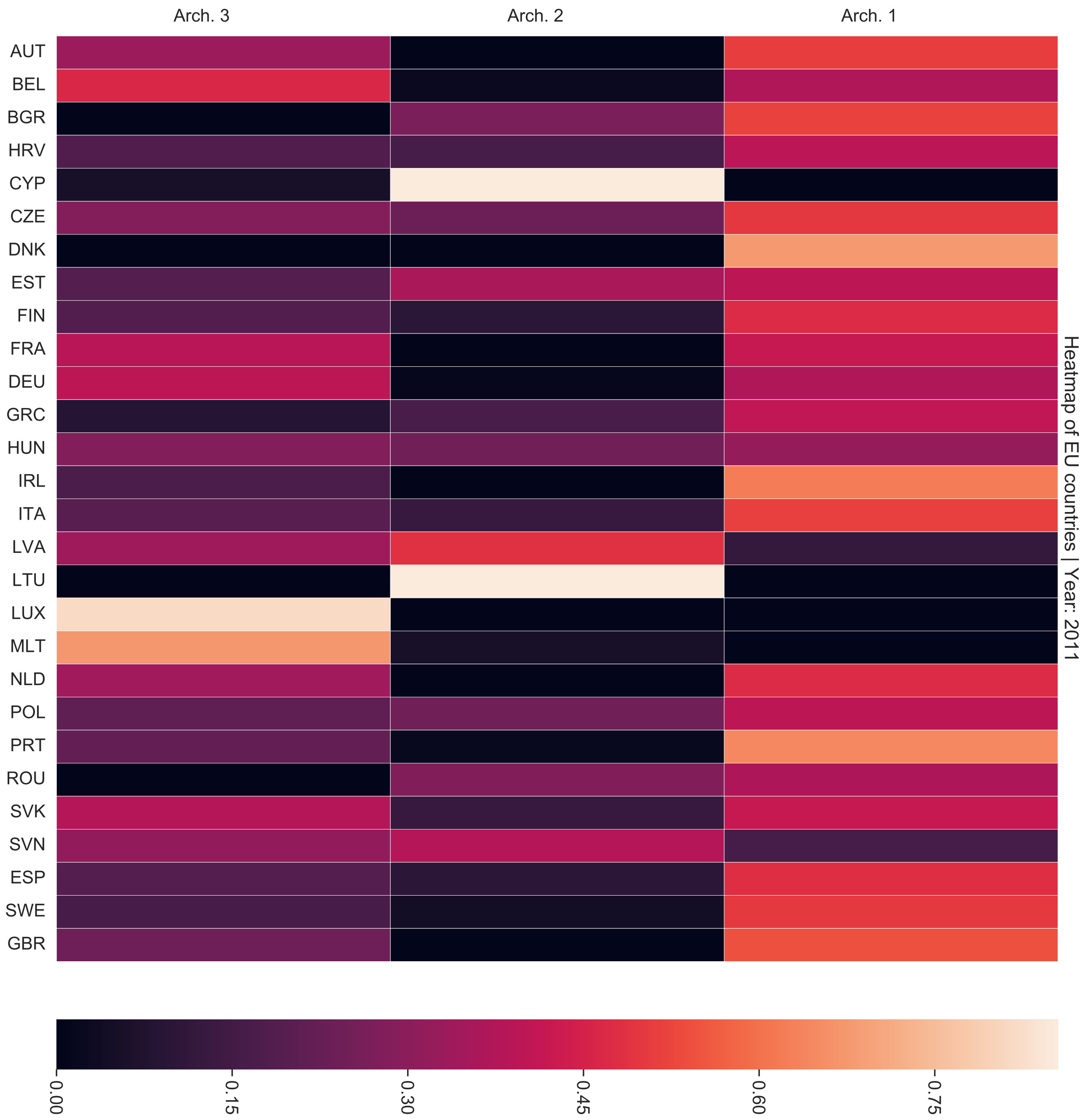

Figure 5.26: Heatmap of the activations of the countries in the cross-sectional EU dataset for the year 2011.

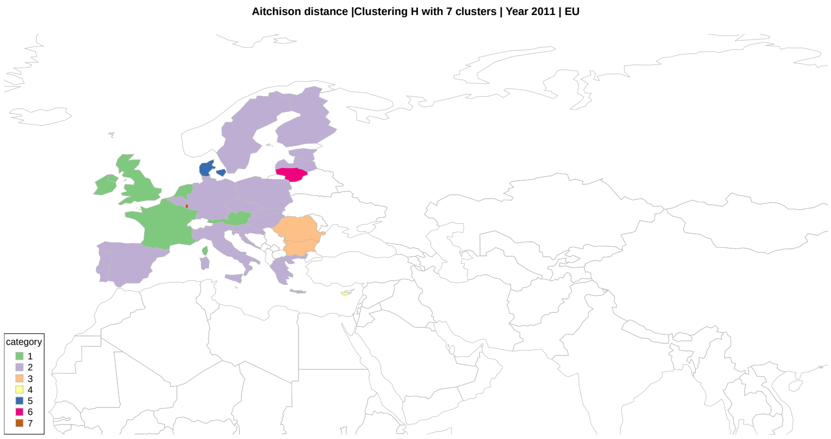
Aitchison distance |Clustering H with 7 clusters | Year 2011 | EU
category
1
2
3
4
5
6
7

Figure 5.27: Clustering obtained considering the cross-sectional EU dataset for the year 2011, via k-medoids and the Aitchison metric, with the number of clusters induced by the best Silhouette score.

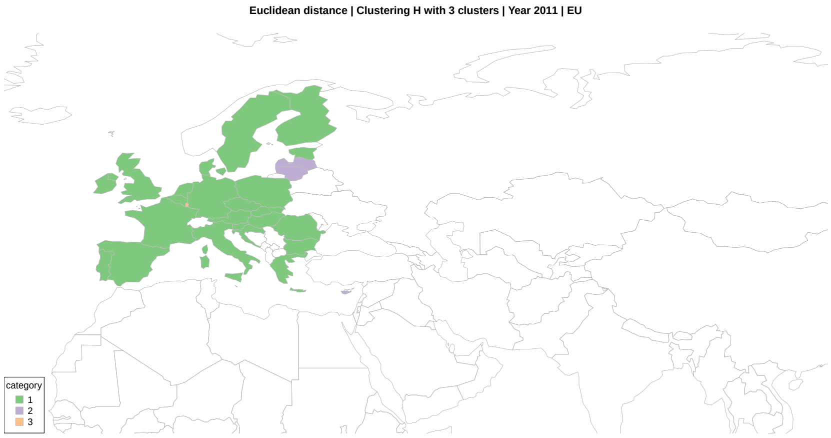
Euclidean distance | Clustering H with 3 clusters | Year 2011 | EU
category
1
2
3

Figure 5.28: Clustering obtained considering the cross-sectional EU dataset for the year 2011, via k-medoids and the Euclidean metric, with the number of clusters induced by the best Silhouette score.

5.3 ANALYSIS OF THE OECD CROSS-SECTIONAL DATASETS

The OECD dataset comprises Australia, Austria, Belgium, Canada, Chile, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel[5], Italy, Japan, Korea, Latvia, Lithuania, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom.

5.3.1 *Archetypes for the OECD cross-sectional datasets*

The rank estimation process yielded again $k = 3$. We report in Figures 5.29 and 5.30 the plots for the years 1995 and 2011 respectively. The Silhouette profiles were reported in Figures 5.31 and 5.32. The results for the Cross-validation and Bi-Cross-Validation are reported in Figures 5.33 and 5.34 for the year 1995 and in Figures 5.35 and 5.36 for the year 2011. We notice a strong similarity between this case and the EU case. Again, we would expect the Bi-Cross-Validation to increase more sharply as we get closer to a model of the same rank of the number of countries under scrutiny. This, together with the Silhouette profiles, again suggests that with a number of sectors this small it is hard to decompose the data.

Focusing on 1995, we report the plots of the archetypes in Figure 5.37, 5.38 and 5.39. The networks induced by the archetypes are reported in Figures 5.40, 5.41 and 5.42. The archetypes for the OECD in 1995 do not suggest any unambiguous interpretation. We find confirmation of what already found in the analysis of the EU dataset, noting that the "Wholesale" and the "R&D" sectors are the most central sectors when we consider the diffusion rapidity of supply-side domestic shock. The relative isolation of the "Education" and "Healthcare" is also confirmed. This was not necessary, as we difference between OECD are greater than those between EU countries. We notice that the first archetype shows stronger activations the sectors of "Wholesale", "R&D", "Financial intermediation", "Electricity, gas and water supply", "Chemical" and "Coke and nuclear fuel". The most relevant sectors in the second case are "Transport", "R&D", "Financial intermediation", "Wholesale". Uniform activations are also presents in the sectors related with the basic industrial infrastructure. The third archetype differs from the first two in its activation of the "Mining" sector.

We report the archetypes for 2011 in Figures 5.43, 5.44 and 5.45. The networks induced by the archetypes are reported in Figures 5.46, 5.47 and 5.48. In this case, the identification of the archetypes seems to

[5] "The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities or third party. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law." (https://www.oecd.org/sti/ind/input-outputtables.htm)

succeed in identifying different patterns. The first archetype is characterised by the pronounced activation of the “R&D" and “Wholesale" sector. The second archetype shows strong activations in the “R&D", “Wholesale", “Real estate" and “Financial intermediation". The third archetype is skewed towards “Basic metals", “Mining" and “Coke and nuclear fuels".

As for why the results are so different, it may be due to significant changes in economic configuration of the OECD countries.

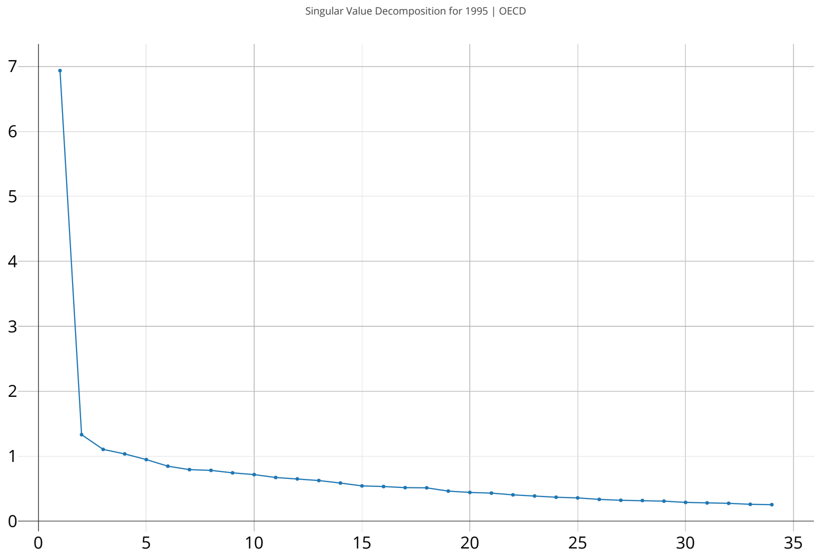

Figure 5.29: Singular Values of the SVD computed for the OECD cross-sectional dataset in 1995.

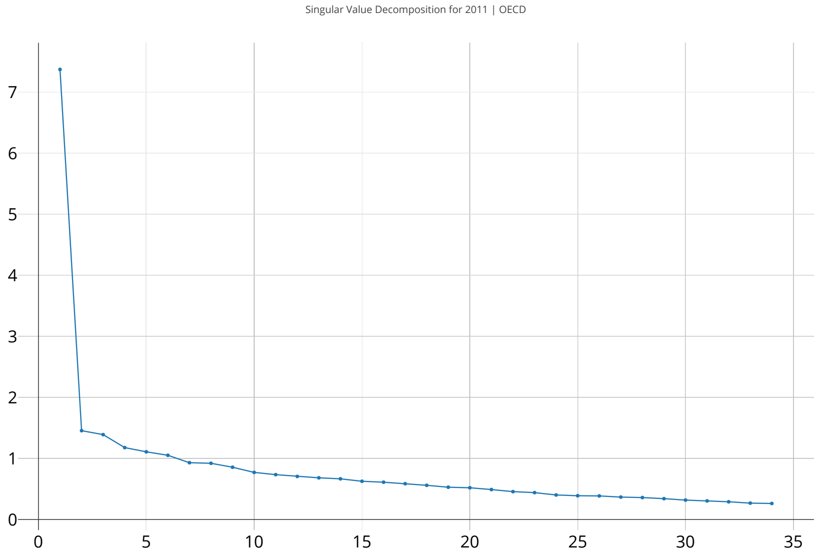

Figure 5.30: Singular Values of the SVD computed for the OECD cross-sectional dataset in 2011.

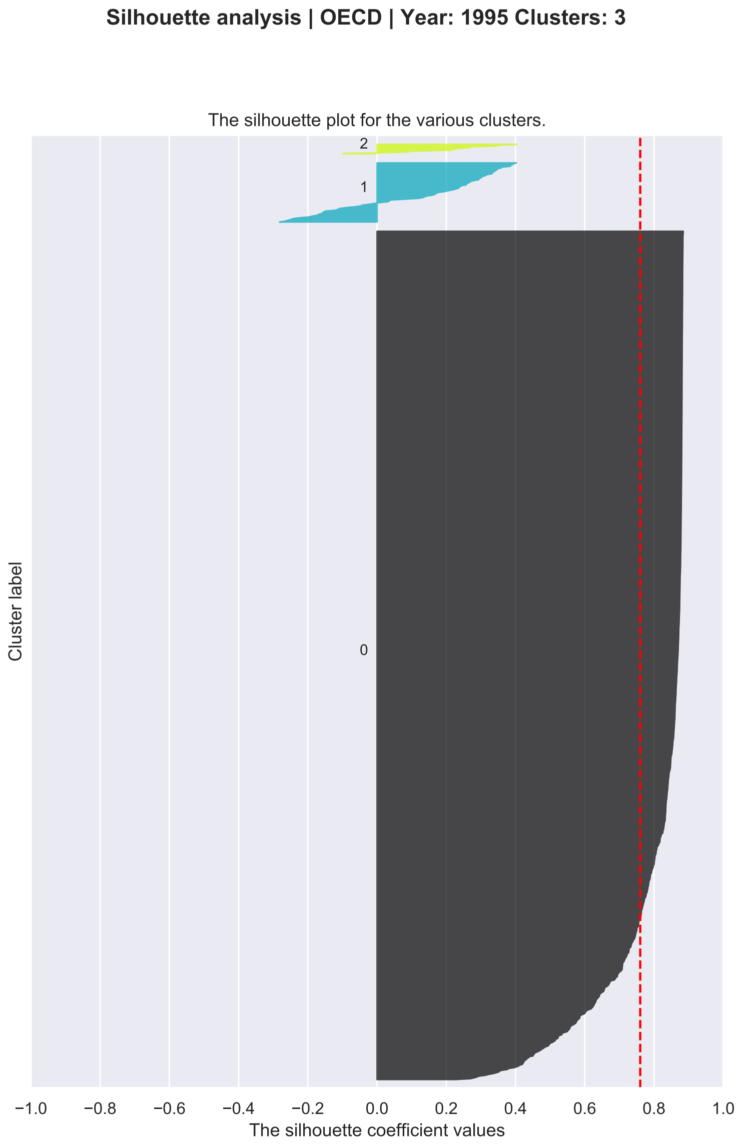

Figure 5.31: Silhouette profiles for the k-means algorithm applying with 3 clusters to the cross-sectional OECD dataset fo the year 1995.

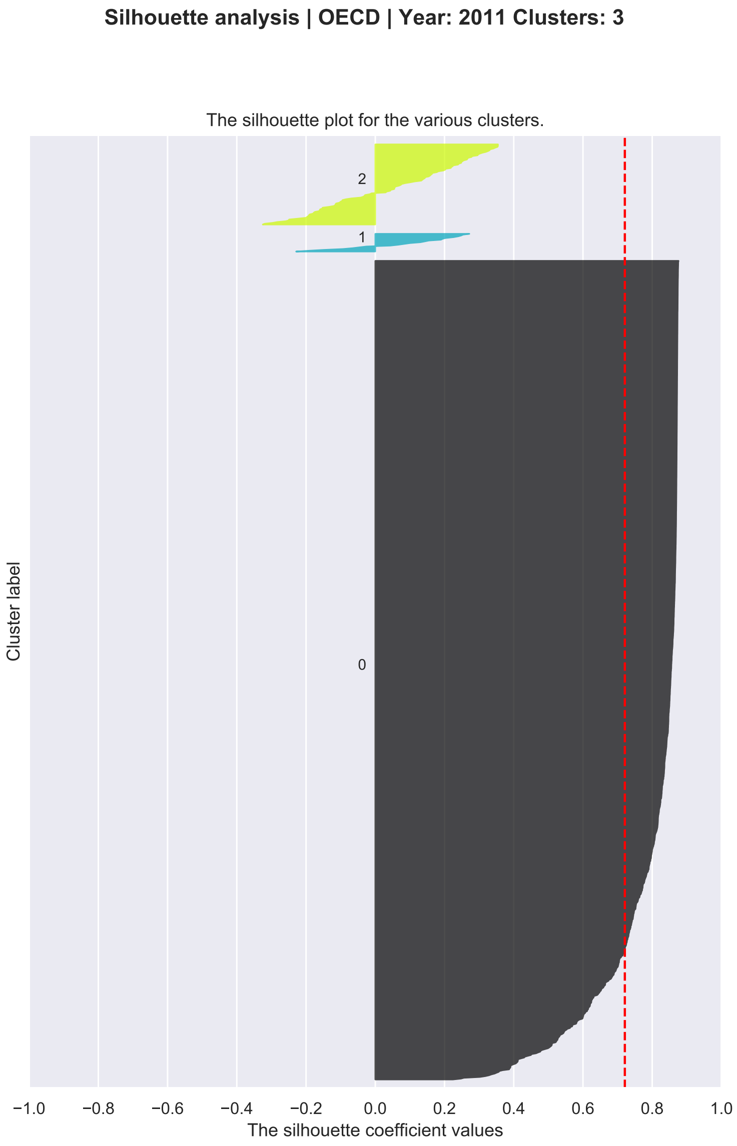

Figure 5.32: Silhouette profiles for the k-means algorithm applying with 3 clusters to the cross-sectional OECD dataset fo the year 2011.

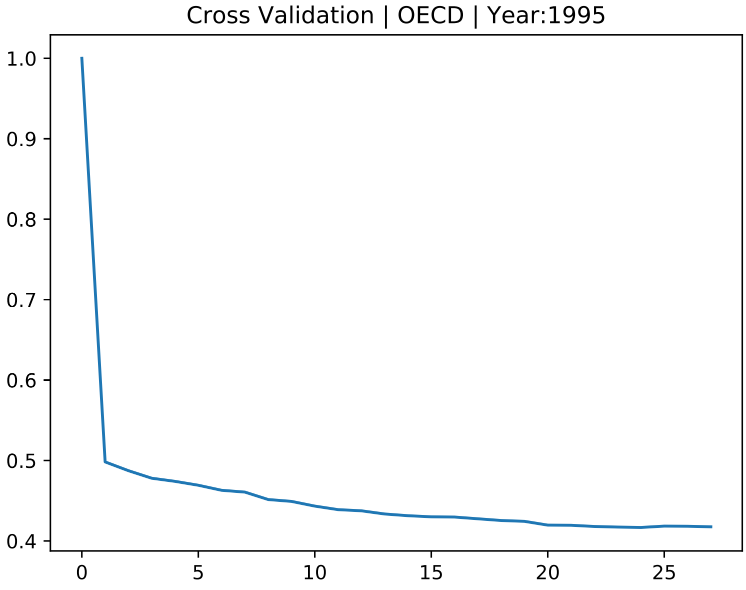

Figure 5.33: Cross-Validation for the cross-sectional OECD dataset for the year 1995.

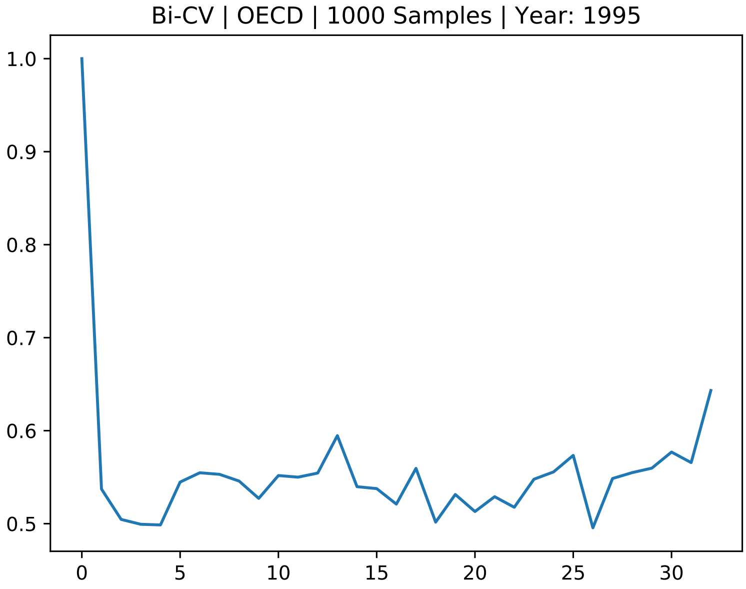

Figure 5.34: Bi-Cross-Validation with simple residuals for the cross-sectional OECD dataset for the year 1995. We considered (3×3) random submatrices and, for each rank, averaged the results of 1000 random trials. Data are normalised considering the MSE of a rank 0 approximation, *i.e.* the Frobenius norm of the sampled matrices alone.

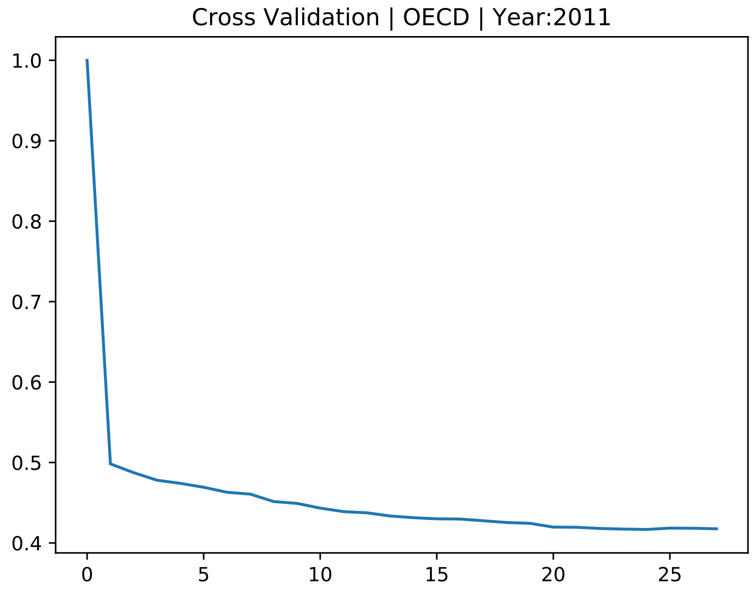

Figure 5.35: Cross-Validation for the cross-sectional OECD dataset for the year 2011.

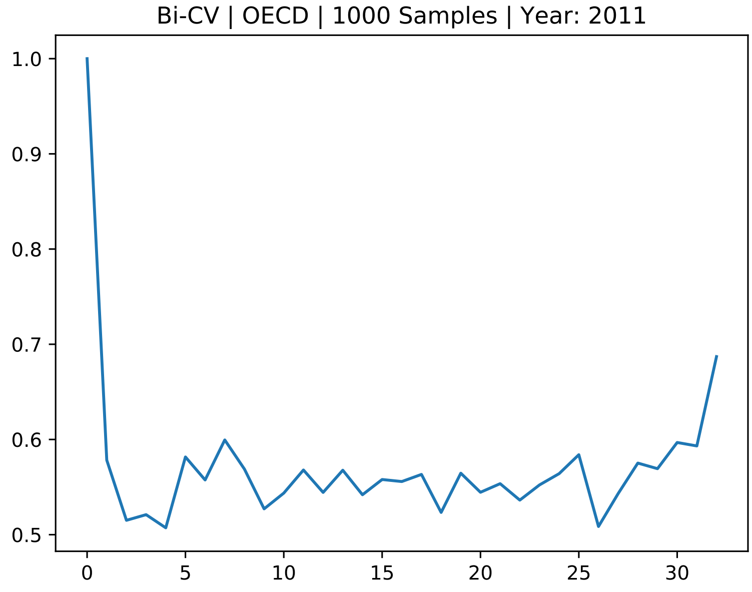

Figure 5.36: Bi-Cross-Validation with simple residuals for the cross-sectional OECD dataset for the year 2011. We considered (3×3) random submatrices and, for each rank, averaged the results of 1000 random trials. Data are normalised considering the MSE of a rank 0 approximation, *i.e.* the Frobenius norm of the sampled matrices alone.

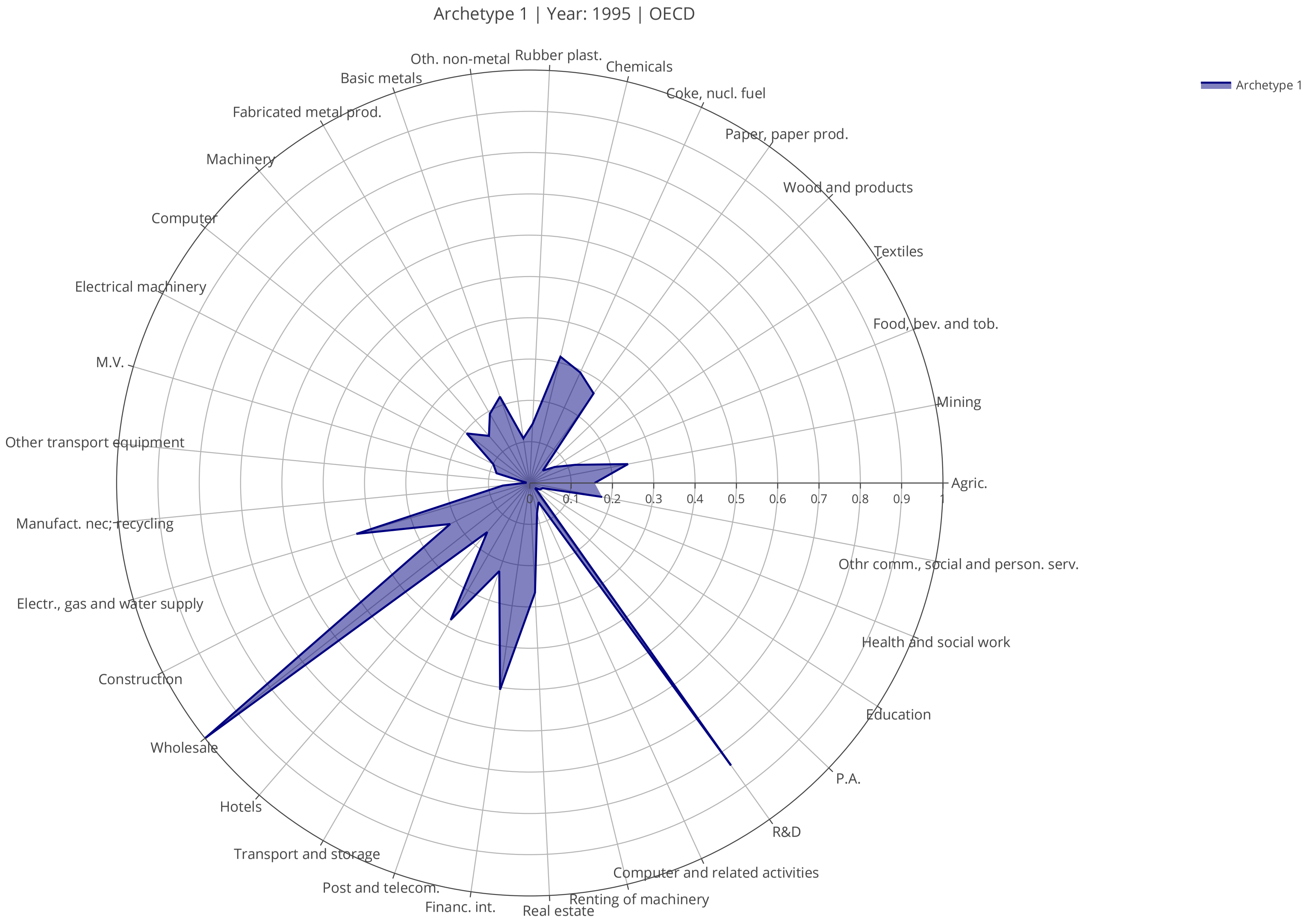

Figure 5.37: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional OECD dataset for the year 1995.

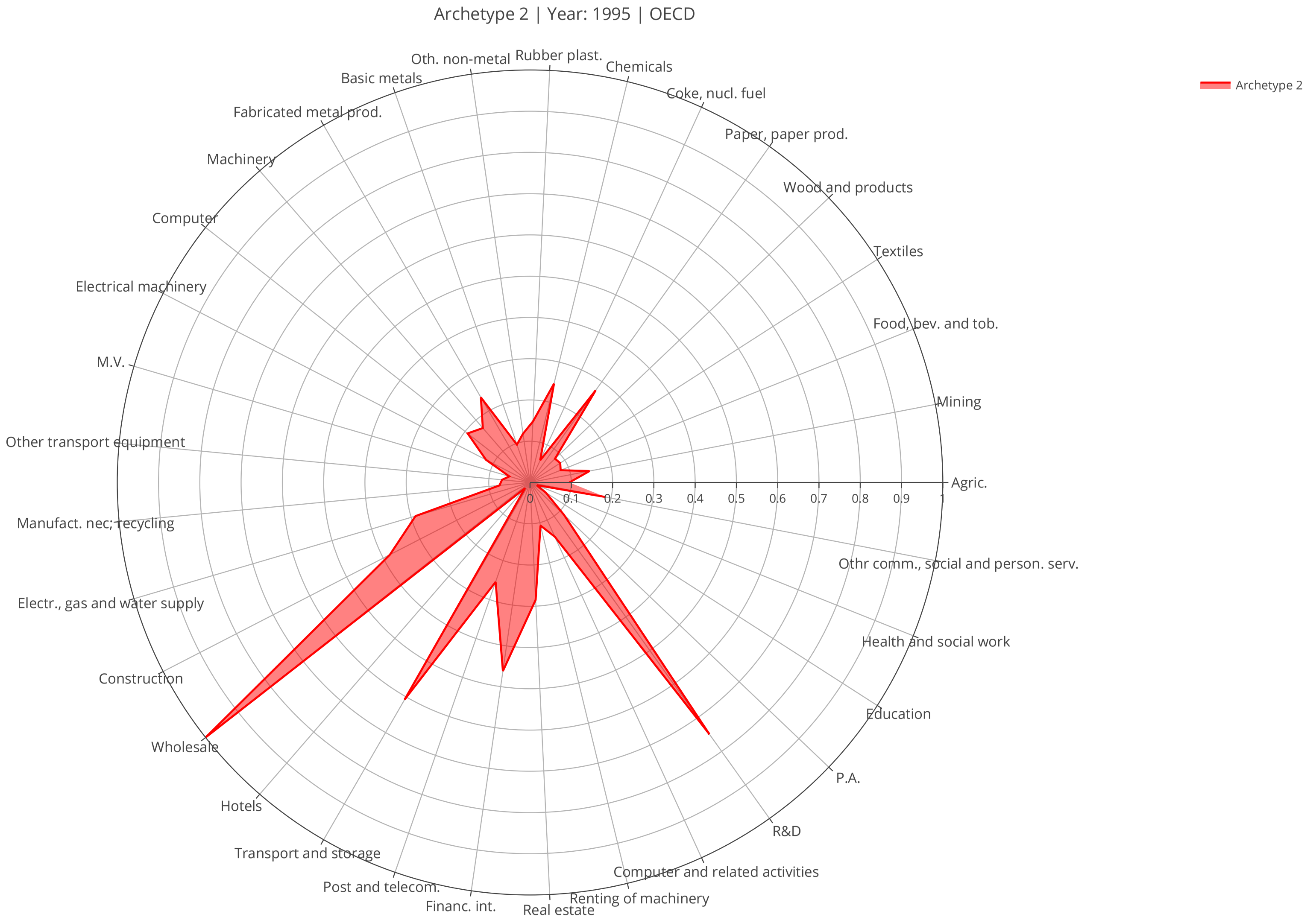

Figure 5.38: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional OECD dataset for the year 1995.

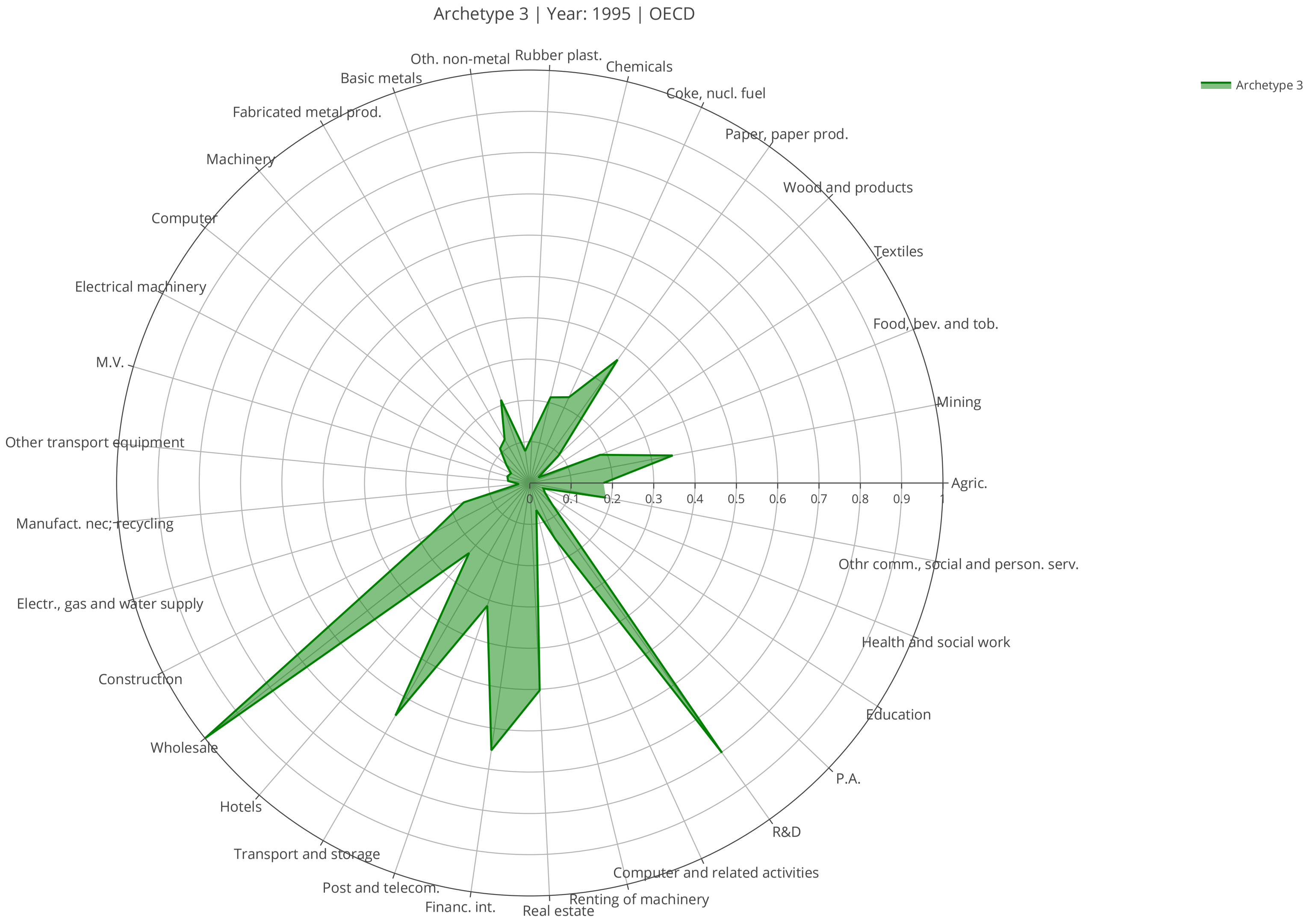

Figure 5.39: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional OECD dataset for the year 1995.

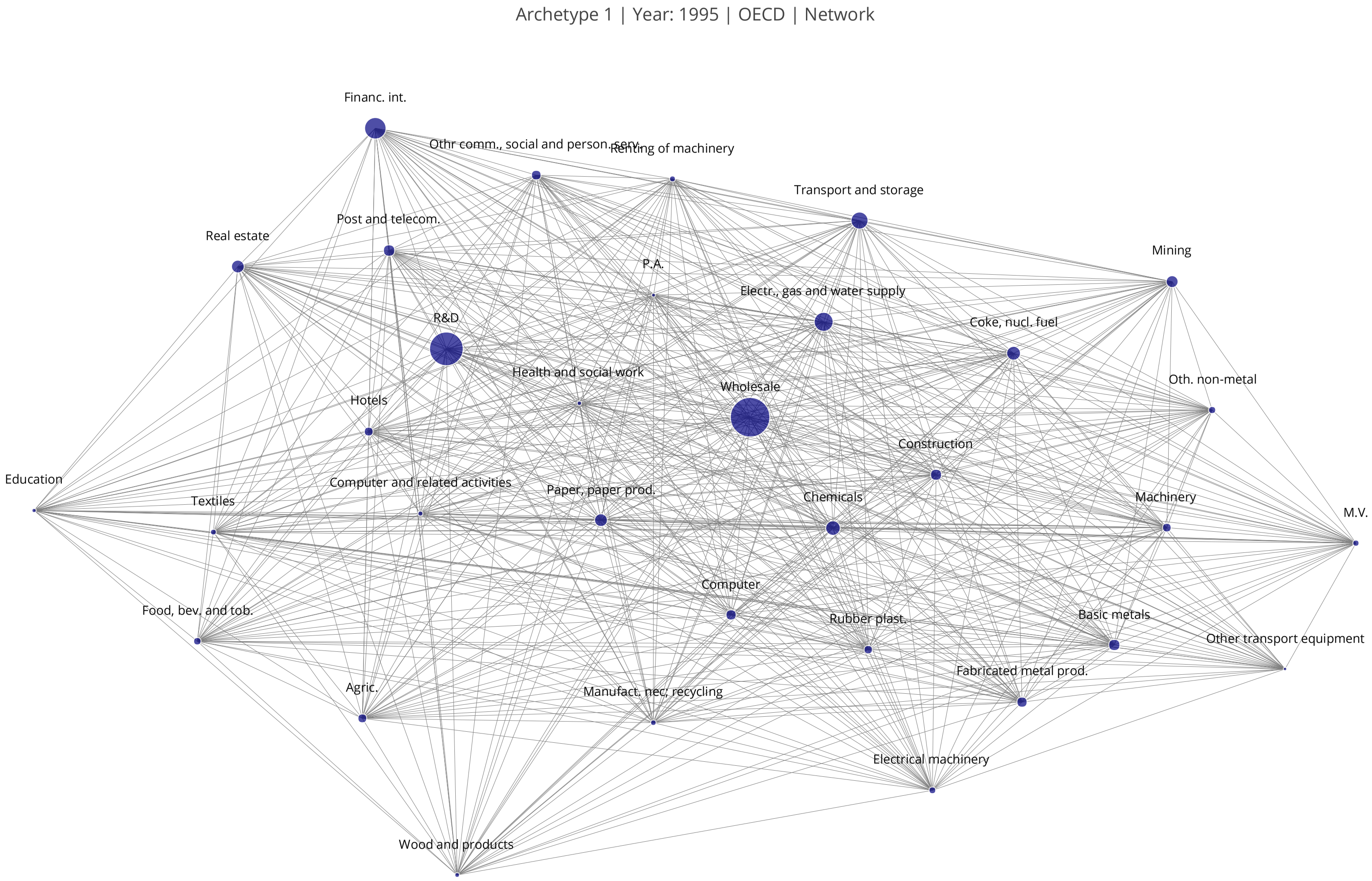

Figure 5.40: Network induced by the first archetype of the cross-sectional OECD dataset for the year 1995. The dimension of the nodes is proportional to their shock-centrality.

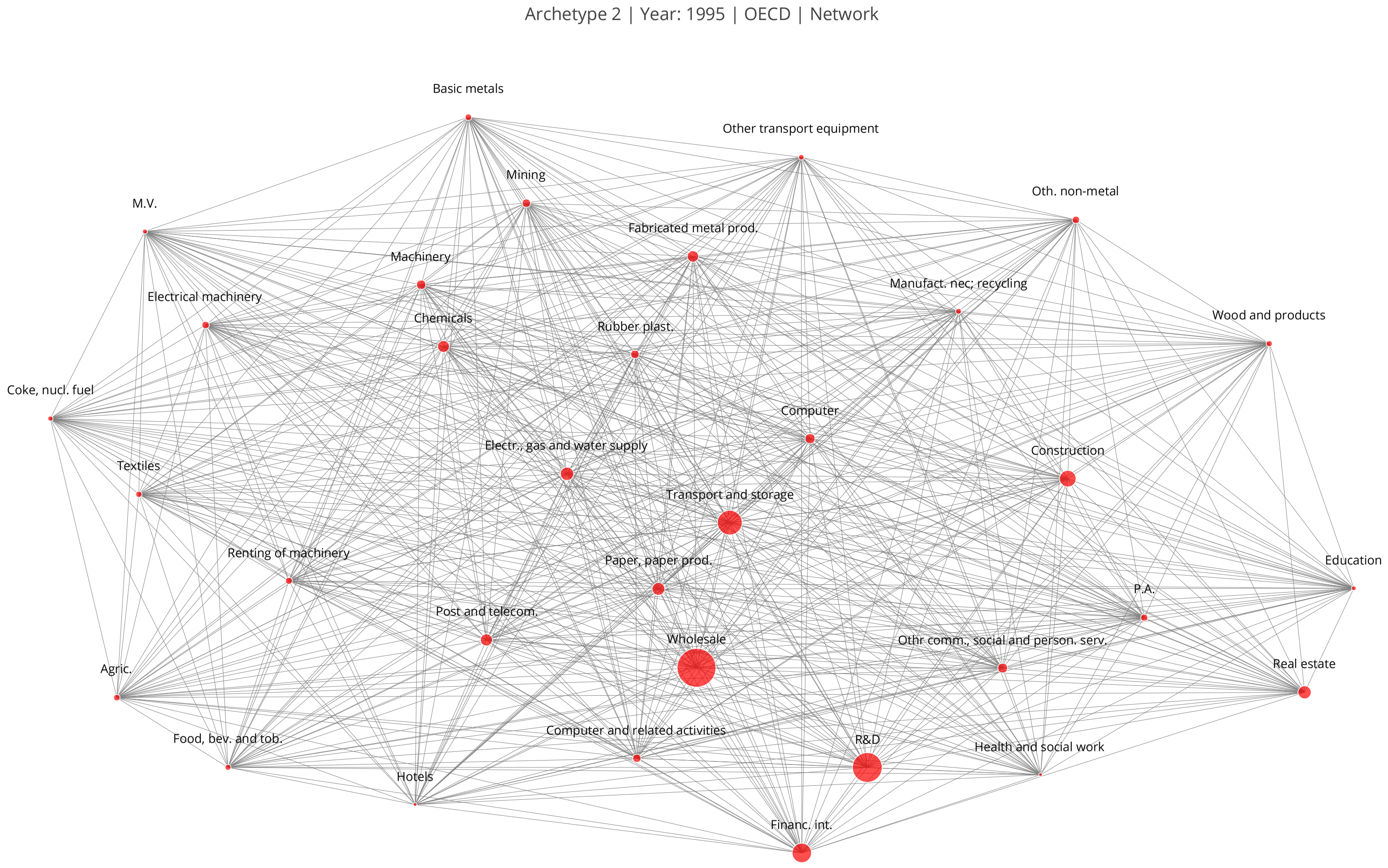

Figure 5.41: Network induced by the second archetype of the cross-sectional OECD dataset for the year 1995. The dimension of the nodes is proportional to their shock-centrality.

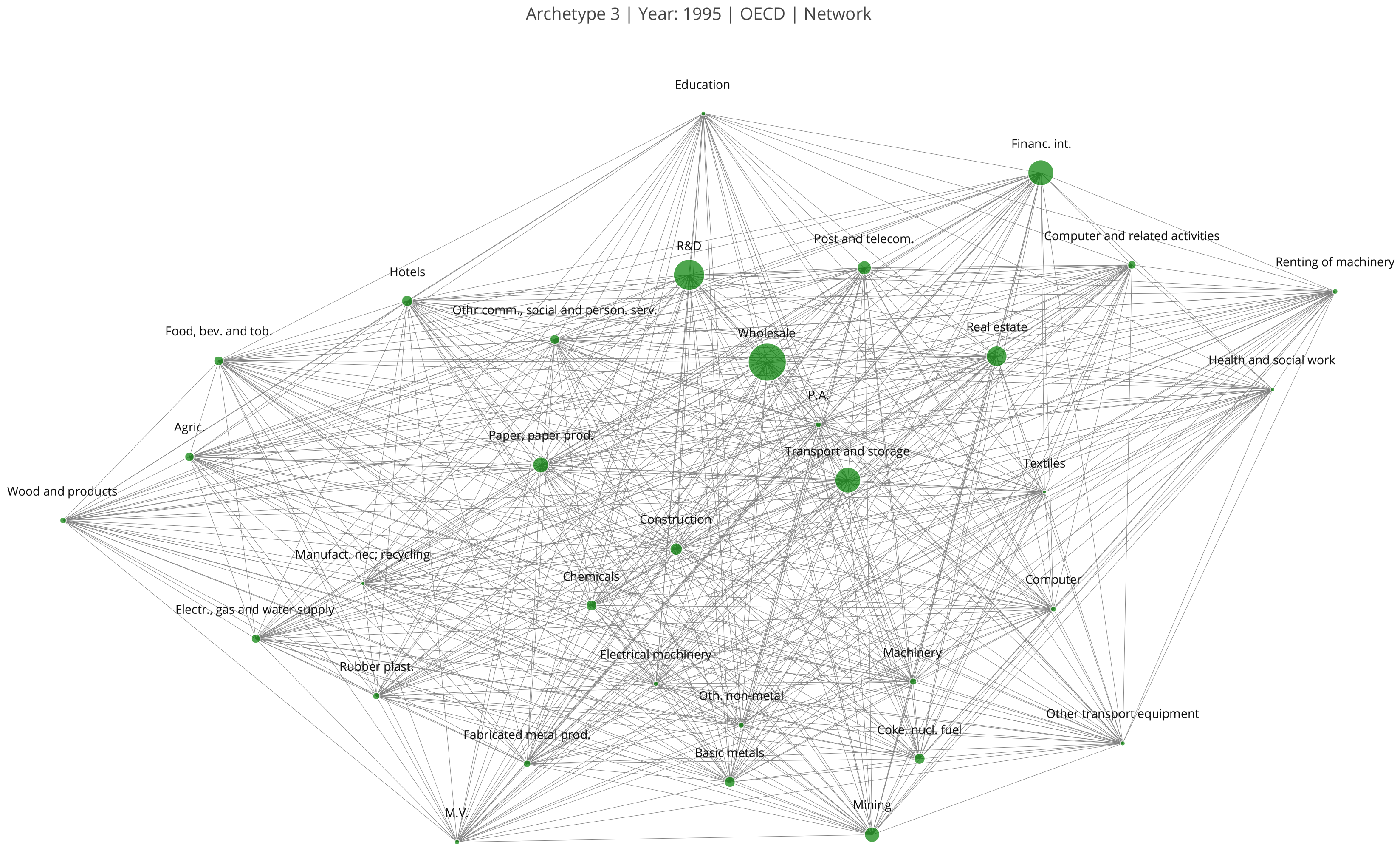

Figure 5.42: Network induced by the third archetype of the cross-sectional OECD dataset for the year 1995. The dimension of the nodes is proportional to their shock-centrality.

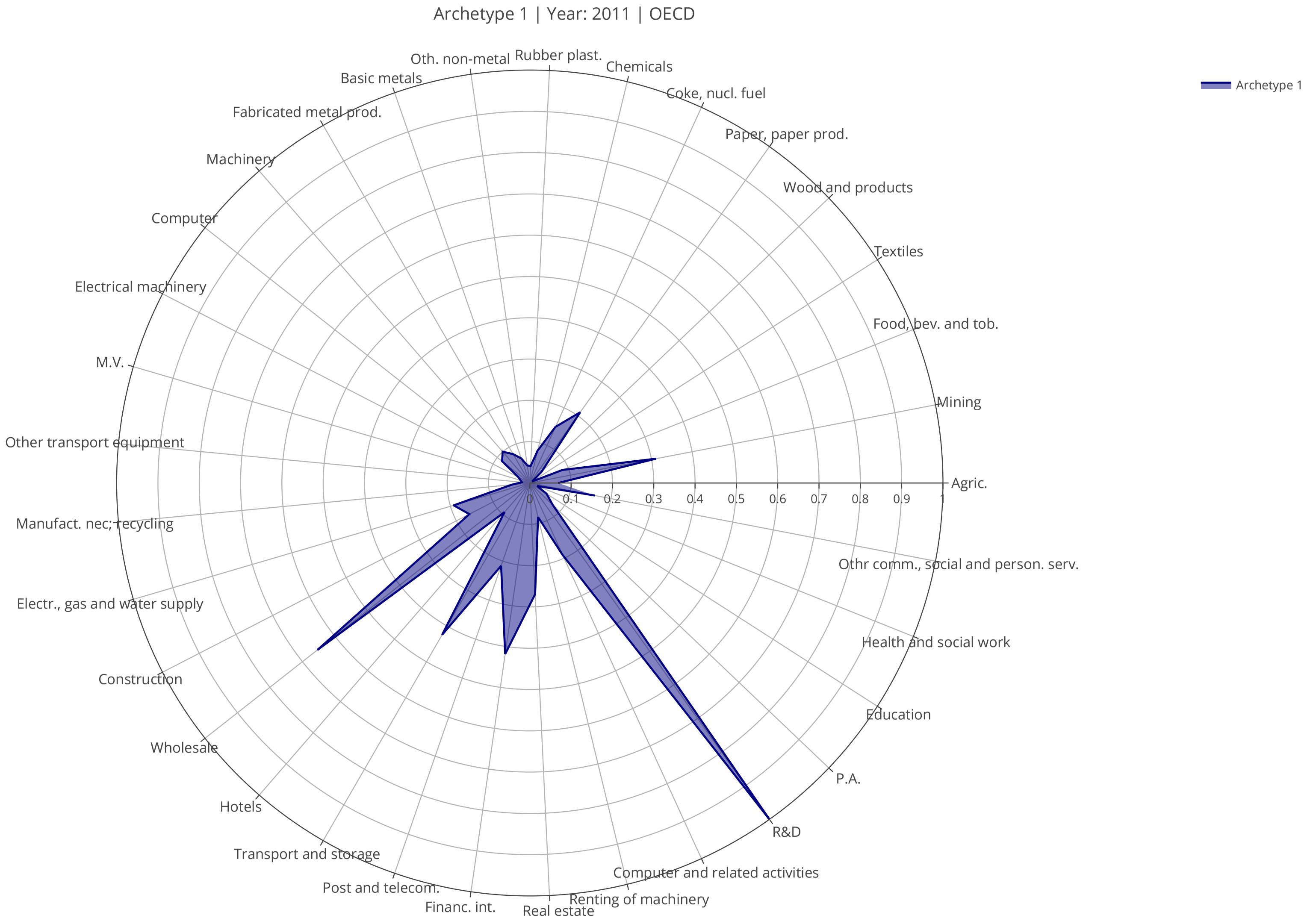

Figure 5.43: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional OECD dataset for the year 2011.

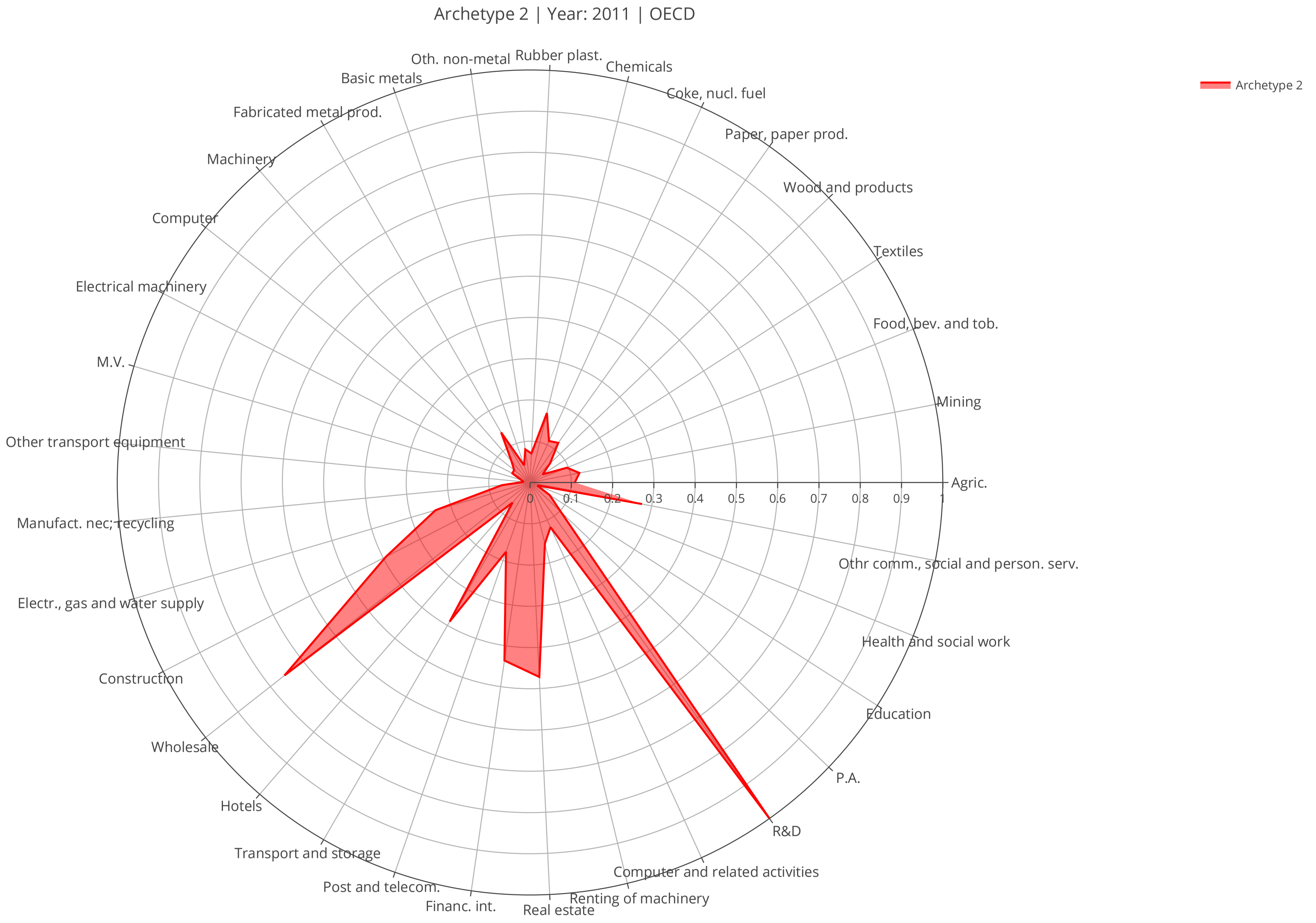

Figure 5.44: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional OECD dataset for the year 2011.

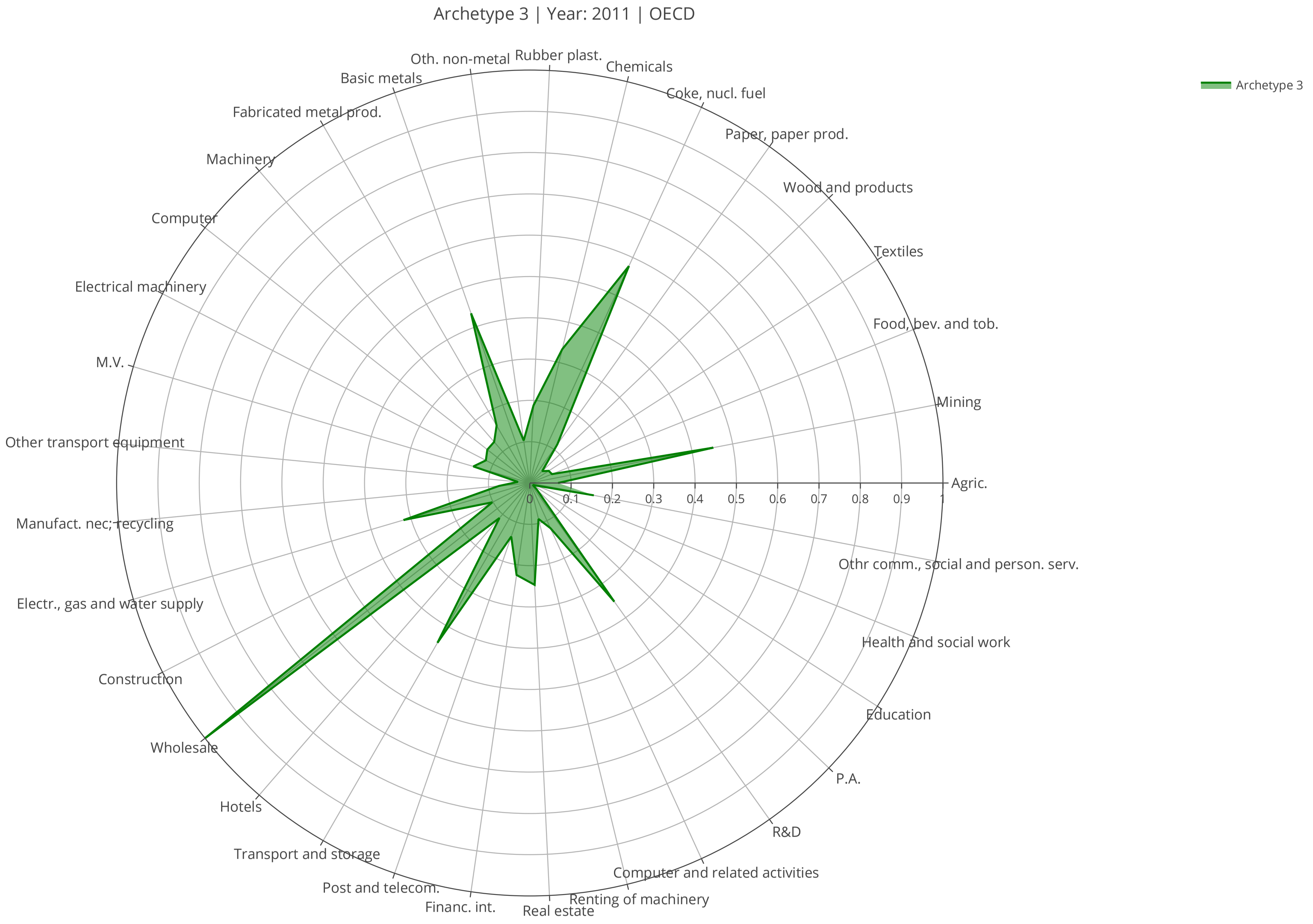

Figure 5.45: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional OECD dataset for the year 2011.

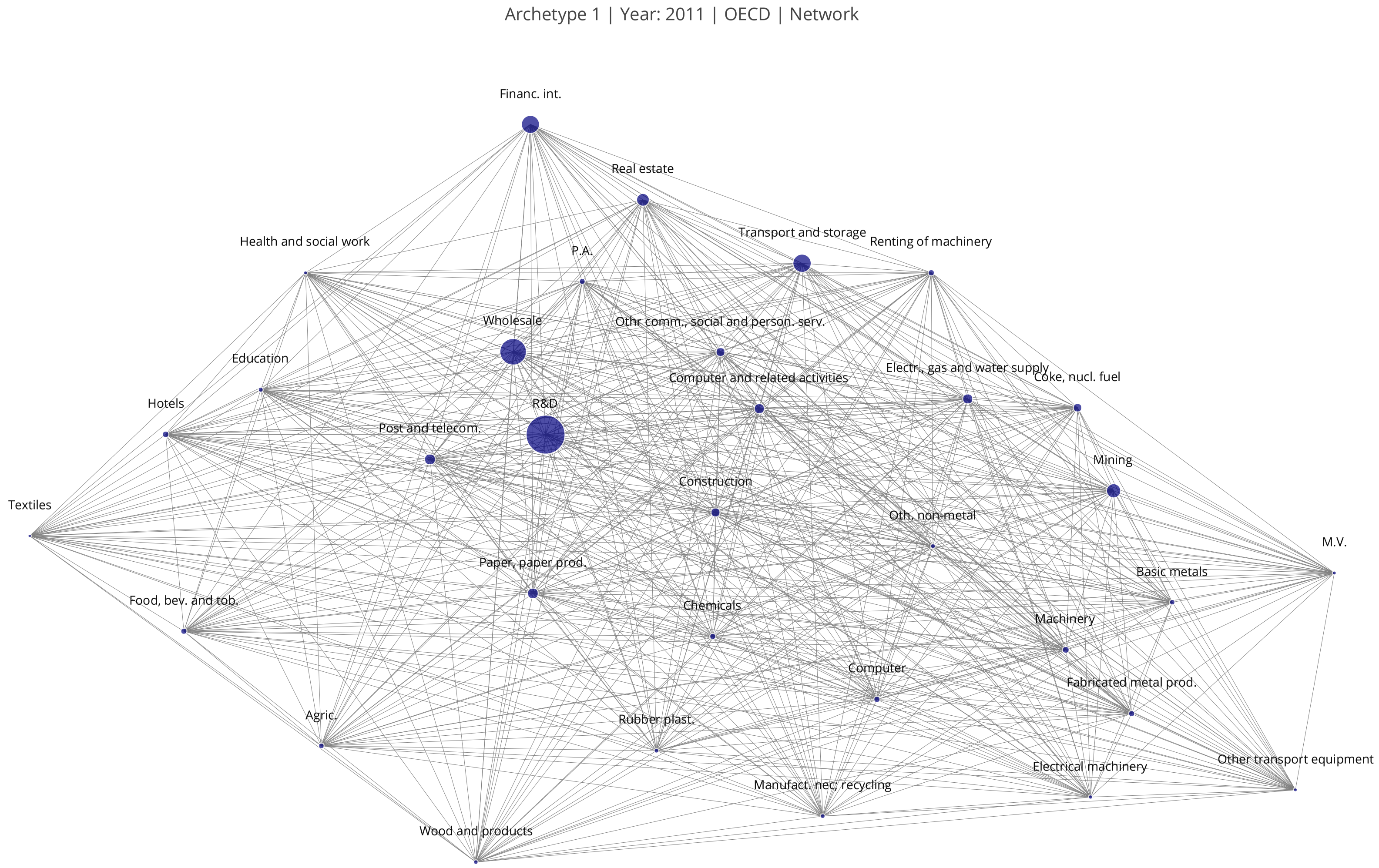

Figure 5.46: Network induced by the first archetype of the cross-sectional OECD dataset for the year 2011. The dimension of the nodes is proportional to their shock-centrality.

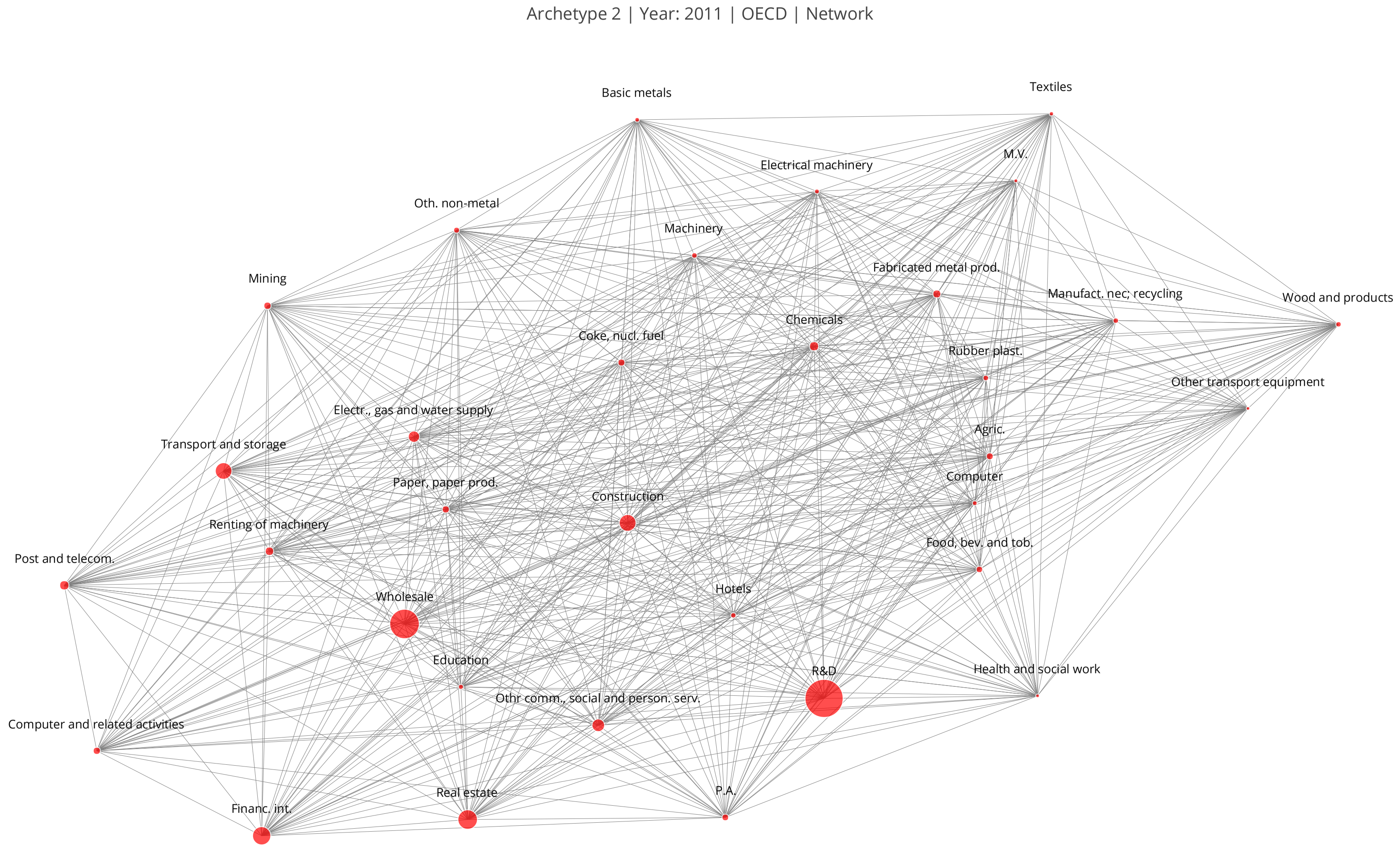

Figure 5.47: Network induced by the second archetype of the cross-sectional OECD dataset for the year 2011. The dimension of the nodes is proportional to their shock-centrality.

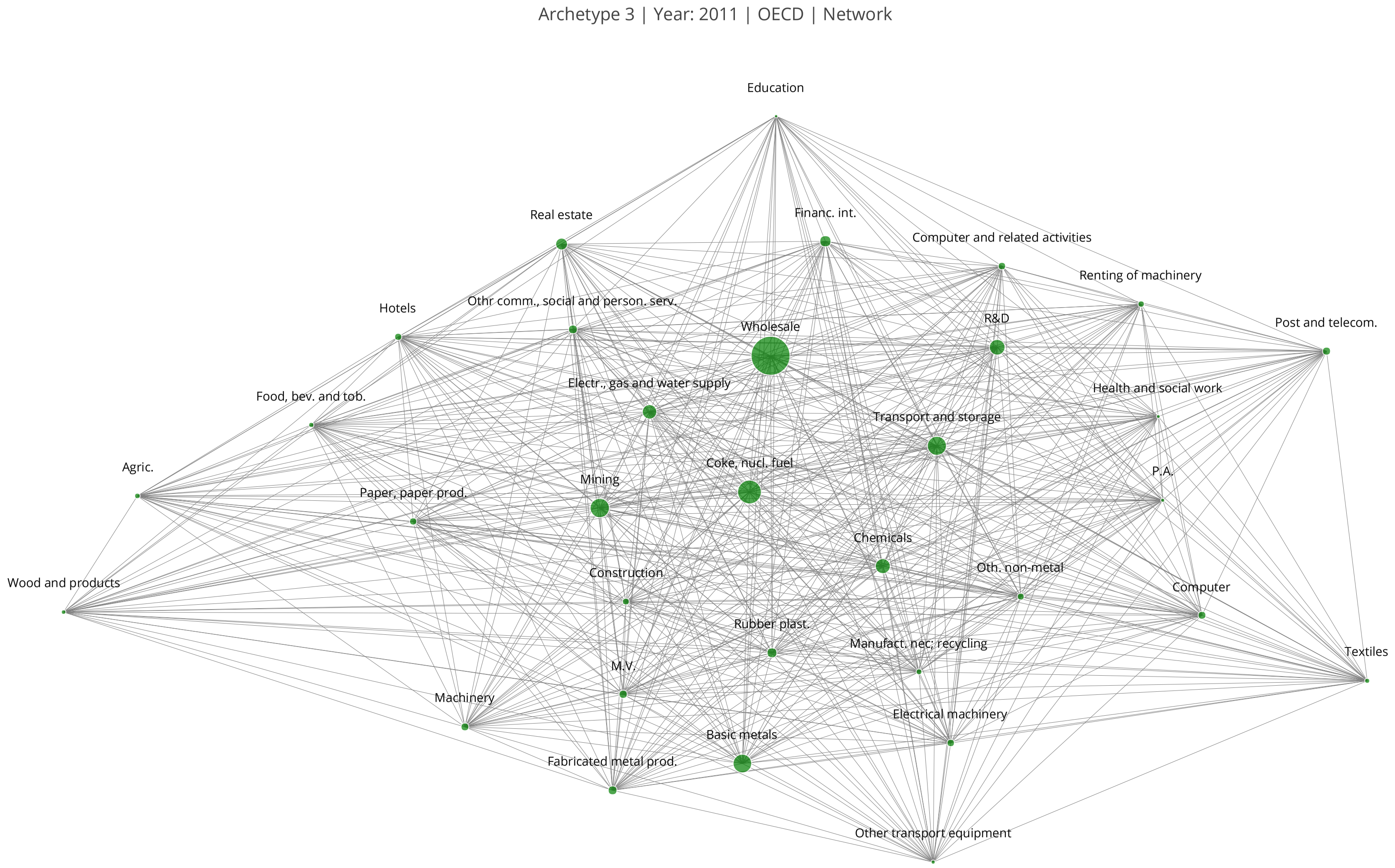

Figure 5.48: Network induced by the third archetype of the cross-sectional OECD dataset for the year 2011. The dimension of the nodes is proportional to their shock-centrality.

Medoid	Archetype 1	Archetype 2	Archetype 3
Spain	0.266	0.351	0.384
Mexico	0.539	0.000	0.461
Luxembourg	0.629	0.371	0.000
Switzerland	0.000	1.000	0.000

Table 5.5: Closure of the activations for the centroids obtained via a k-medoids clustering in the Aitchison geometry for cross-sectional OECD dataset of 1995.

5.3.2 *Clustering for the cross-sectional OECD datasets*

We report in Figure 5.49 and 5.52 the heatmaps with the activations for the countries for 1995 and 2011.

We now focus on the clustering for the year 1995. We report the centroid obtained within the compositional framework in Table 5.5. We notice that once again the Aitchison distance manages to capture differences between countries. Spain is again a medoid, showing coefficients not too different from those in the previous section. This can be considered as an indication of a sort of equilibrium in the economic structure of the country. Together with Spain, in the cluster we find Austria, Italy, Slovakia, the United Kingdom, Germany, Portugal, the Czech Republic, Poland, France, Sweden, Israel, the United States of America, Denmark, Turkey, New Zealand, Norway, Greece, Australia, Belgium, Slovenia, Finland, Island, the Netherlands, Japan and Chile. This group might seem, at first, somewhat heterogeneous. However, considering the geographical contiguity of the countries analysed and the fact that all of them are either developing or developed countries might be considered as sign of the fact that all of them are an equilibrate superposition of the different shock-patterns. In the second cluster, whose medoid is Mexico, we also find Canada and South Korea. Such countries should differ from the first cluster because of either their vulnerability to shocks affecting the extractive sectors or their relationship with the energetic supplies to their industries. The third cluster comprises Estonia, Luxembourg, Hungary and Ireland. This suggests that all four countries are vulnerable to shocks affecting, in particular, the financial sector and the transport sector, in addition to the usual sectors always showing high shock-centrality. To conclude, the fourth cluster contains Switzerland alone. In this case, a high sensitivity to shocks affecting the financial sector seems to be well-justified, given the particular history of this country.

The medoids induced by the Euclidean clustering are reported in Table 5.6. We here have three clusters. The first one, whose medoid is the United Kingdom, also contains Denmark, the Netherlands, Australia, Austria, Italy, New Zealand, Norway, Chile, Portugal, Spain,

Medoid	Archetype 1	Archetype 2	Archetype 3
United Kingdom	0.116	0.339	0.265
Greece	0.413	0.086	0.216
Ireland	0.237	0.602	0.000

Table 5.6: Activations for the centroids obtained via a k-medoids clustering in the Euclidean geometry for cross-sectional OECD dataset of 1995.

Germany, Poland, Slovakia and France. Despite the Silhouette coefficient being worse, we once again find some geographical and economic affinity between countries in this cluster. The second cluster, whose medoid is Greece, contains Japan, Slovenia, Finland, South Korea, Turkey, Israel, Mexico, Hungary, Luxembourg, Canada, the Czech Republic, the United States of America, Belgium and Sweden. The third one, represented by Ireland, contains also Switzerland, Island and Estonia. Again, what we notice is that the Aitchison distance is probably better-suited to cluster the data: it is not the activations *per se* that are of interest, but rather their relationship with each other.

We now turn our attention to the clustering for 2011. We report in Tables 5.7 and 5.8 the medoids for the clusterings. The first cluster comprises Australia, Austria, Belgium, Chile, Czech Republic, Estonia, Finland, France, Germany, Greece, Hungary, Island, Ireland, Italy. Mexico, the Netherlands, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey and the United Kingdom. Such countries show an equal distribution of the activations. We can try to interpret this cluster by noting that this means that such countries show a shock-centrality equally balanced between the extractive, the research and the usual “Wholesale" and “Transport" sectors. The second cluster contains Canada, Japan, South Korea, Norway and the United States of America. In this case, shock centrality is strongest when we consider the extractive and energetic sector, together with the “R&D", “Financial intermediation" and the “Wholesale" sectors. The cluster of New Zealand and Denmark shows higher centrality for the “Real estate" sector, together with the usually central ones. To conclude, Israel and Luxembourg show activations for the archetype mostly connected with “Mining" and “Coke and nuclear fuel".

To aid the visualisation of the clusters obtained, we present the results for 1995 in Figures 5.50 and 5.51 for the Aitchison and Euclidean distances and in Figures 5.53 and 5.54 for 2011.

Medoid	Archetype 1	Archetype 2	Archetype 3
Czech Republic	0.346	0.335	0.319
Canada	0.501	0.000	0.499
New Zealand	0.259	0.741	0.000
Luxembourg	0.000	0.620	0.380

Table 5.7: Closure of the activations for the centroids obtained via a k-medoids clustering in the Aitchison geometry for cross-sectional OECD dataset of 2011.

Medoid	Archetype 1	Archetype 2	Archetype 3
Spain	0.357	0.207	0.210
Slovenia	0.013	0.548	0.227

Table 5.8: Centroids obtained via a k-medoids clustering in the Euclidean geometry for cross-sectional OECD dataset of 2011.

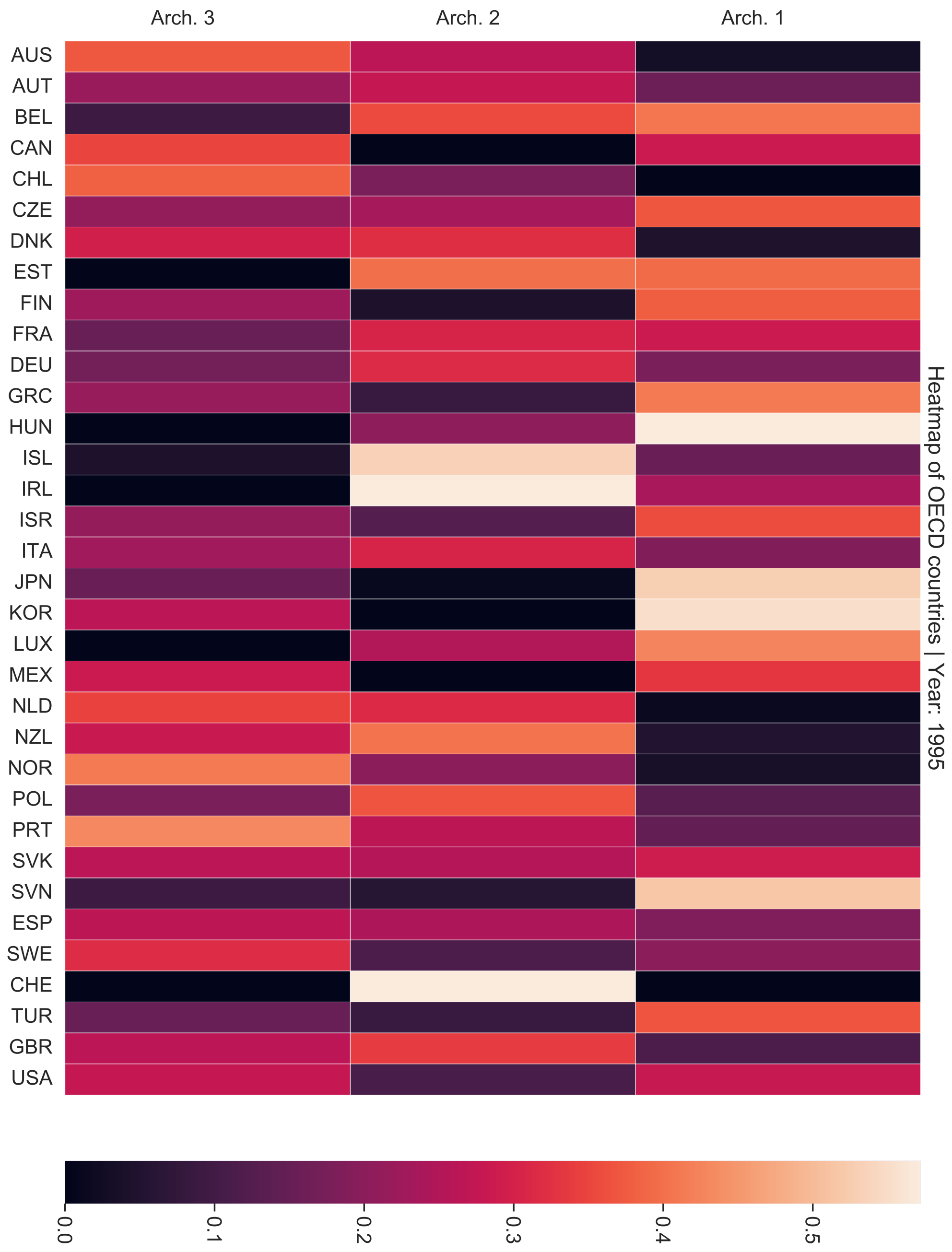

Figure 5.49: Heatmap of the activations of the countries in the cross-sectional OECD dataset for the year 1995.

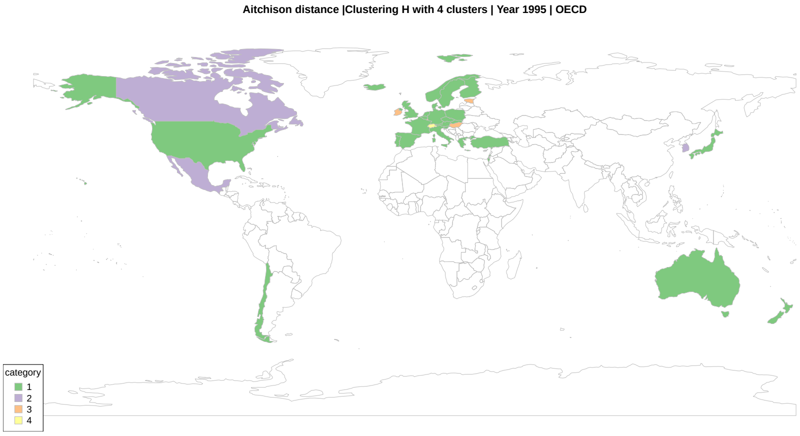
Aitchison distance |Clustering H with 4 clusters | Year 1995 | OECD
category
1
2
3
4

Figure 5.50: Clustering obtained considering the cross-sectional OECD dataset for the year 1995, via k-medoids and the Aitchison metric, with the number of clusters induced by the best Silhouette score.

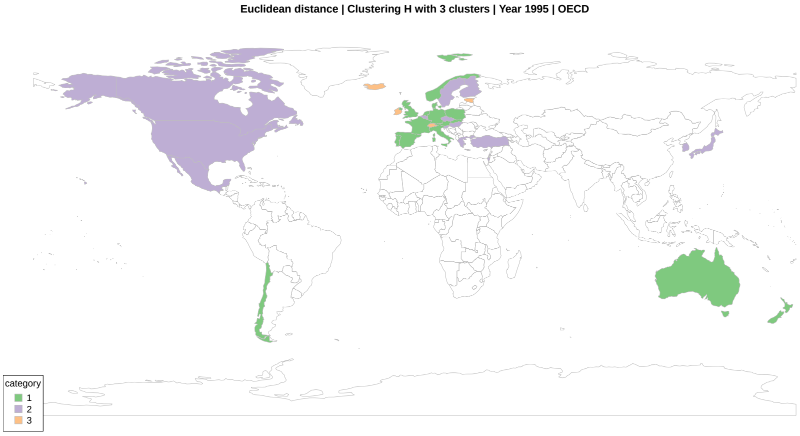
Euclidean distance | Clustering H with 3 clusters | Year 1995 | OECD
category
1
2
3

Figure 5.51: Clustering obtained considering the cross-sectional OECD dataset for the year 1995, via k-medoids and the Euclidean metric, with the number of clusters induced by the best Silhouette score.

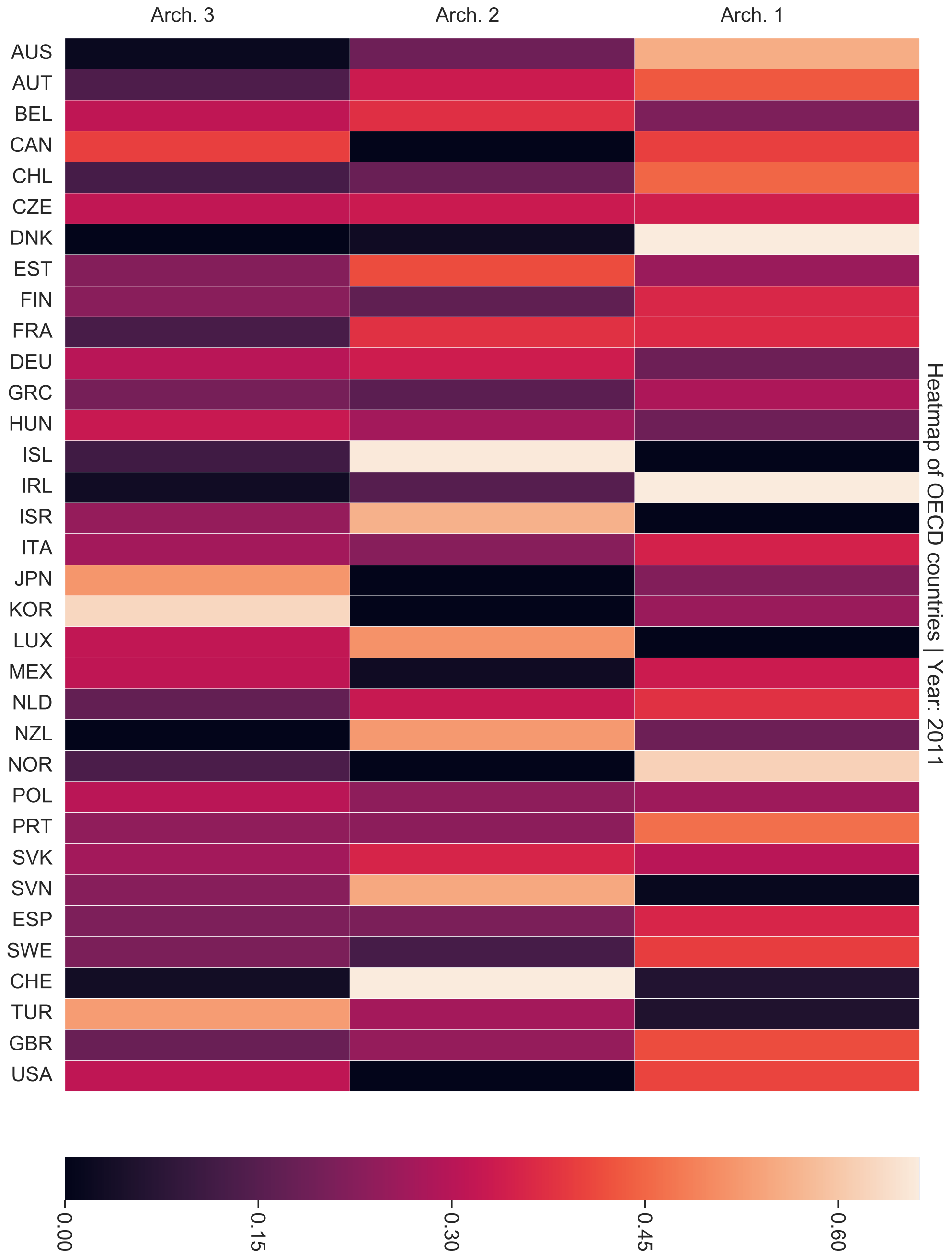

Figure 5.52: Heatmap of the activations of the countries in the cross-sectional OECD dataset for the year 2011.

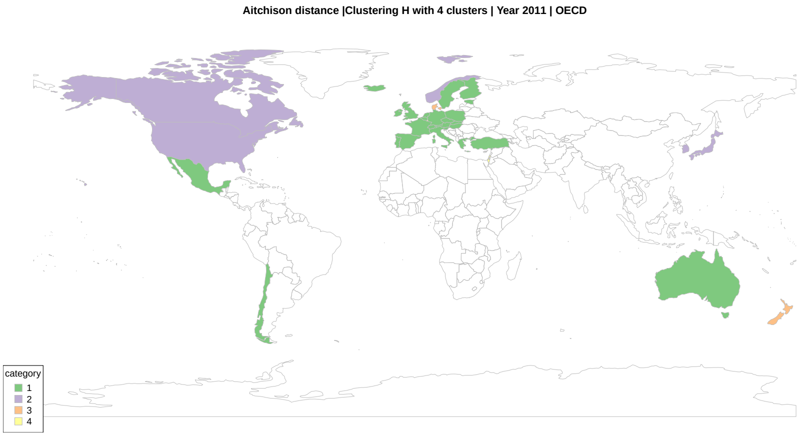
Aitchison distance |Clustering H with 4 clusters | Year 2011 | OECD
category
🟩 1
🟪 2
🟧 3
🟨 4

Figure 5.53: Clustering obtained considering the cross-sectional OECD dataset for the year 2011, via k-medoids and the Aitchison metric, with the number of clusters induced by the best Silhouette score.

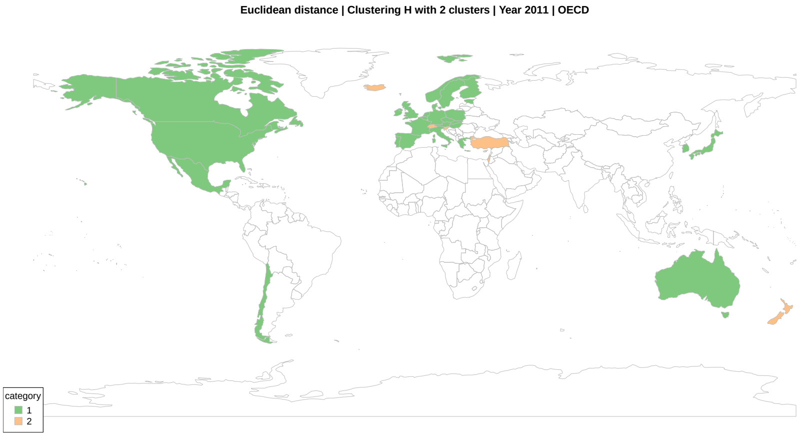
Euclidean distance | Clustering H with 2 clusters | Year 2011 | OECD
category
🟩 1
🟧 2

Figure 5.54: Clustering obtained considering the cross-sectional OECD dataset for the year 2011, via k-medoids and the Euclidean metric, with the number of clusters induced by the best Silhouette score.

5.4 ANALYSIS OF THE WORLD CROSS-SECTIONAL DATASETS

The World dataset comprises the countries of the OECD dataset, together with Argentina, Brazil, Brunei Darussalam, Bulgaria, Cambodia, China (People's Republic of), Colombia, Costa Rica, Croatia, Cyprus, India, Indonesia, Hong Kong, Kazakhstan, Malaysia, Malta, Morocco, Peru, Philippines, Romania, Russia, Saudi Arabia, Singapore, South Africa, Chinese Taipei, Thailand, Tunisia, Vietnam and a fictitious *Rest of the World* aggregate.

5.4.1 *Archetypes for the World cross-sectional datasets*

Rank estimation for the World dataset yielded, once again, $k = 3$. We show in Figures 5.55 and 5.56 the SVD for dataset. The Silhouette profiles are reported in Figures 5.57 and 5.58. To conclude, the plots for the Cross-validation and the Bi-Cross-Validation with simple residuals are presented in Figures 5.59 and 5.60 for 1995 and in Figures 5.61 and 5.62. We see another confirmation of the fact that Cross-Validation and Bi-Cross-Validation seems to fall short of providing us with a definite value for k. However, after pooling the various techniques and presenting the result to the panel of experts in economics and industrial management, we decided to maintain the same value.

We report the archetypes found for 1995 in Figures 5.63, 5.64 and 5.65 and the induced networks in Figures 5.66, 5.67 and 5.68. As far as 2011 is concerned, we report the archetypes 5.69, 5.70 and 5.71 and the induced networks in Figures 5.72, 5.73 and 5.74.

We begin by noting that the greater variability of the dataset is captured by the composition. Increasing the diversity of the economies under scrutiny allows for a better characterisation of the fundamental elements of the overall system. The first archetype presents major activations in the “R&D" and “Wholesale" sectors, suggesting a measure of the degree of reached economic development. The second archetypes presents major centrality in the “Financial intermediation", “Mining", “Basic metals", “Coke and nuclear fuels" and “Chemicals" sectors. In this case we notice the centrality of the extractive and basic industrial apparatus together with some financial shock vulnerability. Finally, the third archetype is mostly defined by the “Agriculture", “Coke and nuclear fuel" and “Chemicals" sectors. In this case, we notice a vulnerability on the extractive sector too, albeit related to an apparently less industrialised set of sectors.

Considering the results for 2011, we confirm what already seen in 1995. The decomposition succeeds in separating archetypes that are easy to interpret. We also notice again that the situation in 2011 presents some differences with respect to 1995, even though we do not attempt to find reasons as for why this might happen. In the first archetype, we see a sharp dominance of the “Mining" and “Coke and nuclear fuel" sector, to an unprecedented level of polarisation. The second archetype is essentially characterised by the “Wholesale" and “R&D" sector, suggesting a certain level of economic maturity. The third one, to conclude, is significantly marked by the “R&D", “Wholesale", “Transport" and “Financial intermediation", sectors that are key to assess the economic volatility of growth in an economic system.

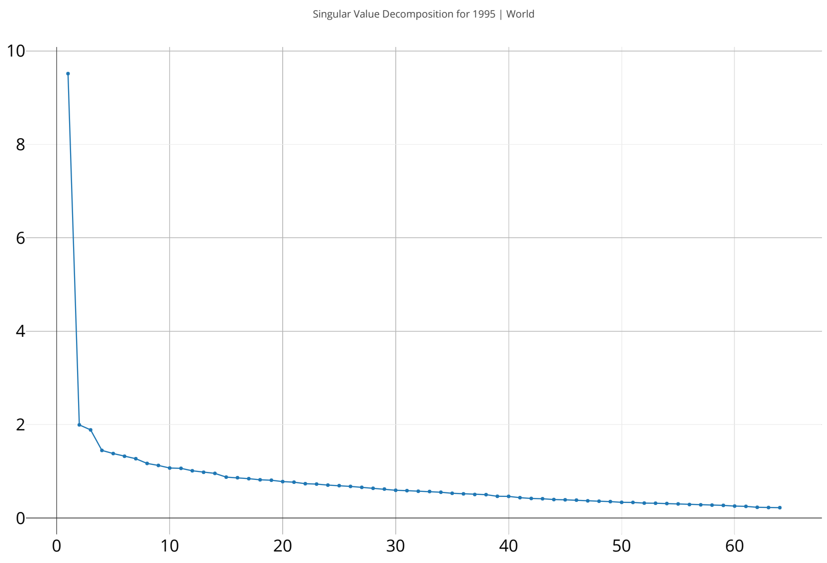

Figure 5.55: Singular Values of the SVD computed for the World cross-sectional dataset in 1995.

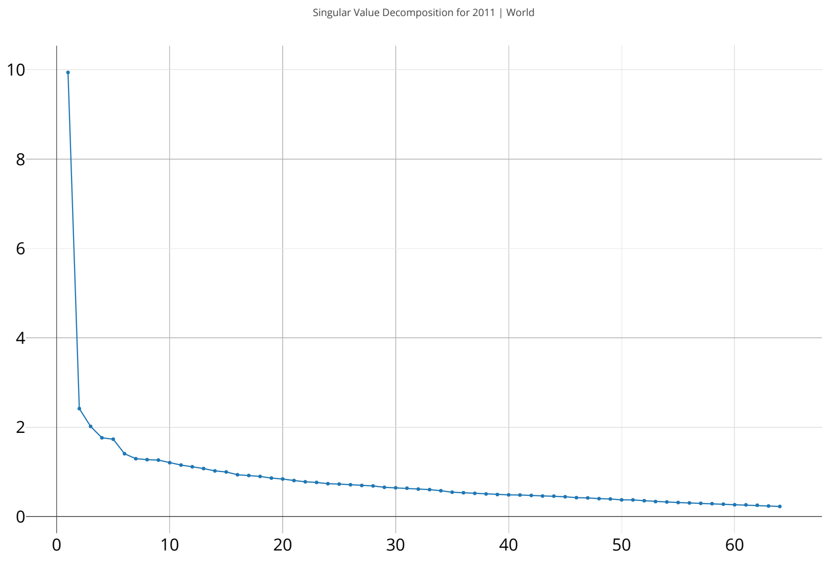

Figure 5.56: Singular Values of the SVD computed for the World cross-sectional dataset in 2011.

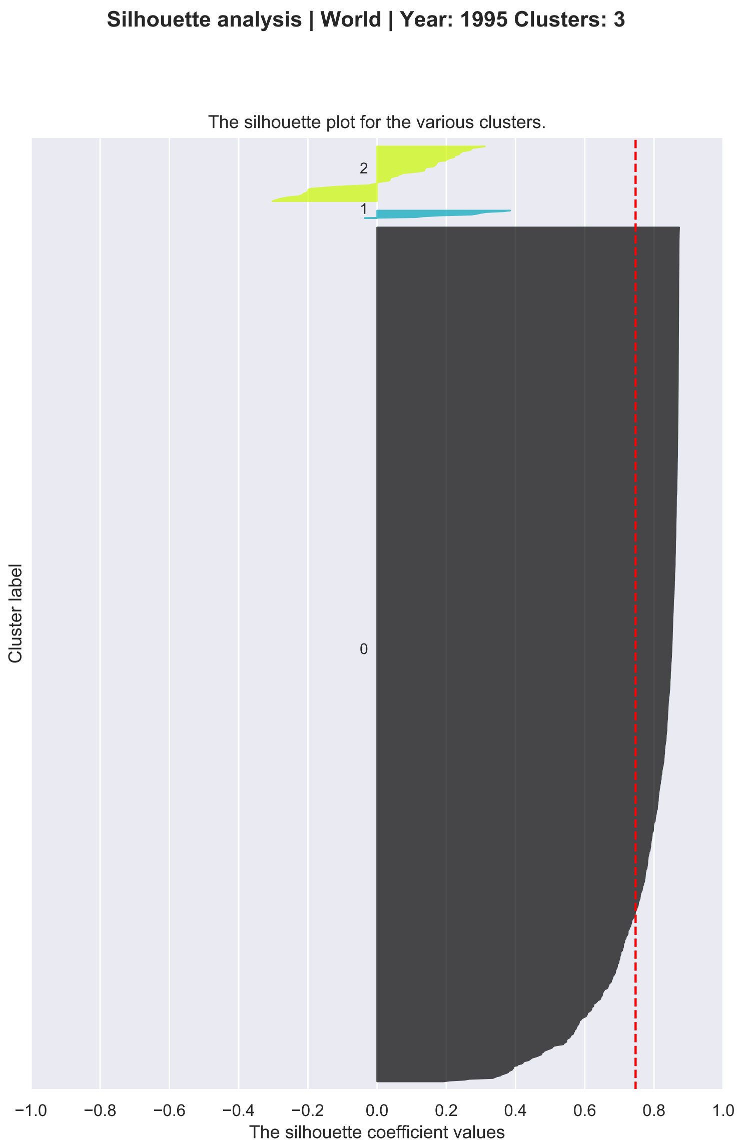

Figure 5.57: Silhouette profiles of the k-means algorithm with 3 clusters for the cross-sectional World dataset of 1995.

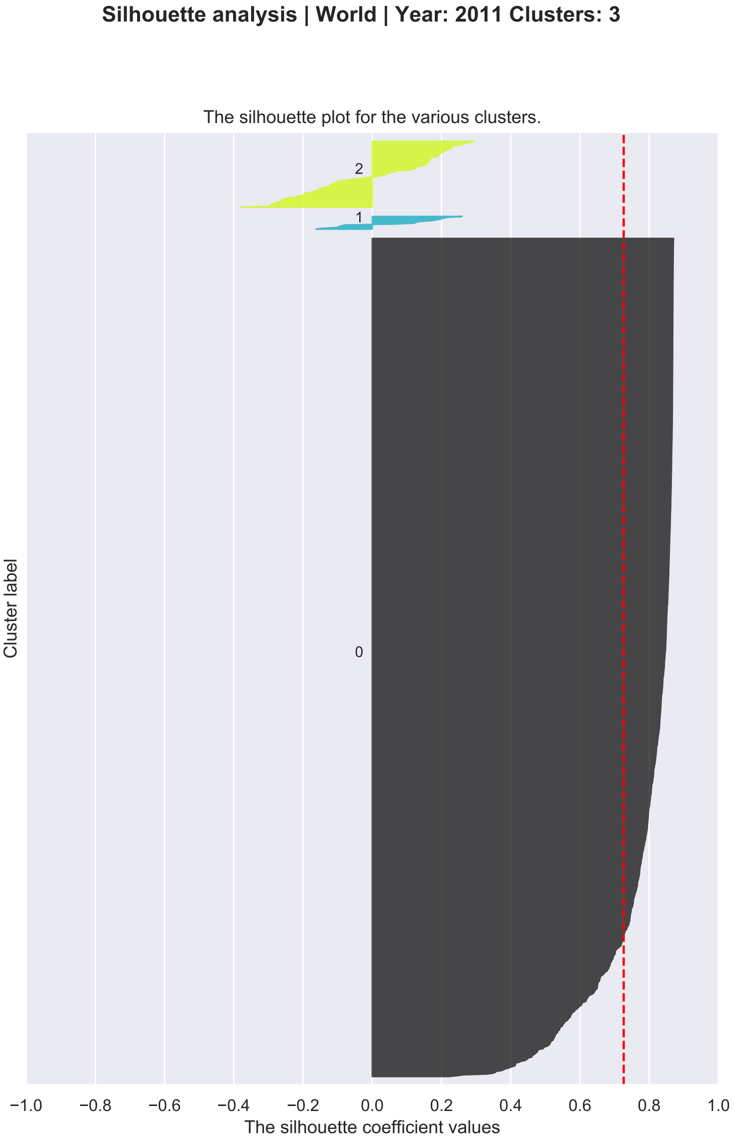

Figure 5.58: Silhouette profiles of the k-means algorithm with 3 clusters for the cross-sectional World dataset of 2011.

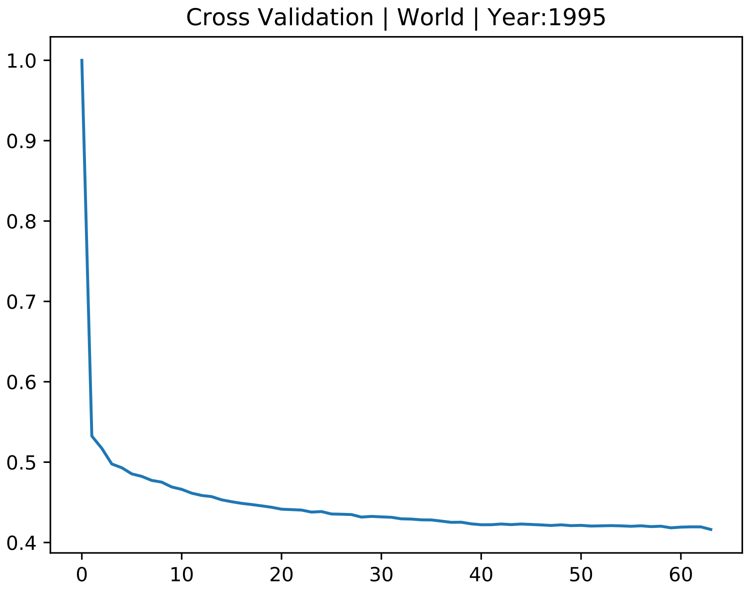

Figure 5.59: Cross-Validation for the cross-sectional World dataset for the year 1995.

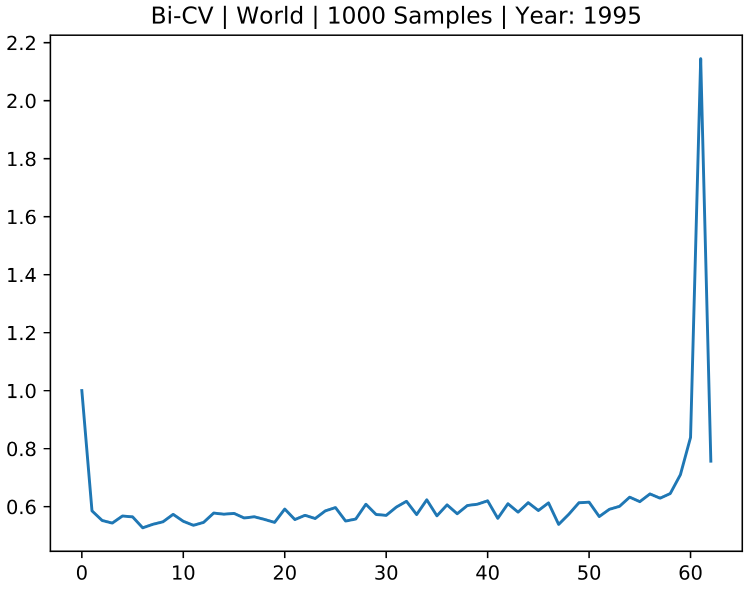

Figure 5.60: Bi-Cross-Validation with simple residuals for the cross-sectional World dataset for the year 1995. We considered (3×3) random submatrices and, for each rank, averaged the results of 1000 random trials. Data are normalised considering the MSE of a rank 0 approximation, *i.e.* the Frobenius norm of the sampled matrices alone.

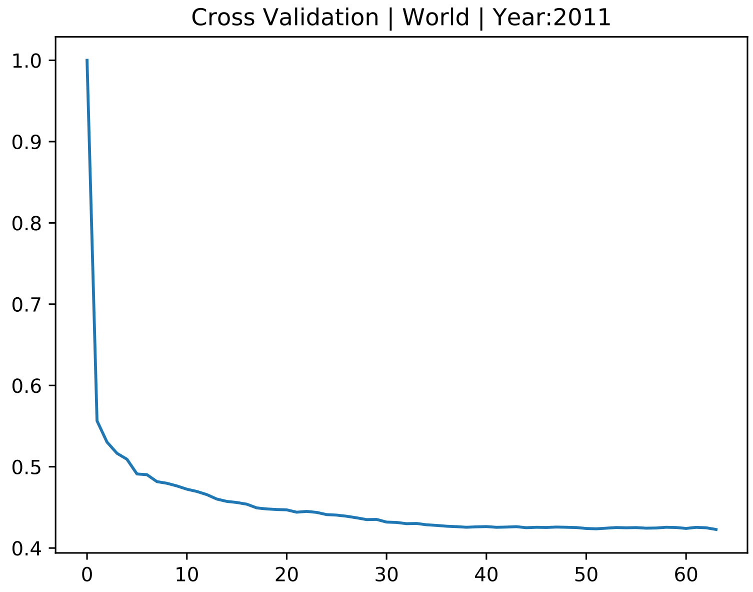

Figure 5.61: Cross-Validation for the cross-sectional World dataset for the year 2011.

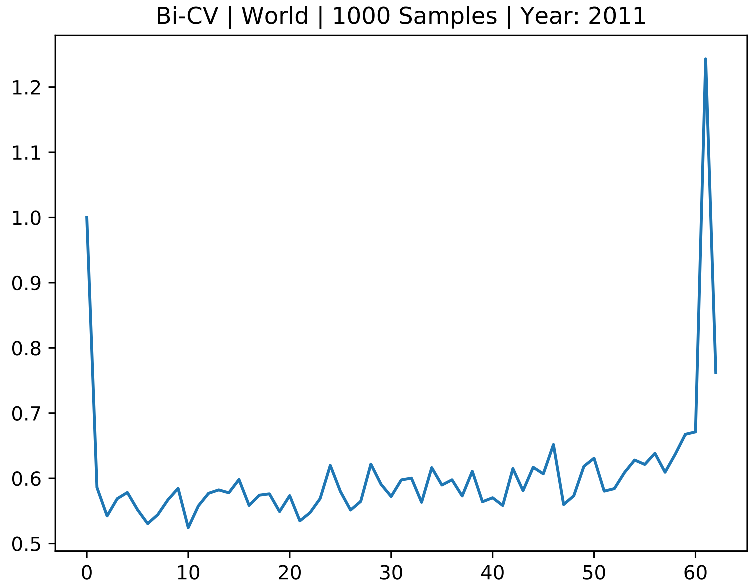

Figure 5.62: Bi-Cross-Validation with simple residuals for the cross-sectional World dataset for the year 2011. We considered (3×3) random submatrices and, for each rank, averaged the results of 1000 random trials. Data are normalised considering the MSE of a rank 0 approximation, *i.e.* the Frobenius norm of the sampled matrices alone.

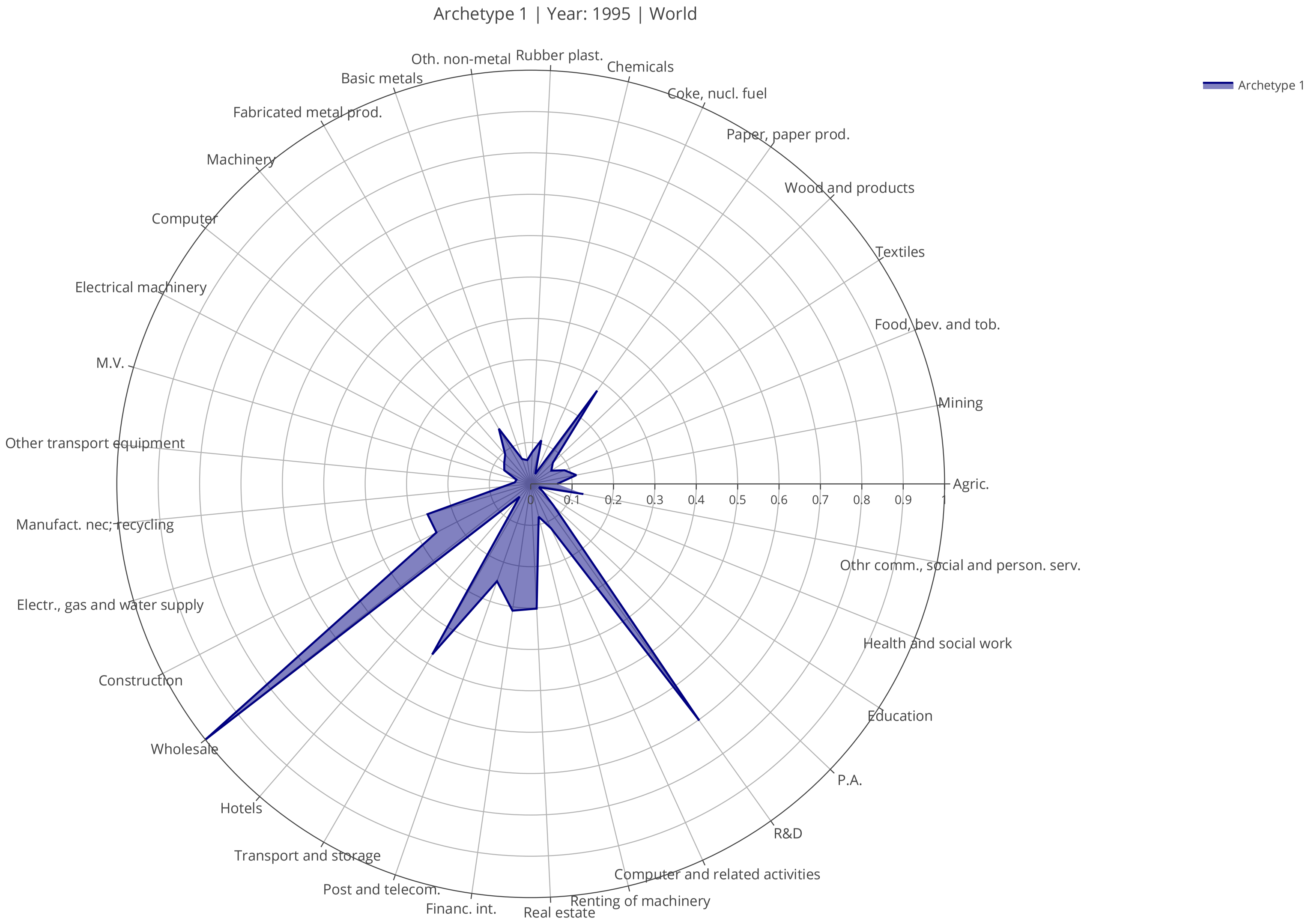

Figure 5.63: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional World dataset for the year 1995.

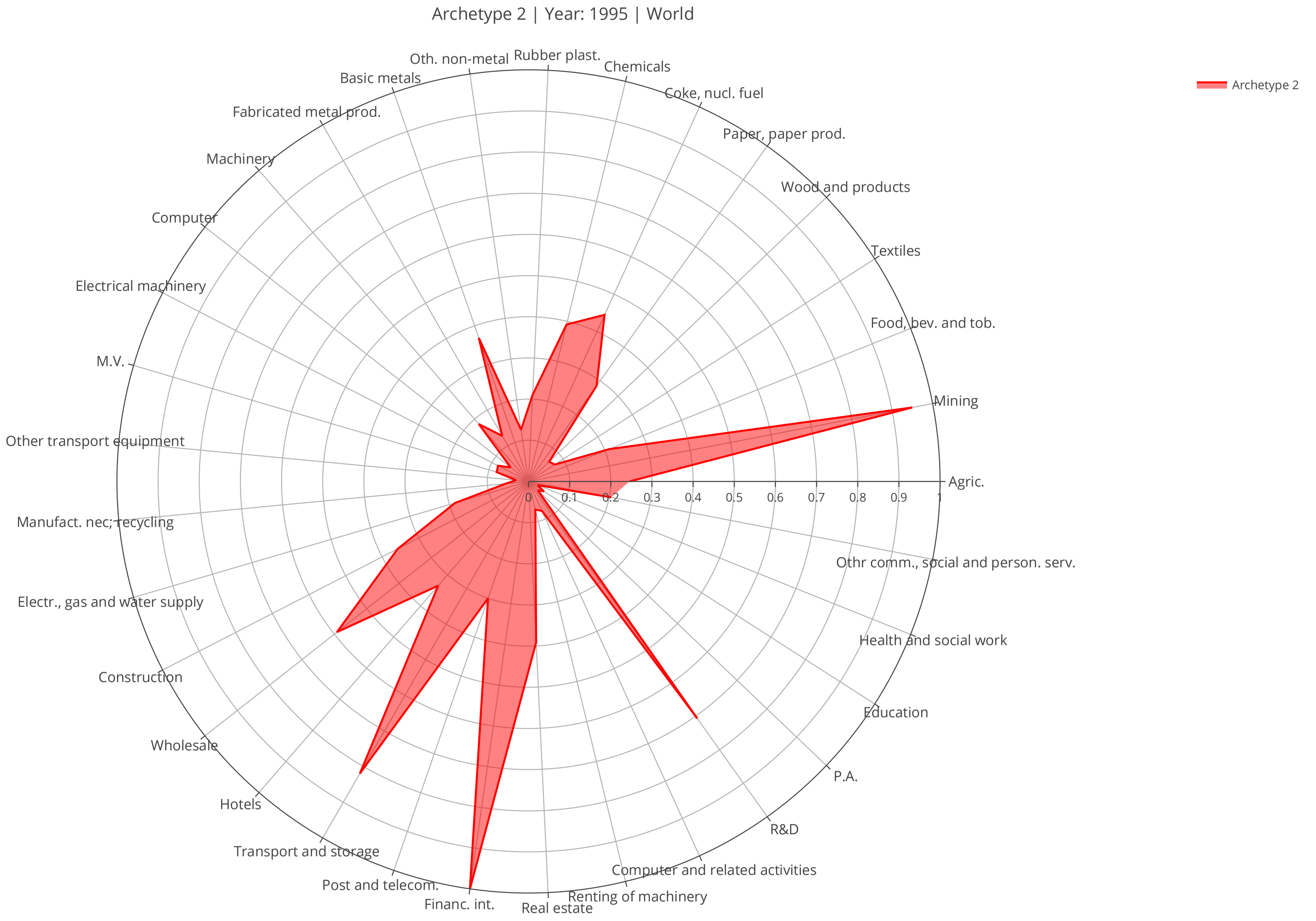

Figure 5.64: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional World dataset for the year 1995.

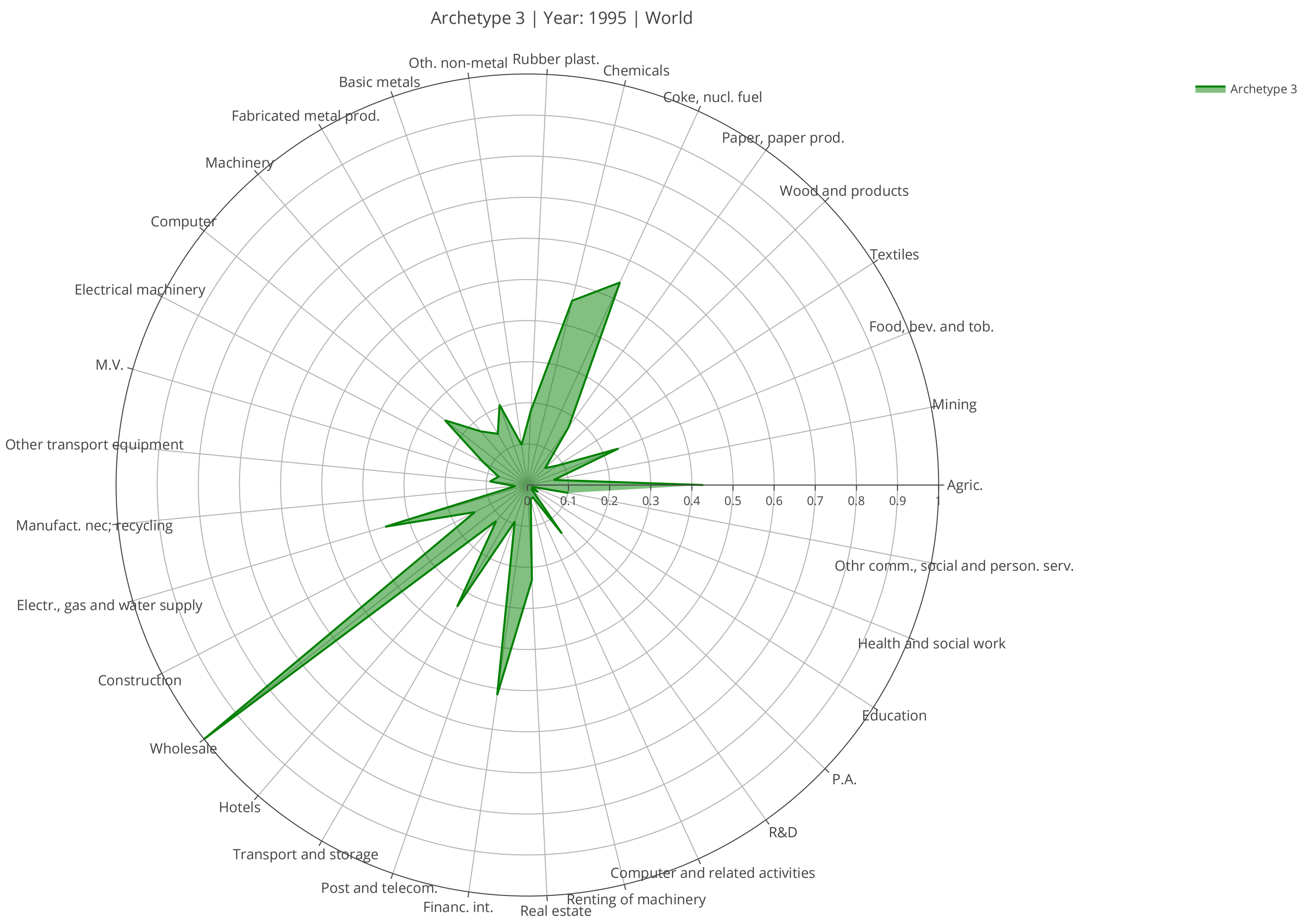

Figure 5.65: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional World dataset for the year 1995.

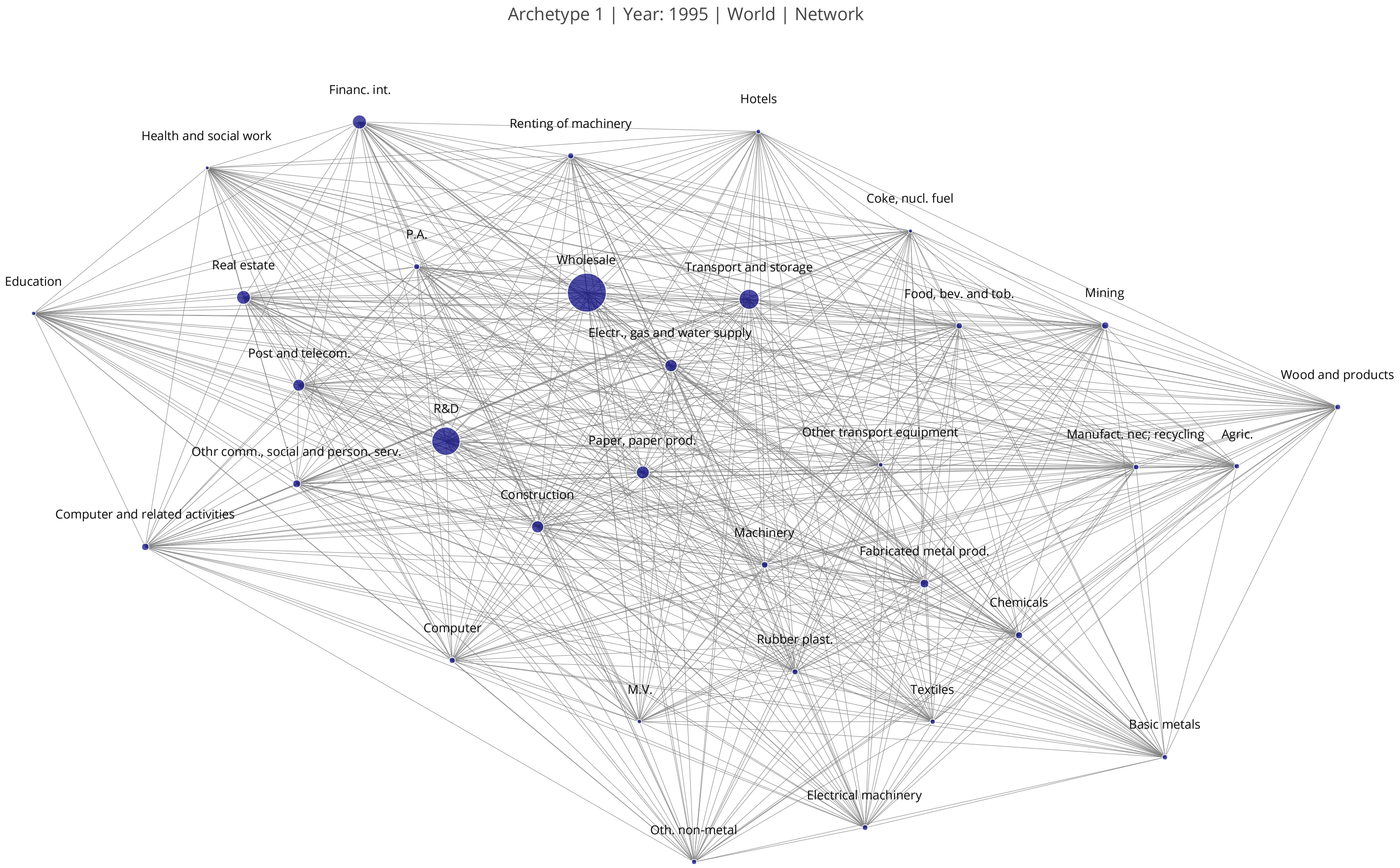

Figure 5.66: Network induced by the first archetype of the cross-sectional World dataset for the year 1995. The dimension of the nodes is proportional to their shock-centrality.

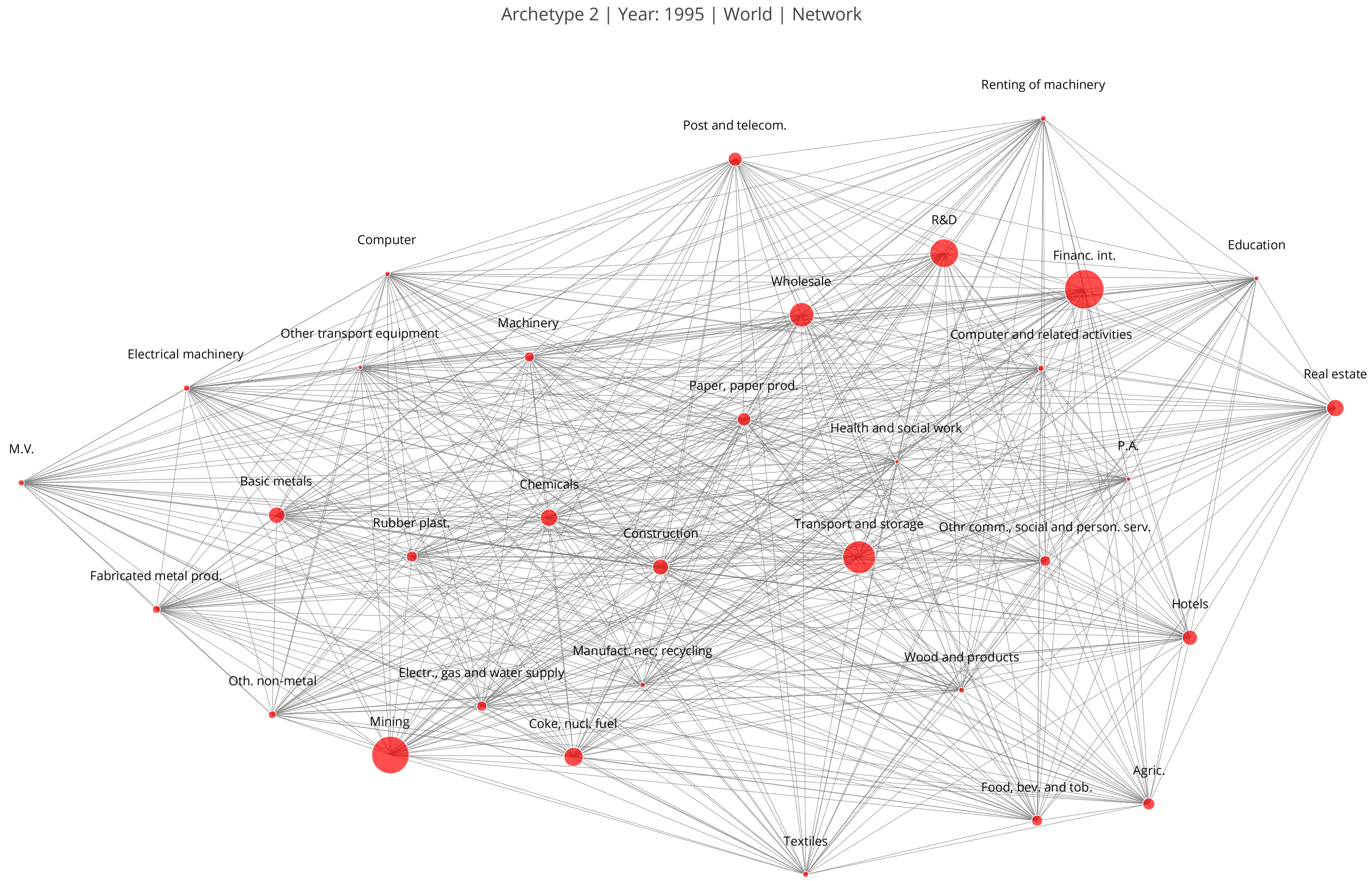

Figure 5.67: Network induced by the second archetype of the cross-sectional World dataset for the year 1995. The dimension of the nodes is proportional to their shock-centrality.

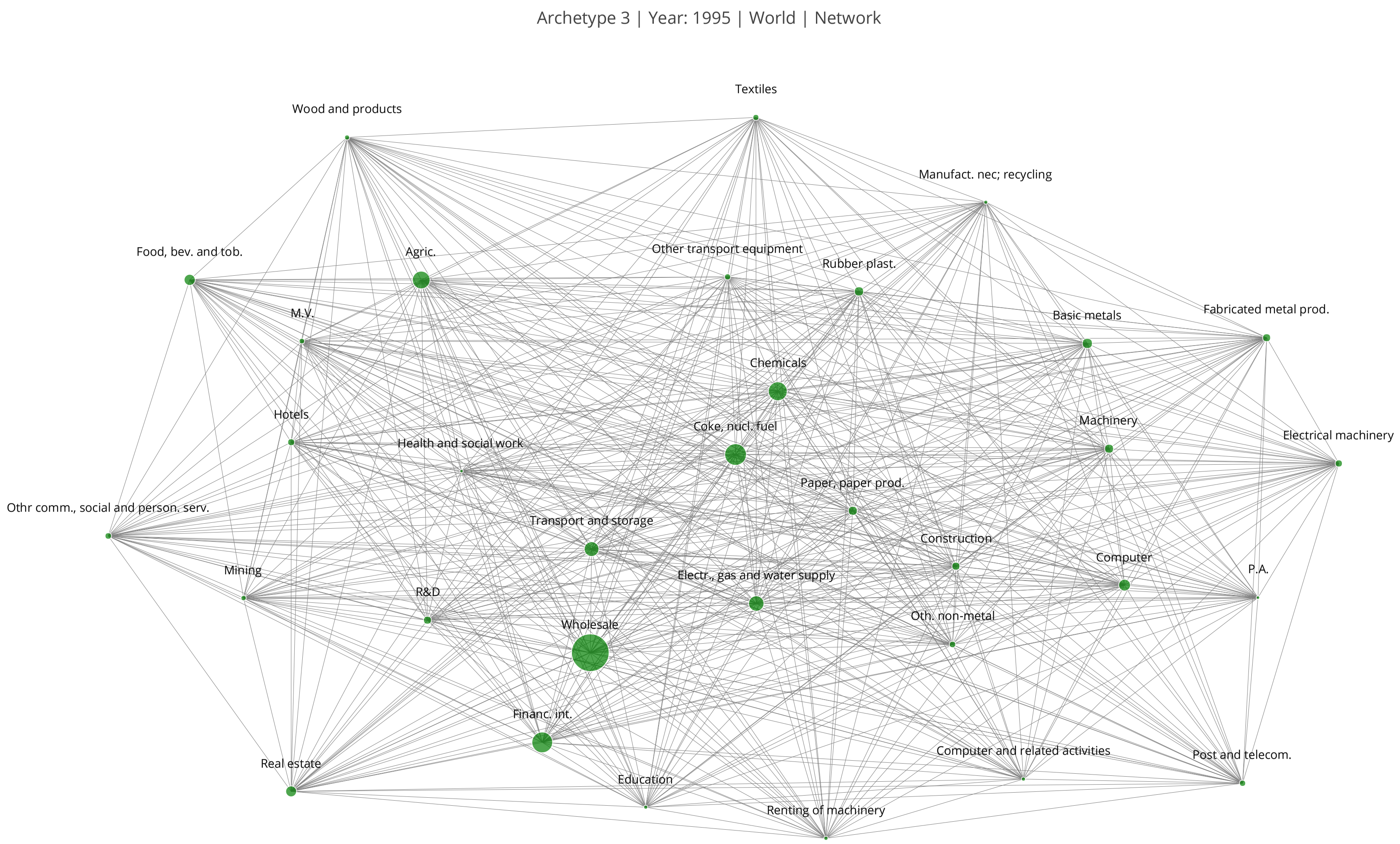

Figure 5.68: Network induced by the third archetype of the cross-sectional World dataset for the year 1995. The dimension of the nodes is proportional to their shock-centrality.

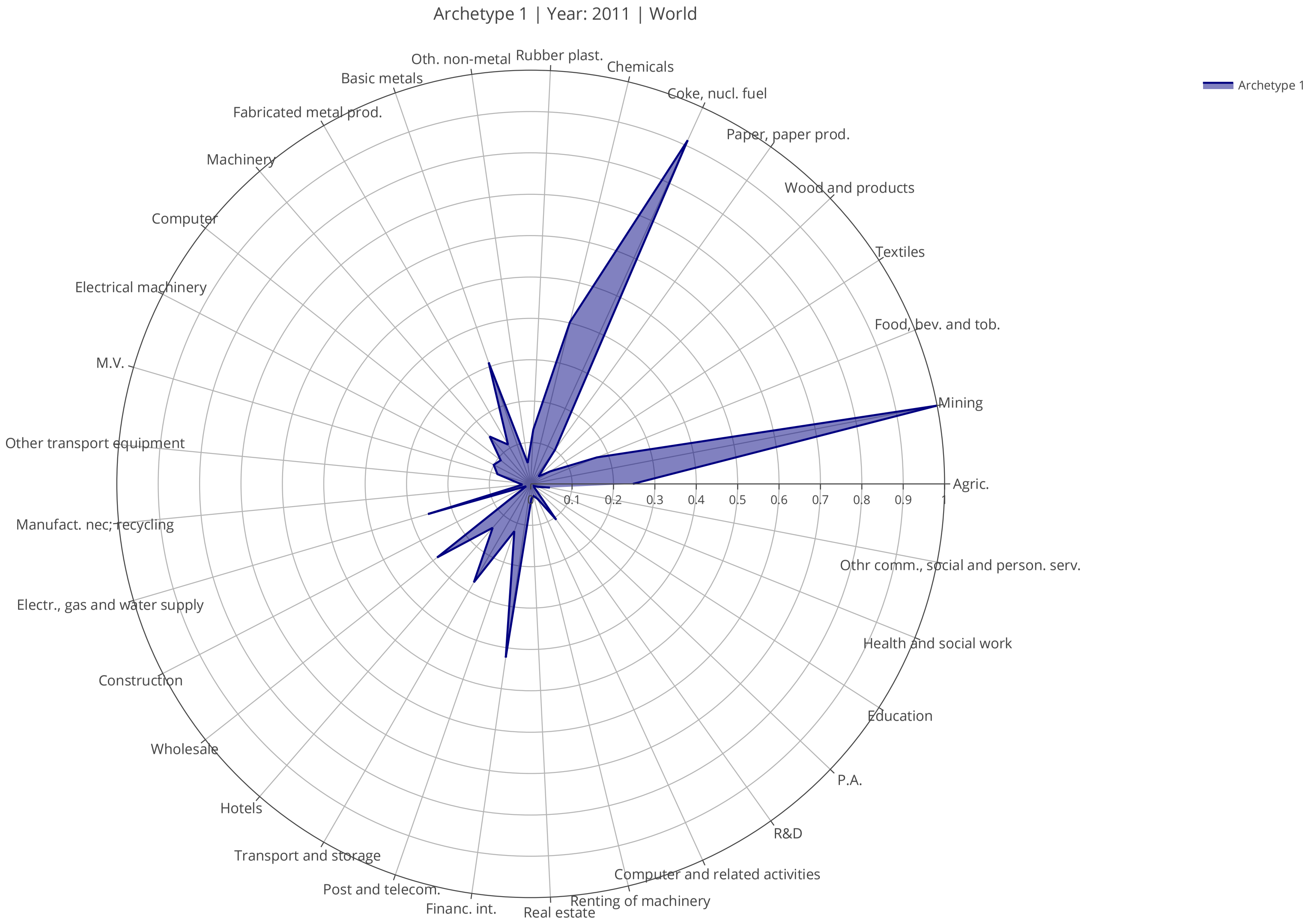

Figure 5.69: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional World dataset for the year 2011.

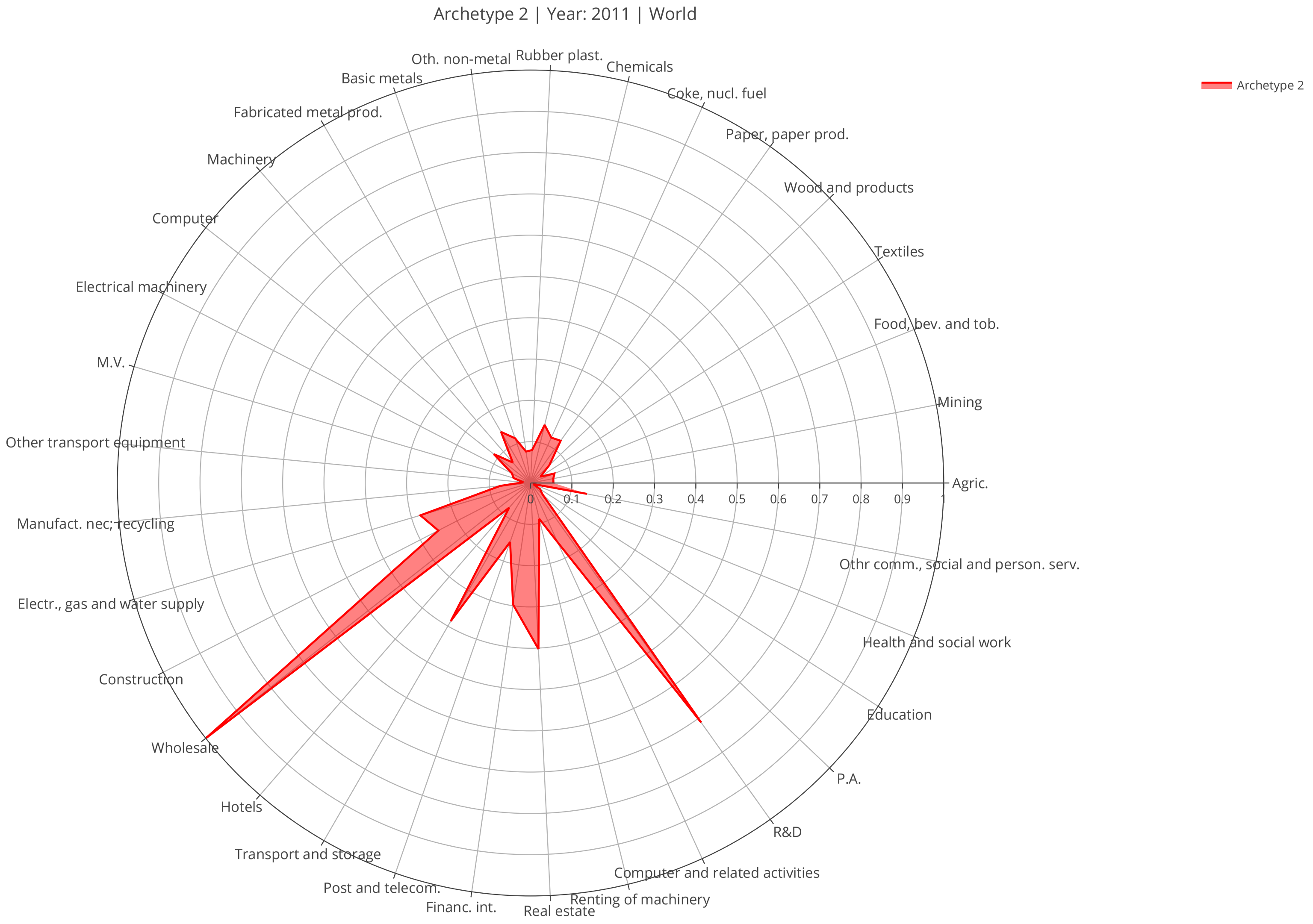

Figure 5.70: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional World dataset for the year 2011.

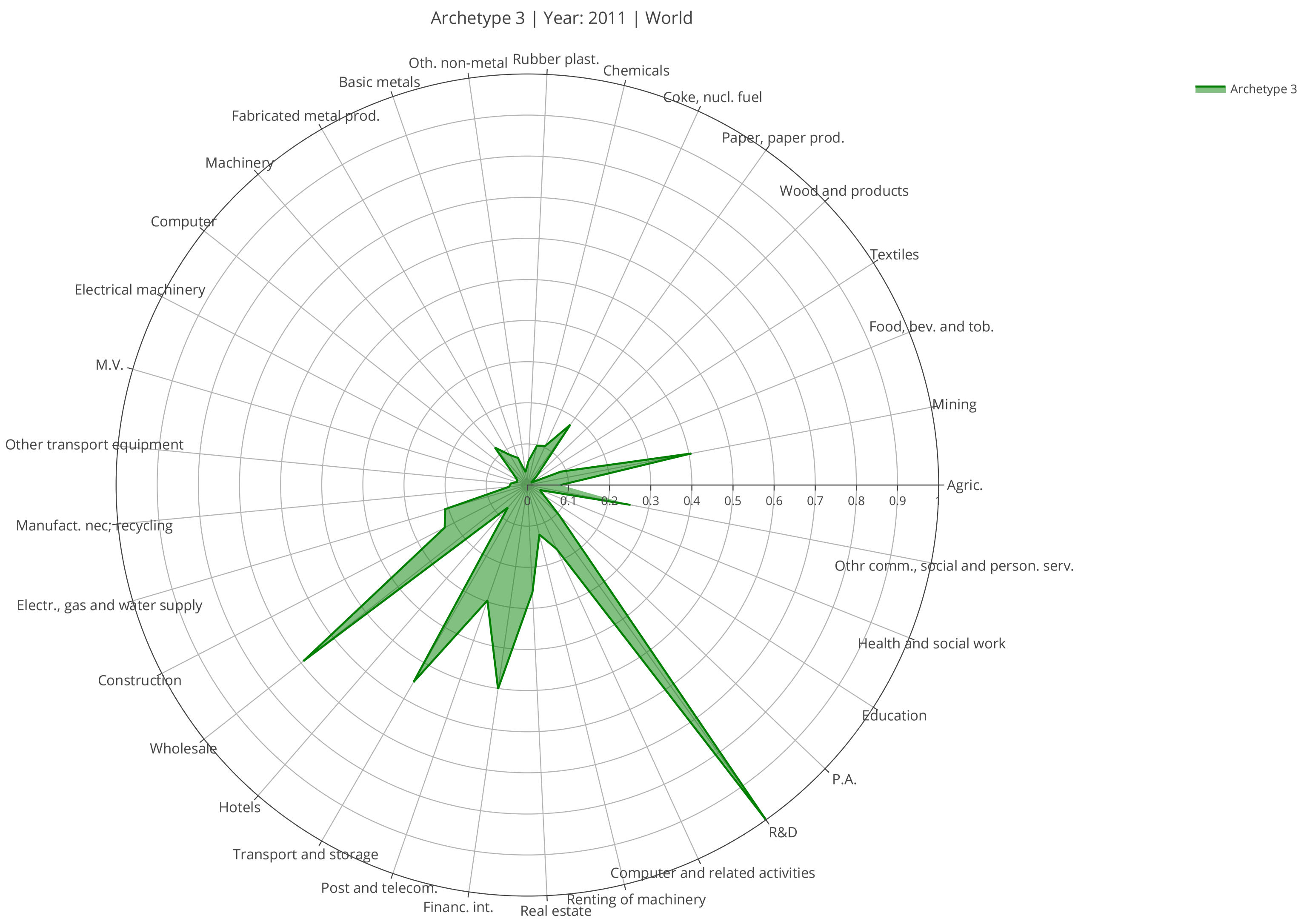

Figure 5.71: Radar chart showing the shock-induced random walk centrality measure for the archetypes of the cross-sectional World dataset for the year 2011.

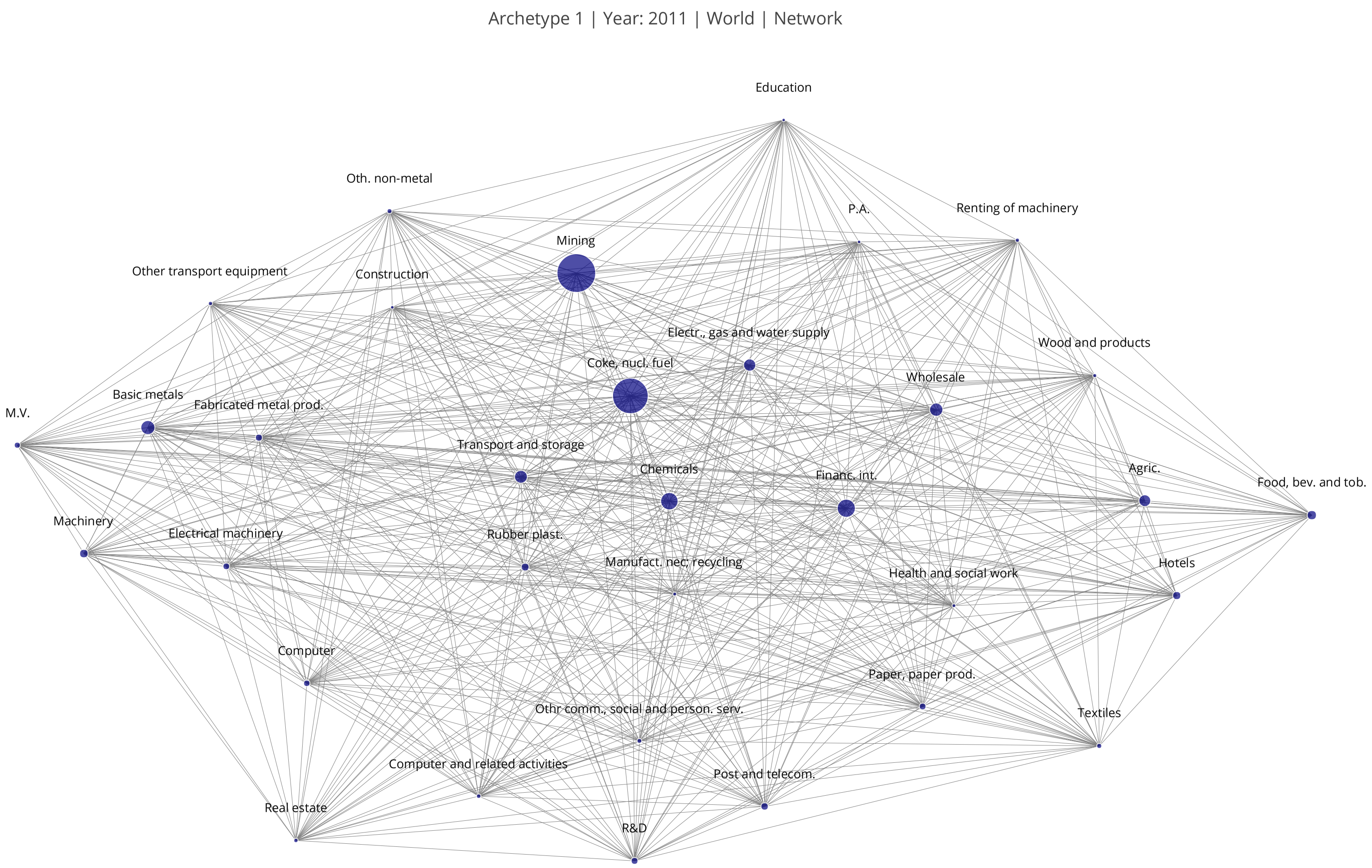

Figure 5.72: Network induced by the first archetype of the cross-sectional World dataset for the year 2011. The dimension of the nodes is proportional to their shock-centrality.

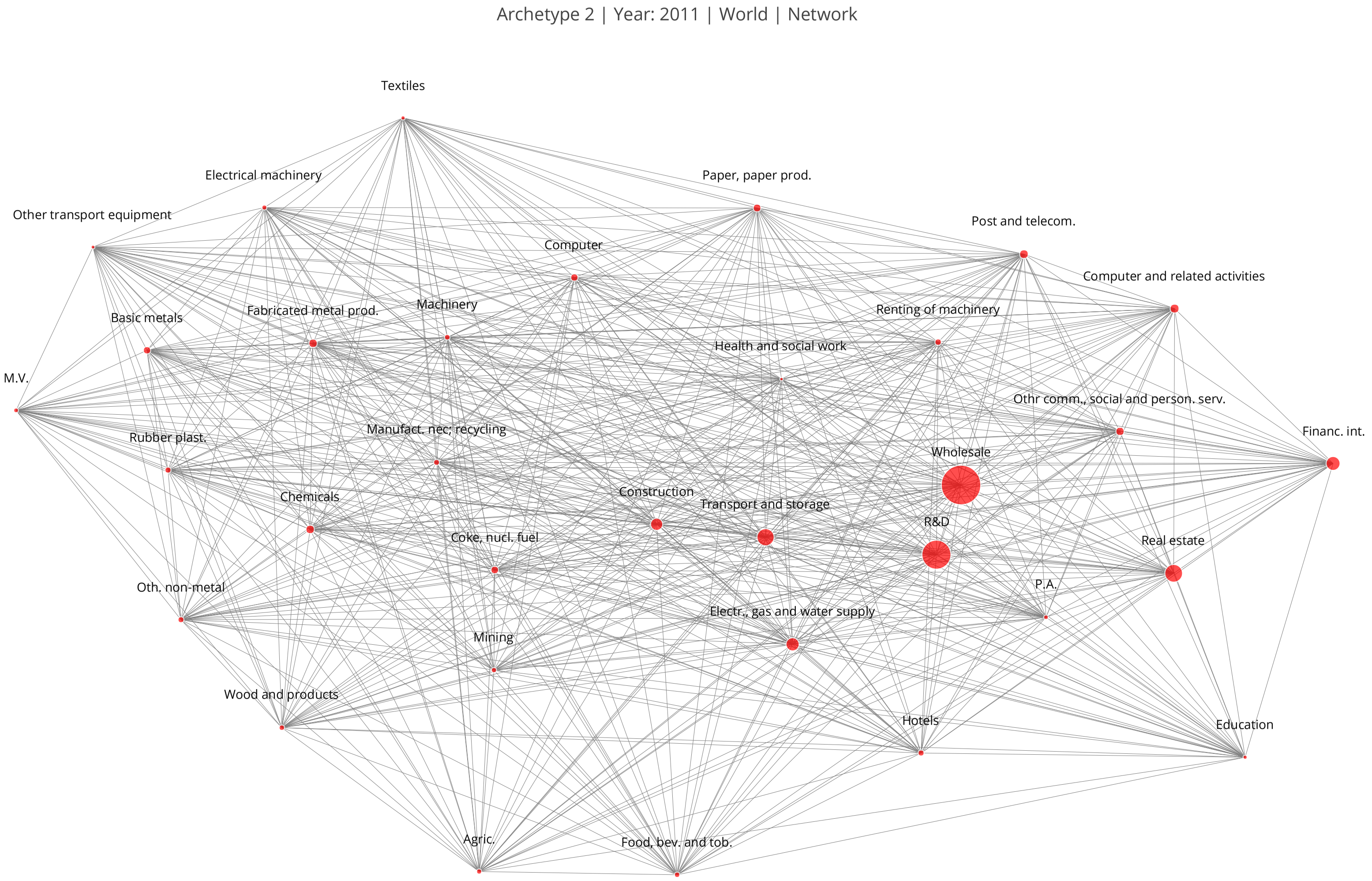

Figure 5.73: Network induced by the second archetype of the cross-sectional World dataset for the year 2011. The dimension of the nodes is proportional to their shock-centrality.

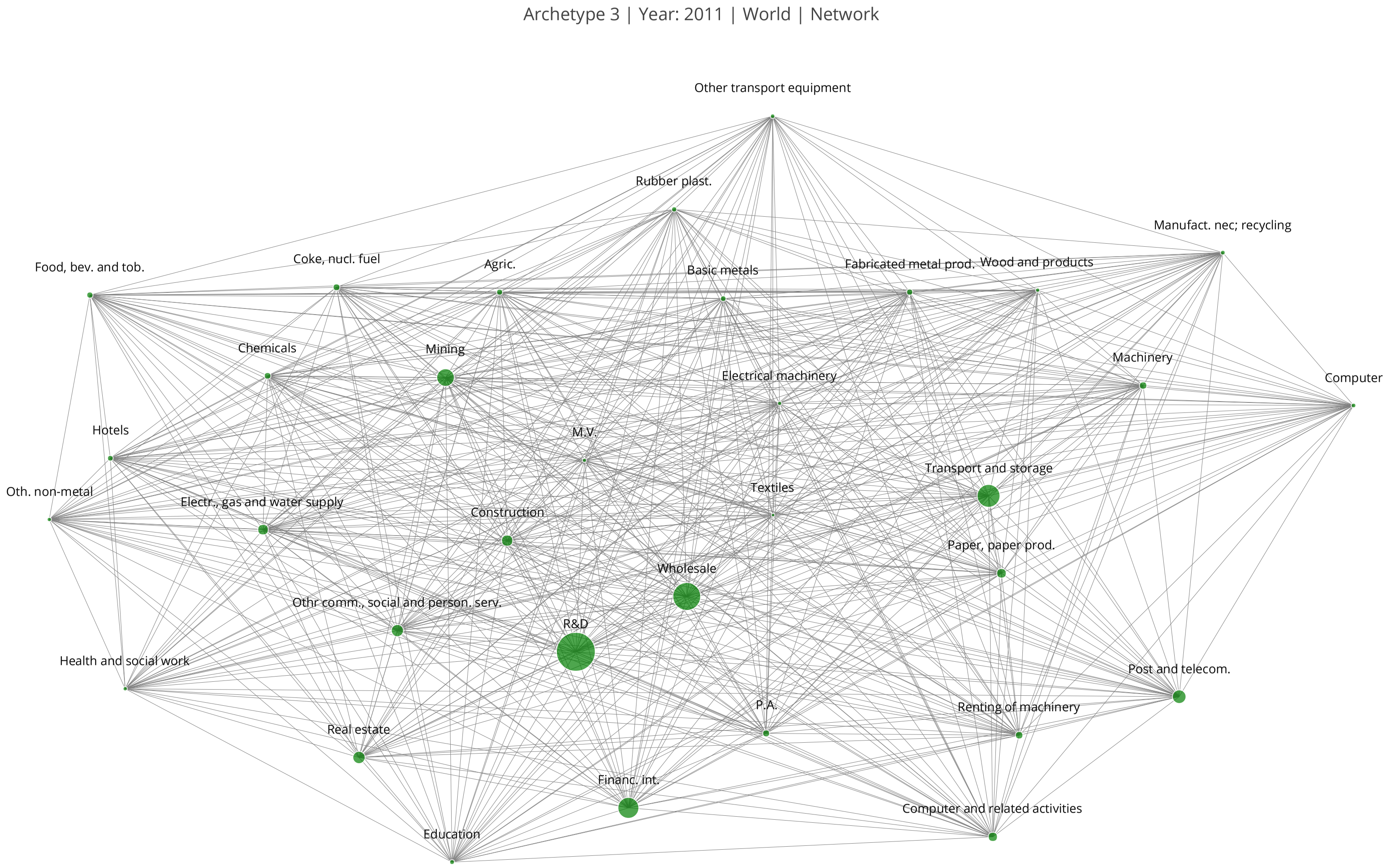

Figure 5.74: Network induced by the third archetype of the cross-sectional World dataset for the year 2011. The dimension of the nodes is proportional to their shock-centrality.

5.4.2 *Clustering for the cross-sectional World datasets*

The heatmaps with the activations are reported in Figures 5.75 and 5.78.

We focus first on 1995. We report in Tables 5.9 and 5.10 the centroids with the respective activations. The compositional clustering yields four clusters. In the first one we find Australia, Chile, Denmark, the Netherlands, Brunei Darussalam, Peru, Saudi Arabia. In the second clusters there are Austria, Belgium, Canada, Czech Republic, Estonia, Finland, France, Germany, Greece, Hungary, Island, Ireland, Israel, Italy, Japan, South Korea, Luxembourg, Mexico, New Zealand, Norway, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom, the United States of America, Argentina, Bulgaria, Brazil, China, Colombia, Cyprus, Hong Kong, Croatia, Indonesia, India, Lithuania, Malaysia, Romania, Russia, Singapore, Thailand, Tunisia, Chinese Taipei, Vietnam, South Africa and the Rest of the World. The third cluster comprises Latvia, Cambodia and Malta. Finally, the fourth cluster is formed by Morocco and the Philippines.

We notice that the second cluster is the most populated and that it is also the one represented by Croatia, whose components are somewhat equally spread out. This means that the majority of the countries presents a structure which is in between the archetypes. We expect shocks in those countries to be propagated at high velocities to the usual “R&D" and “Wholesale" sectors. This is also true for the other archetypes. This might suggests that those economies are well-connected, if we consider the industrial relationship between their sectors. The first cluster, represented by Saudi Arabia, is different: in this case, we expect, for instance, “Agriculture" to be a peripheral sector. Again, there may be different reasons as for why this happens. A possibility is that in the above countries this sector is not connected in input to many other, thus suffering less from supply-side shocks. The third cluster is represented by Cambodia. Here we have another case: we expect those countries to be less exposed as far as the basic industrial sectors are concerned. Finally, Morocco and the Philippines adhere almost perfectly to the third archetype: in this case, we can consider them as having as central sectors “Chemical", “Agriculture" and “Coke and nuclear fuel."

The Euclidean clustering in this case yields a different result. The number of clusters is greater and equals seven. As already states, we find neighbourhoods of countries considering the magnitude and order of the various activations. In the first cluster we find Australia, Chile, the Netherlands, Norway, South Korea, Spain, Sweden, Argentina, Brunei Darussalam, Peru, Romania, Saudi Arabia and South Africa. We have a small impact of the first two archetypes, thus expecting less centrality of the respective important sectors. The second cluster is composed by Australia, Denmark, Germany, Italy, Luxem-

Medoid	Archetype 1	Archetype 2	Archetype 3
Saudi Arabia	0.479	0.521	0.000
Croatia	0.391	0.319	0.290
Cambodia	0.448	0.000	0.552
Philippines	0.000	0.080	0.920

Table 5.9: Closure of the activations for the centroids obtained via a k-medoids clustering in the Aitchison geometry for cross-sectional World dataset of 1995.

Medoid	Archetype 1	Archetype 2	Archetype 3
Saudi Arabia	0.259	0.282	0.000
Poland	0.349	0.155	0.104
Island	0.330	0.027	0.274
Greece	0.239	0.243	0.206
Malaysia	0.090	0.318	0.340
Switzerland	0.532	0.003	0.054
Vietnam	0.052	0.012	0.842

Table 5.10: Centroids obtained via a k-medoids clustering in the Euclidean geometry for cross-sectional World dataset of 1995.

bourg, New Zealand, Poland, the United Kingdom, Russia. In this case, we have a slightly weaker participation of the second archetype and an increase of the first one. The third cluster is composed by Belgium, Estonia, France, Hungary, Island, Ireland, Latvia, Costa Rica, Cambodia, Lithuania, Malta and Singapore. In this case, countries suffer from a minimal participation of archetype 2. In the fourth cluster we have Canada, Czech Republic, Finland, Greece, Israel, Japan, Mexico, Portugal, Slovakia, Slovenia, Turkey, the United States, Bulgaria, Colombia, Cyprus, Croatia, the Rest of the World. We notice that these countries are somewhat equilibrated in their position in the space of the activations. The fifth cluster has as members South Korea, China, Indonesia, India, Malaysia, Morocco, Thailand, Tunisia, Chinese Taipei. We see this cluster present stronger activations towards the second and third archetypes, thus having an economic structure in which the “Financial intermediation", “Mining" and “Agriculture". The sixth cluster is formed Switzerland and Hong Kong. Here we have countries that shows strong activation in the “Wholesale" and “R&D" sectors. Vietnam and Philippines form the seventh and last cluster, showing a very strong activation for the third archetype.

It is interesting to see that both metrics groups clusters that are separated in terms of ratios: it means countries are scattered into the space in blobs.

Medoid	Archetype 1	Archetype 2	Archetype 3
Greece	0.317	0.358	0.324
Cambodia	0.533	0.467	0.000
New Zealand	0.000	0.463	0.537
Indonesia	0.413	0.000	0.587

Table 5.11: Closure of the activations for the centroids obtained via a k-medoids clustering in the Aitchison geometry for cross-sectional World dataset of 2011.

Medoid	Archetype 1	Archetype 2	Archetype 3
Finland	0.169	0.187	0.243
South Korea	0.548	0.172	0.074

Table 5.12: Centroids obtained via a k-medoids clustering in the Euclidean geometry for cross-sectional World dataset of 2011.

Let us consider the clustering for 2011. We report in Tables 5.11 and 5.12 the medoids for the Aitchison and the Euclidean metrics.

The compositional clustering finds four clusters. The first one is the most numerous and comprises Australia, Austria, Belgium, Canada, Chile, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Island, Ireland, Italy, Japan, South Korea, Lativa, Mexico, the Netherlands,Norway, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, the United Kingdom, the United States of America, Argentina, Bulgaria, Brazil, China, Colombia, Hong Kong, Croatia, India, Lithuania, Malta, Malaysia, Romania, Russia, Saudi Arabia, Thailand, Tunisia, Chinese Taipei, South Africa. Given the obvious differences between the above countries, it is of interest to see that those countries are represented by Greece, a point whose activations are somewhat equilibrated between the three archetypes. This means that in the above countries we find an equivalent importance, in terms of supply-side shock propagation, of the extractive, tertiary and “Financial intermediation" sectors. It is also interesting to notice that many of the countries in the cluster are actually members of the OECD and the others show geographical contiguity. The second cluster contains Luxembourg, Turkey, Costa Rica, Cyprus, Cambodia, Philippines, Vietnam. Those countries are supposedly countries whose shock-nodes are concentrated on the first two archetypes, thus excluding, somewhat surprisingly for the case of Luxembourg, “Financial intermediation" (although it should be noted we are considering supply-side domestic shocks). The third cluster is composed by New Zealand, Switzerland and Singapore, which all share the shock-central nodes of the second and third clusters, namely the “Financial intermediation" node, which is very much understandable for Switzerland and Singapore and the

“R&D" and “Wholesale" nodes, perhaps more relevant for New Zealand. Finally, Brunei Darussalam, Indonesia, Morocco, Peru and the Rest of the World are active on the first and third archetype and therefore their “Mining" sector, as well as the “Coke and nuclear fuel" is not very central according to our shock-centrality point of view.

The Euclidean clustering groups countries into two clusters: we expect countries belonging to the same group to be close to each other in the space. The first cluster is formed by Australia, Austria, Belgium, Canada, Chile, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Island, Ireland, Israel, Italy, Latvia, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, Switzerland, the United Kingdom, the United States of America, Argentina, Bulgaria, Brazil, Brunei Darussalam, Colombia, Costa Rica, Cyprus, Hong Kong, Croatia, Indonesia, Cambodia, Lithuania, Malta, Peru, Romania, Russia, Saudi Arabia, Singapore, South Africa, the Rest of the World. In this case, we have a low but equally spread set of activations, indicating a somewhat equilibrated yet weak participation of the archetypes. The second one is formed by Japan, South Korea, Turkey, China, Indonesia, Morocco, Philippines, Thailand, Tunisia, Chinese Taipei and Taiwan. In this case, countries are more concentrated towards the first archetype, with high activation.

We conclude the analysis for the World dataset by showing in Figures 5.76 and 5.77 the clustering yielded by the two different metrics in 1995 and the equivalent results for 2011 in Figures 5.79 and 5.80.

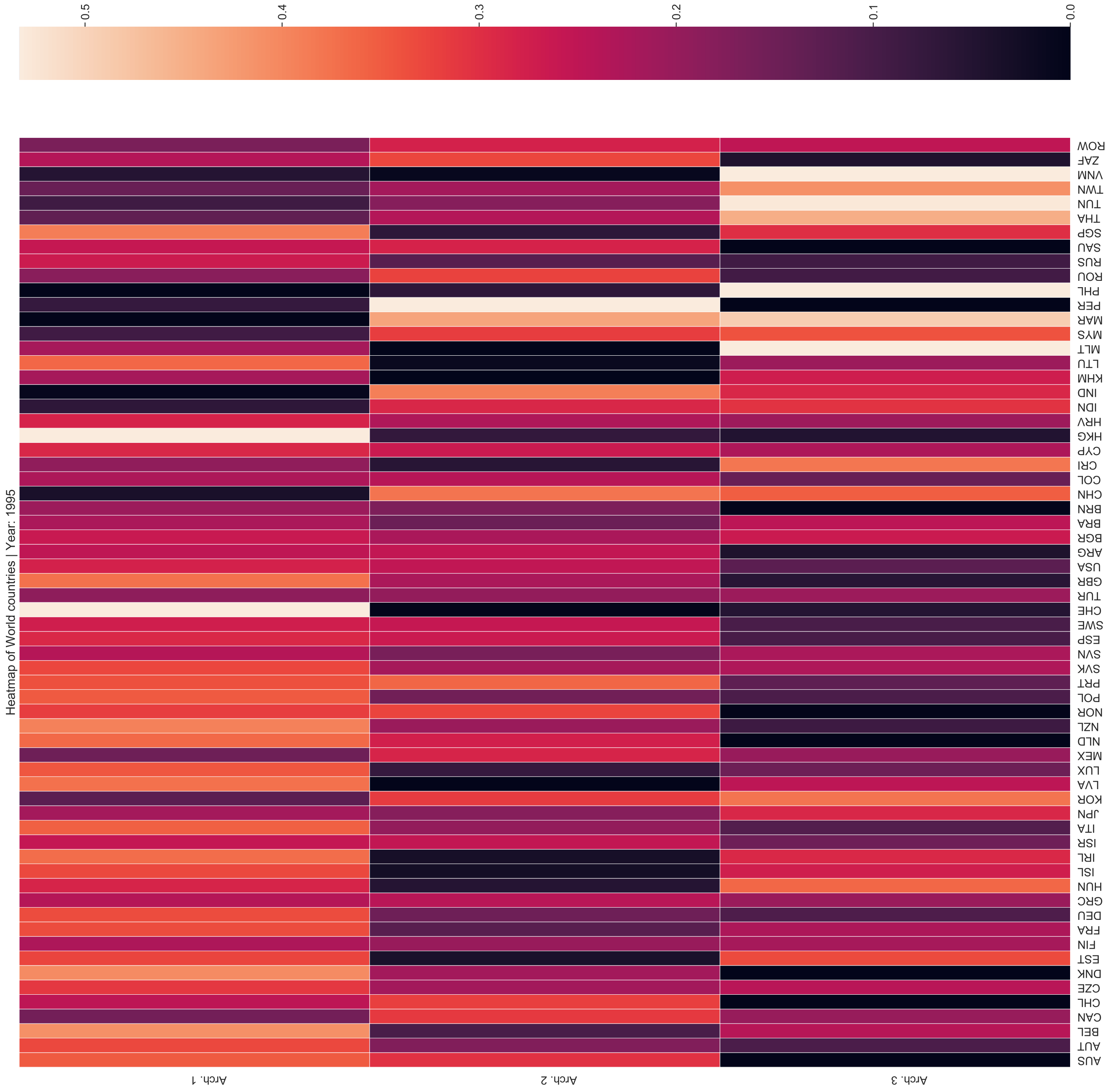

Figure 5.75: Heatmap of the activations of the countries in the cross-sectional World dataset for the year 1995.

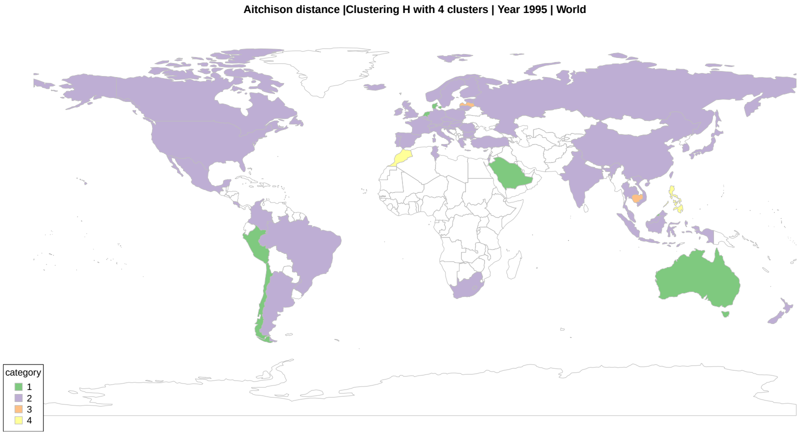
Aitchison distance |Clustering H with 4 clusters | Year 1995 | World
category
1
2
3
4

Figure 5.76: Clustering obtained considering the cross-sectional World dataset for the year 1995, via k-medoids and the Aitchison metric, with the number of clusters induced by the best Silhouette score.

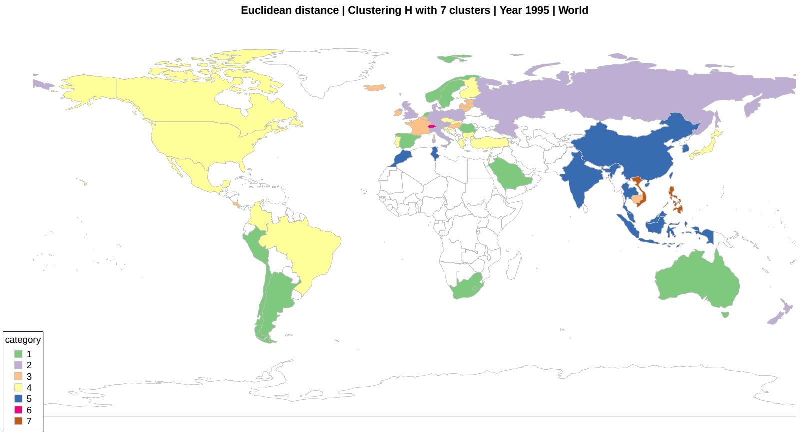
Euclidean distance | Clustering H with 7 clusters | Year 1995 | World
category
1
2
3
4
5
6
7

Figure 5.77: Clustering obtained considering the cross-sectional World dataset for the year 1995, via k-medoids and the Euclidean metric, with the number of clusters induced by the best Silhouette score.

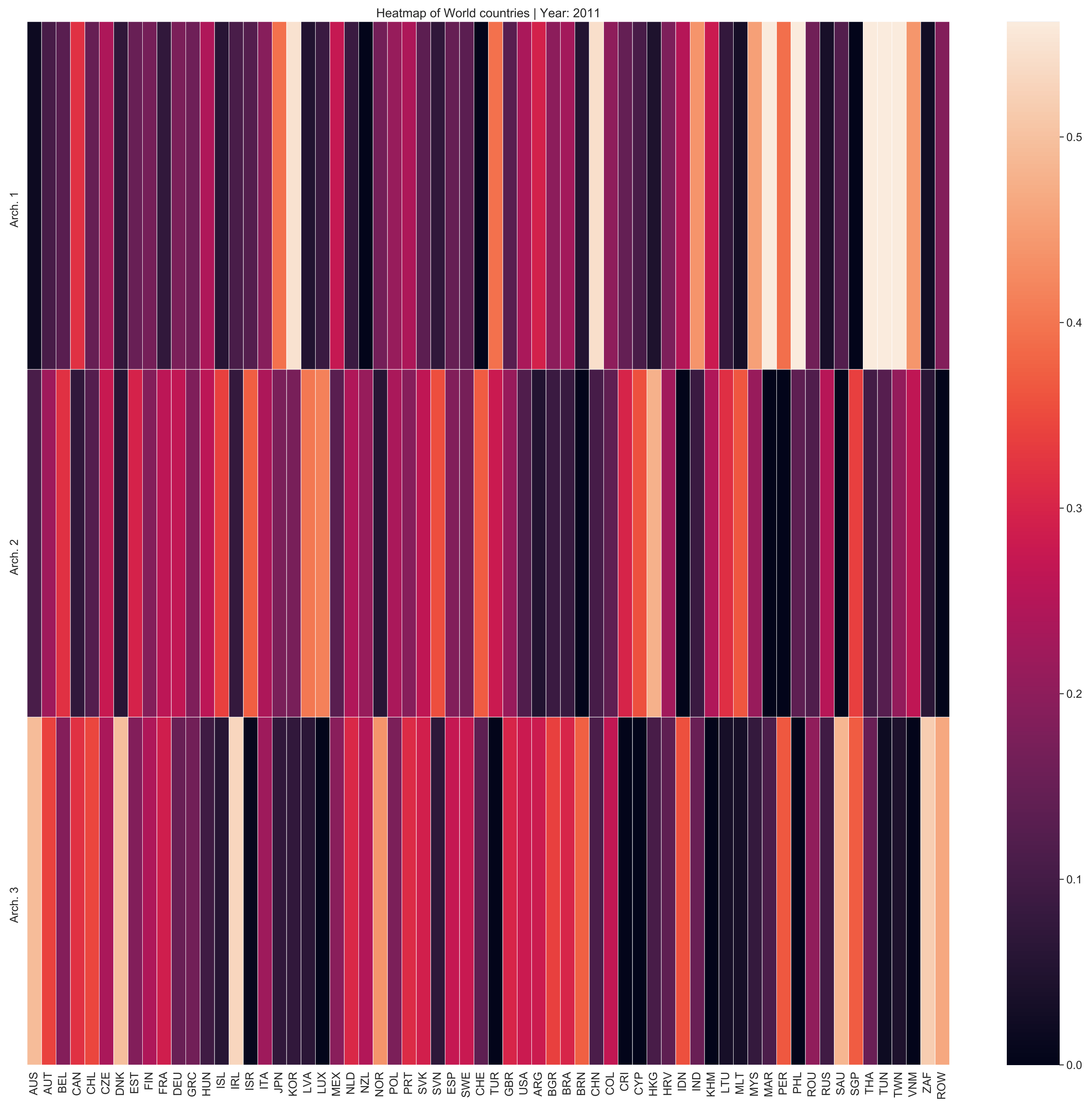

Figure 5.78: Heatmap of the activations of the countries in the cross-sectional World dataset for the year 2011.

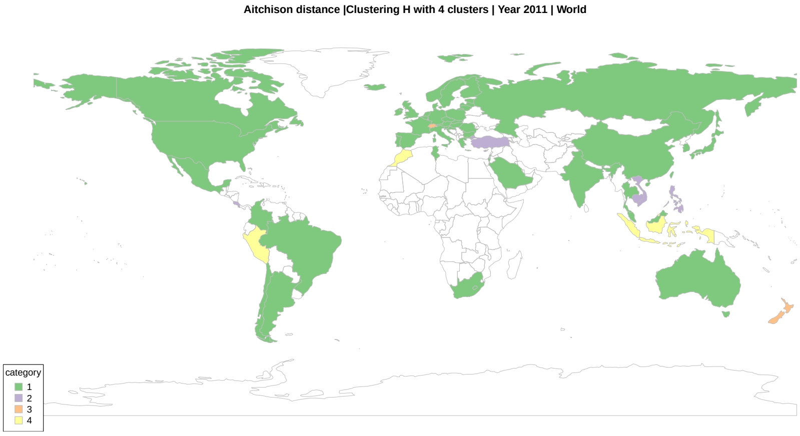
Aitchison distance |Clustering H with 4 clusters | Year 2011 | World
category
1
2
3
4

Figure 5.79: Clustering obtained considering the cross-sectional World dataset for the year 2011, via k-medoids and the Aitchison metric, with the number of clusters induced by the best Silhouette score.

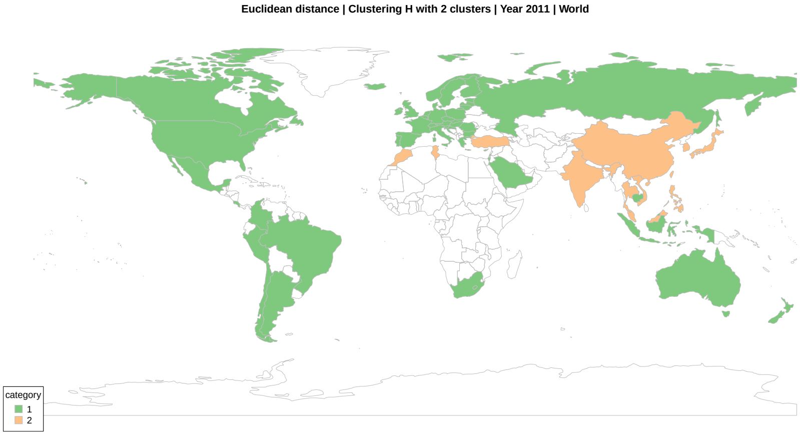
Euclidean distance | Clustering H with 2 clusters | Year 2011 | World
category
1
2

Figure 5.80: Clustering obtained considering the cross-sectional World dataset for the year 2011, via k-medoids and the Euclidean metric, with the number of clusters induced by the best Silhouette score.

5.5 SOME CONSIDERATIONS ON THE ANALYSIS

The cross-sectional datasets for 1995 and 2011 for the EU, the OECD and the World datasets have been analysed. The first thing we have noticed is that the identification of the archetypes proved to be more ambiguous in the EU and the OECD datasets. This might be due to the level of sectorial aggregation of the data, which does not allow for a clear identification of the archetypes, as it is suggested by the fact that when we broaden the scope of the analysis such issue seems to be less impactful. Another possibility might be that the shock propagation model we proposed in Section 2.4.0.1 fails to provide meaningful insight on the economies at this level of aggregation. However, the sectors that have usually emerged as central are in line with our expectation, confirming the validity of the approach. We also note that the rank estimation process yielded the same results for the three datasets, albeit with some ambiguities. Rank estimation is one of the most complicated aspects of NMF, especially when such estimation is expected to capture patterns of interpretable meaning. It is for this reason that we decided to be as conservative as possible, pooling the whole spectrum of employed techniques and assessing the validity of the results with a pool of experts. Another possibility could have been that of increasing the rank of the decomposition up to the maximum meaningful number. However, such an approach would deviate from our endeavour of finding a purely quantitative criterion: clustering is useful if it automates a representation of data which is then found to be meaningful. However, the consistency of some archetypes throughout the whole datasets is encouraging, as it shows that we are capturing some actual features of the economic systems we wish to describe. This holds in particular when considering the World dataset in comparison with either the EU and the OECD ones: we are capturing at least in part the structure of countries that are member of the latter two organisations and then noticing different structures in other archetypes. In this sense, despite the limitations given by the presence of archetypes that were not clearly distinguishable, the approach seems to be valid. It would be of natural interest to try and apply it on the other two international Input-Output tables datasets, the WIOD and the Eurostat sets, to compare the results.

Considering the clusterings, we notice that the Aitchison distance appears to be better suited for this kind of analysis. This is probably because, by construction, compositional clusterings aggregate considering *ratios* and this allows for an easier interpretation of the groups we find, especially if it is not straightforward to understand the role particular magnitudes play, as in this case. Yet, we also notice that it is important to explicitly set what we deem important before performing the clustering as the results can differ greatly according to the metric employed. We also notice that the clustering induced by the Aitchison metric provides clusters that are more homogeneous with the respect to those obtained considering the Euclidean distance.

Finally, we observe that the main limit of the framework here presented is given by the superposition of the issues affecting the two separate techniques. Despite finding clusters that effectively group together countries with geographical and economic similarities, the interpretation of the results remain somewhat elusive in various cases. This probably constitutes the most interesting finding of this work. On the one hand, this might be due to the fact that subclusters are being wrongly put together. This might be due to an erroneous estimation of the rank k of the NMF or to the scarce efficacy of the k-medoids algorithm in this framework. On the other one, it should be noticed that this is a clustering problem. This means that we do not have access to some “ground truth" and therefore we need to rely on intuition, aided by some measures of quality of the clustering as the one we have used, to assess how well the algorithm is scoring. As a matter of fact, formulating expectations regarding the output is a task beyond the scope of this work. Yet, it is in this lack of expectation that we believe this work can be of use, as it provides with an analysis of various countries based exclusively on their domestic industrial relations and can be interesting when compared with other quantitative or qualitative research methods. We believe this will become increasingly true in the future, as Input-Output Tables datasets gains in accuracy and detail

6 CONCLUSION AND FUTURE DEVELOPMENTS

We have presented a novel methodological framework aimed at clustering Input-Output tables. Given the complexity of the problem, our strategy has been that of decoupling the problem into two parts, separating clustering from interpretation. This was accomplished by employing non-negative matrix factorisation, a dimensionality reduction technique whose non-negativity constraints suited the data. The main drawback of this technique, however, is the estimation of the dimensionality of the low-rank decomposition we are willing to pursue. Despite pooling different techniques and implementing both a Cross-Validation and a Bi-Cross-Validation scheme, the results were mixed. In this sense, it might be interesting to embed the technique into a Bayesian framework in order to exploit prior knowledge in a more rigorous fashion and, possibly, to obtain a posterior distribution for the rank k. An alternative might be that of introducing sparsity constraints in the loss function of the NMF: the tuning of the parameters could then be conducted applying the same Cross-Validation algorithms already developed and implemented in this work. The identification of different archetypes also suffered from some ambiguities, especially in the EU case. This may due to the fact the countries belonging to the EU are similar to each other. Yet, this is probably the dataset that allows for the greater level of speculation because of its tight geographical scope. It might therefore be of interest to analyse the experimental Eurostat dataset and assess whether results show variations. The same consideration can be made regarding the WIOD dataset and the new revision of the OECD we have used, covering more recent years. One of the strongest point of our method is, in facts, its generality: any Input-Output tables database can be employed in the analysis. Another direction for the work would consist in pooling datasets for different years, in order to take into account the temporal dimensions. Despite constituting a somewhat unorthodox approach in econometrics, if anything because of its novelty, it might be regarded as an attempt to introduce a comparative static analysis to the data, a method widely employed in economics.

The interpretation of the archetypes was conducted considering their reaction to small supply-side shocks. One of the benefits of this framework is its modularity. Different approaches are possible without changing the substance of the statistical process. In this sense, it might be possible to either sophisticate and elaborate on the model we presented, for instance combining both supply and demand side shocks. Another possibility consists of analysing the archetypes by considering how the economies they describe would evolve over time, for instance by employing growth models. We opted for a shock measure because of the renewed interest of the scientific community on such approaches in recent times and because of its interpretability.

As far as the actual clustering of the data is concerned, the main limitation of the analysis consisted in the scarce direct interpretability of some clusters. In our view, this problem presents no immediate solution. A tentative solution might consist in comparing the result of this analysis with the results of other and different quantitative analysis that are ordinarily used in economic analysis and to couple them with some historical qualitative consideration to assess the explanatory power of our method.

In this work, we have found the Aitchison metric to yield more homogeneous and interpretable results than a standard Euclidean distance. This was confirmed by the Silhouette coefficients computed in all the cases under scrutiny. We deem the latter a satisfactory confirmation of our hypotheses and intuitions, at least when considering a k-medoids algorithm. There exists, however, a plethora of different clustering techniques that might be of help. Again, we believe that it might be of interest to consider a Bayesian approach to clustering, as it allows to estimate posterior distributions for the number of clusters: a tool that can be of great help in this particular instance. Another possibility would be that of considering algorithms that do not necessarily cluster all the points in the dataset, adapting them to a compositional framework. An example might be the DBSCAN algorithm ([28]). Otherwise, fuzzy clustering algorithms might be key for finding different structures in the same group of states.

A possible objection to the approach followed throughout this work might be the lack of homogeneity in the mathematical assumptions made during the various phases of the analysis, as if assumptions were cherry-picked depending on contingent needs. While this is true, the main design idea behind the analysis was that of trying to cluster data that do not inherently either present an immediate or an intuitive interpretation by following what seemed reasonable. The flexibility that stems from the relaxation of the homogeneity of the mathematical framework was fundamental to shed some light on a problem otherwise of difficult solution. Statistics was applied as a tool, notably given the novelty of this particular approach to this particular problem. This is especially true if we consider that in this work we do not aim at substituting entirely consolidated and widely employed methodologies used in economic analysis, but rather to construct an auxiliary approach.

To conclude, given the computational cost of NMF and Cross-Validation, it might be interesting to rewrite the code in a more efficient language, such as C++. This might be done expanding already existing tools, such as ([25]).

BIBLIOGRAPHY

[1] Daron Acemoglu, Ufuk Akcigit and William Kerr. *Networks and the Macroeconomy: An Empirical Exploration*. Tech. rep. 2015 (cit. on p. 18).

[2] K. J. Arrow and G. Debreu. 'Existence of an Equilibrium for a Competitive Economy'. In: *Econometrica* 22.3 (1954), p. 265 (cit. on p. 8).

[3] S. Baldone. *Produzione e distribuzione del reddito*. il Mulino, 1976 (cit. on p. 9).

[4] Michael W. Berry, Murray Browne, Amy N. Langville, V. Paul Pauca and Robert J. Plemmons. 'Algorithms and applications for approximate nonnegative matrix factorization'. In: *Computational Statistics & Data Analysis* 52.1 (2007), pp. 155–173 (cit. on pp. 26, 28).

[5] D. Billheimer, P. Guttorp and W. F. Fagan. 'Statistical Interpretation of Species Composition'. In: *Journal of the American Statistical Association* 96.456 (2001), pp. 1205–1214 (cit. on p. 38).

[6] J. M. Bioucas-Dias and J. M. P. Nascimento. 'Estimation of signal subspace on hyperspectral data'. In: *Image and Signal Processing for Remote Sensing XI*. Ed. by Lorenzo Bruzzone. SPIE, 2005 (cit. on p. 31).

[7] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer New York, 23rd Aug. 2016. 760 pp. ISBN: 9781493938438 (cit. on p. 31).

[8] Victor Bisot, Romain Serizel, Slim Essid and Gael Richard. 'Acoustic scene classification with matrix factorization for unsupervised feature learning'. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016 (cit. on p. 26).

[9] Fischer Black. *Business Cycles and Equilibrium*. John Wiley and Sons, Inc, 3rd Dec. 2009. 197 pp. ISBN: 0470499176 (cit. on p. 21).

[10] Florian Blöchl, Fabian J. Theis, Fernando Vega-Redondo and Eric O'N. Fisher. 'Vertex centralities in input-output networks reveal the structure of modern economies'. In: *Physical Review E* 83.4 (2011) (cit. on pp. 3, 20).

[11] Béla Bollobás. *Random Graphs*. Cambridge University Press, 2001 (cit. on pp. 21, 22).

[12] Stephen P. Borgatti. 'Centrality and network flow'. In: *Social Networks* 27.1 (2005), pp. 55–71 (cit. on p. 21).

[13] C. Boutsidis and E. Gallopoulos. 'SVD-based initialization: A head start for nonnegative matrix factorization'. In: *Pattern Recognition* 41.4 (2008), pp. 1350–1362 (cit. on p. 30).

[14] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer-Verlag GmbH, 10th Nov. 2010. ISBN: 0387709134 (cit. on p. 38).

[15] J.-P. Brunet, P. Tamayo, T. R. Golub and J. P. Mesirov. 'Metagenes and molecular pattern discovery using matrix factorization'. In: *Proceedings of the National Academy of Sciences* 101.12 (2004), pp. 4164–4169 (cit. on p. 30).

[16] G. Caldarelli and A. Chessa. *Data Science and Complex Networks*. Oxford University Press, 29th Sept. 2016. 144 pp. ISBN: 9780199639601 (cit. on p. 14).

[17] V. M. Carvalho and A. Thabaz-Salehi. 'Production Networks: A Primer'. In: *The Annual Review of Economics* (2018) (cit. on pp. 14, 18).

[18] Vasco M. Carvalho. 'From Micro to Macro via Production Networks'. In: *Journal of Economic Perspectives* 28.4 (2014), pp. 23–48 (cit. on p. 14).

[19] Felix Chayes. *Ratio Correlation*. University of Chicago Press, 1st Sept. 1971. 108 pp. ISBN: 0226102203 (cit. on p. 35).

[20] M. Chu, F. Diele, R. Plemmons and S. Ragni. 'Optimality, computation and interpretation of nonnegative matrix factorizations'. In: *SIAM Journal of Matrix Analysis and Applications* (2004) (cit. on p. 26).

[21] A. Cichocki and R. Zdunek. 'Non-negative Matrix Factorization with Quasi-Newton Optimization'. In: *Artificial Intelligence and Soft Computing - ICAISC 2006*. Ed. by Leszek Rutkowski, Ryszard Tadeusiewicz, Lotfi A. Zadeh and Jacek M. Żurada. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 870–879. ISBN: 978-3-540-35750-6 (cit. on p. 29).

[22] A. Cichocki, R. Zdunek and S. Amari. 'Hierarchical ALS Algorithms for Nonnegative Matrix and 3D Tensor Factorization'. In: *Independent Component Analysis and Signal Separation*. Ed. by Mike E. Davies, Christopher J. James, Samer A. Abdallah and Mark D. Plumbley. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 169–176. ISBN: 978-3-540-74494-8 (cit. on p. 29).

[23] A. Cichocki, R. Zdunek and S. Amari. 'Nonnegative Matrix and Tensor Factorization [Lecture Notes]'. In: *IEEE Signal Processing Magazine* 25.1 (2008), pp. 142–145 (cit. on p. 26).

[24] A. Cichocki, R. Zdunek, A. H. Phan and S. Amari. *Nonnegative Matrix and Tensor Factorizations*. John Wiley & Sons, Ltd, 2009 (cit. on p. 28).

[25] Ryan R. Curtin, Marcus Edel, Mikhail Lozhnikov, Yannis Mentekidis, Sumedh Ghaisas and Shangtong Zhang. 'mlpack 3: a fast, flexible machine learning library'. In: *Journal of Open Source Software* 3 (26 2018), p. 726 (cit. on p. 130).

[26] Margaret E. Daube-Witherspoon and Gerd Muehllehner. 'An Iterative Image Space Reconstruction Algorthm Suitable for Volume ECT'. In: *IEEE Transactions on Medical Imaging* 5.2 (1986), pp. 61–66 (cit. on p. 28).

[27] Chris Ding, Tao Li and Michael I. Jordan. 'Nonnegative Matrix Factorization for Combinatorial Optimization: Spectral Clustering, Graph Matching, and Clique Finding'. In: *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008 (cit. on p. 26).

[28] M. Ester, H. Kriegel, J. Sander and X. Xu. 'A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise'. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon: AAAI Press, 1996, pp. 226–231 (cit. on p. 130).

[29] Attila Frigyesi and Mattias Höglund. 'Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes.' In: *Cancer informatics* 6 (2008), pp. 275–292. ISSN: 1176-9351 (cit. on pp. 26, 30).

[30] T. Fujimoto and R. R. Ranade. 'Two characterizations of inverse-positive matrices: the Hawkins-Simon condition and the Le Chatelier-Braun principle.' eng. In: *The Electronic Journal of Linear Algebra* 11 (2004), pp. 59–65 (cit. on p. 14).

[31] N. Gillis and F. Glineur. 'Accelerated Multiplicative Updates and Hierarchical ALS Algorithms for Nonnegative Matrix Factorization'. In: *Neural Computation* 24.4 (2012), pp. 1085–1105 (cit. on p. 29).

[32] Nicolas Gillis. 'Nonnegative matrix factorization : complexity, algorithms and applications'. PhD thesis. Université Catholique de Louvain, 2011 (cit. on pp. 25, 27, 28).

[33] Nicolas Gillis. 'The Why and How of Nonnegative Matrix Factorization'. In: (21st Jan. 2014). arXiv: `http://arxiv.org/abs/1401.5226v2 [stat.ML]` (cit. on p. 31).

[34] G. H. Golub and C. F. Van Loan. *Matrix Computations*. J. Hopkins University Press, 7th Jan. 2013. ISBN: 1421407949 (cit. on p. 23).

[35] Alexander Gray. *The Development of Economic Doctrine*. John Wiley and Sons, Inc., 1931 (cit. on p. 10).

[36] L. Grippo and M. Sciandrone. 'On the convergence of the block nonlinear Gauss–Seidel method under convex constraints'. In: *Operations Research Letters* 26.3 (2000), pp. 127–136 (cit. on p. 29).

[37] N. Guan, D. Tao, Z. Luo and B. Yuan. 'NeNMF: An Optimal Gradient Method for Nonnegative Matrix Factorization'. In: *IEEE Transactions on Signal Processing* 60.6 (2012), pp. 2882–2898 (cit. on p. 29).

[38] Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag New York Inc., 9th Feb. 2009. 745 pp. ISBN: 978-0387848570 (cit. on pp. 25, 39).

[39] D. Hawkins and H. A. Simon. 'Note: Some Conditions of Macroeconomic Stability'. In: *Econometrica* 17.3/4 (1949), p. 245 (cit. on p. 14).

[40] Lucie N. Hutchins, Sean M. Murphy, Priyam Singh and Joel H. Graber. 'Position-dependent motif characterization using non-negative matrix factorization'. In: *Bioinformatics* 24.23 (2008), pp. 2684–2690 (cit. on p. 30).

[41] W. Isard. 'Interregional and Regional Input-Output Analysis: A Model of a Space-Economy'. In: *The Review of Economics and Statistics* 33.4 (1951), pp. 318–328. ISSN: 00346535, 15309142 (cit. on p. 9).

[42] Richard Johnson and Dean Wichern. *Applied Multivariate Statistical Analysis: Pearson New International Edition*. Sixth. Pearson Education Limited, 2014. 776 pp. ISBN: 1292024941 (cit. on p. 25).

[43] Eric Jones, Travis Oliphant, Pearu Peterson et al. *SciPy: Open source scientific tools for Python*. [Online; accessed <today>]. 2001 (cit. on p. 42).

[44] S. Kagawa, S. Okamoto, S. Suh, Y. Kondo and K. Nansai. 'Finding environmentally important industry clusters: Multiway cut approach using nonnegative matrix factorization'. In: *Social Networks* 35.3 (2013), pp. 423–438 (cit. on p. 3).

[45] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Interscience, 1990. ISBN: 978-0471878766 (cit. on p. 39).

[46] D. Kim, S. Sra and I. S. Dhillon. 'Fast Newton-type Methods for the Least Squares Nonnegative Matrix Approximation Problem'. In: *SIAM International Conference on Data Mining (SDM)*. 0. 2007 (cit. on p. 28).

[47] H. Kim and H. Park. 'Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis'. In: *Bioinformatics* 23.12 (2007), pp. 1495–1502 (cit. on p. 29).

[48] Daniel D. Lee and H. Sebastian Seung. 'Algorithms for Non-negative Matrix Factorization'. In: *Proceedings of the 13th International Conference on Neural Information Processing Systems*. NIPS'00. Denver, CO: MIT Press, 2000, pp. 535–541 (cit. on pp. 25, 26, 28).

[49] Daniel D. Lee and H. Sebastian Seung. 'Learning the parts of objects by non-negative matrix factorization'. In: *Nature* 401.6755 (1999), pp. 788–791 (cit. on pp. 25, 26).

[50] W. W. Leontief. 'Quantitative Input and Output Relations in the Economic Systems of the United States'. In: *The Review of Economics and Statistics* 18.3 (1936), p. 105 (cit. on pp. 3, 10).

[51] W. W. Leontief. *The Structure of american Economy, 1919-1939: An Empirical Application of Equilibrium Analysis*. Oxford University Press, 1951. ISBN: 978-0196311265 (cit. on p. 10).

[52] C. Lin. 'On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization'. In: *IEEE Transactions on Neural Networks* 18.6 (2007), pp. 1589–1596 (cit. on p. 28).

[53] C. Lin. 'Projected Gradient Methods for Nonnegative Matrix Factorization'. In: *Neural Computation* 19.10 (2007), pp. 2756–2779 (cit. on pp. 28, 29).

[54] L. Lovász. 'Random Walks On Graphs: A Survey'. In: *Combinatorics, Paul Erdos is Eighty* (1993) (cit. on p. 23).

[55] Robert E. Lucas. 'Understanding business cycles'. In: *Carnegie-Rochester Conference Series on Public Policy* 5 (1977), pp. 7–29 (cit. on p. 20).

[56] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert and K. Hornik. *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.7-1. Zurich, Switzerland, 2018 (cit. on p. 43).

[57] Karl Marx. *A History of Economic Theories*. The Langland Press, 1952 (cit. on p. 10).

[58] A. Mas-Colell, M. D. Whinston and J. R. Green. *Microeconomic Theory*. Oxford University Press, 1st Sept. 1995. 998 pp. ISBN: 9780195102680 (cit. on p. 5).

[59] James McNerney, Brian D. Fath and Gerald Silverberg. 'Network structure of inter-industry flows'. In: *Physica A: Statistical Mechanics and its Applications* 392.24 (2013), pp. 6427–6441 (cit. on p. 3).

[60] Ronald E. Miller and Peter D. Blair. *Input-Output Analysis: Foundations and Extensions*. Second. Cambridge University Press, 11th Aug. 2009. 750 pp. ISBN: 978-0-521-51713-3 (cit. on p. 12).

[61] Ronald E. Miller and Peter D. Blair. *Input-Output Analysis: Foundations and Extentions*. Prentice Hall, 1984. ISBN: 0-13-466715-8 (cit. on p. 9).

[62] H. Minc. *Non-negative Matrices*. Wiley, 1988. ISBN: 0-471-83966-3 (cit. on p. 14).

[63] Leon N. Moses. 'The Stability of Interregional Trading Patterns and Input-Output Analysis'. In: *The American Economic Review* 45.5 (1955), pp. 803–826. ISSN: 00028282 (cit. on p. 9).

[64] Mark Newman. *Networks: An Introduction*. Oxford University Press, 2010. ISBN: 978-0199206650 (cit. on pp. 14, 20).

[65] F. Nielsen and R. Bhatia, eds. *Matrix Information Geometry*. Springer, 7th Aug. 2012 (cit. on p. 39).

[66] J. Nocedal and S. Wright. *Numerical Optimization*. Springer-Verlag GmbH, 1st Sept. 2006. ISBN: 0387303030 (cit. on pp. 27, 30).

[67] A. B. Owen and P. O. Perry. 'Bi-cross-validation of the SVD and the nonnegative matrix factorization'. In: *The Annals of Applied Statistics* 3.2 (2009), pp. 564–594 (cit. on p. 31).

[68] Pentti Paatero and Unto Tapper. 'Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values'. In: *Environmetrics* 5.2 (1994), pp. 111–126 (cit. on p. 25).

[69] Vera Pawlowsky-Glahn, Juan J. Egozcue and Raimon Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, Ltd, 27th Mar. 2015. 272 pp. ISBN: 1118443063 (cit. on p. 35).

[70] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay. 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on pp. 26, 42).

[71] Almarin Phillips. 'The Tableau Economique as a Simple Leonief Model'. In: *Quarterly Journal of Economics* 69.1 (Feb. 1955), p. 137 (cit. on p. 10).

[72] A. M. Prieto and J. L. Zofío. 'Network DEA efficiency in input–output models: With an application to OECD countries'. In: *European Journal of Operational Research* 178.1 (2007), pp. 292–304 (cit. on p. 3).

[73] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019 (cit. on p. 43).

[74] Thijs ten Raa. *The Economics of Input-Output Analysis*. Cambridge University Press, 23rd Apr. 2014. 197 pp. ISBN: 978-0-521-84179-5 (cit. on pp. 9, 12).

[75] O. Rifki, H. Ono and S. Kagawa. 'The robustest clusters in the input–output networks: global $$\hbox {CO}_2$$ CO 2 emission clusters'. In: *Journal of Economic Structures* 6.1 (2017) (cit. on p. 3).

[76] P. J. Rousseeuw. 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis'. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65 (cit. on p. 42).

[77] W. Rudin. *Real and complex analysis*. Third. McGraw-Hill International Edition, 2001. ISBN: 978-0070542334 (cit. on p. 36).

[78] P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha and M. Hoffman. 'Static and Dynamic Source Separation Using Nonnegative Factorizations: A unified view'. In: *IEEE Signal Processing Magazine* 31.3 (2014), pp. 66–75 (cit. on p. 26).

[79] Steven Squires, Adam Prügel-Bennett and Mahesan Niranjan. 'Rank Selection in Nonnegative Matrix Factorization using Minimum Description Length'. In: *Neural Computation* 29.8 (2017), pp. 2164–2176 (cit. on p. 31).

[80] V. Y. F. Tan and C. Fevotte. 'Automatic Relevance Determination in Nonnegative Matrix Factorization with the β-Divergence'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.7 (2013), pp. 1592–1605 (cit. on p. 26).

[81] Vincent Y. F. Tan and Cédric Févotte. 'Automatic Relevance Determination in Nonnegative Matrix Factorization with the β-Divergence'. In: (25th Nov. 2011). arXiv: `http://arxiv.org/abs/1111.6085v3 [stat.ML]` (cit. on p. 31).

[82] M. Templ, K.Hron and P. Filzmoser. *robCompositions: an R-package for robust statistical analysis of compositional data*. John Wiley and Sons, 2011, pp. 341–355. ISBN: 978-0-470-71135-4 (cit. on p. 44).

[83] Hal R. Varian. *Microeconomic Analysis*. Norton & Company, 11th Mar. 1992. 42 pp. ISBN: 978-0393957358 (cit. on pp. 5, 9).

[84] J. von Neumann, O. Morgenstern and H. W. Kuhn. *Theory of Games and Economic Behavior*. Princeton University Press, 19th Mar. 2007. 776 pp. ISBN: 9780691130613 (cit. on p. 9).

[85] Léon Walras. *Elements of Pure Economics*. Homewood, 1954 (cit. on p. 10).

[86] S. van der Walt, S. C. Colbert and G. Varoquaux. 'The NumPy Array: A Structure for Efficient Numerical Computation'. In: *Computing in Science & Engineering* 13.2 (2011), pp. 22–30 (cit. on p. 42).

[87] H. Wang, C. Wang, H. Zheng, H. Feng, R. Guan and W. Long. 'Updating Input–Output Tables with Benchmark Table Series'. In: *Economic Systems Research* 27.3 (2015), pp. 287–305 (cit. on p. 3).

[88] Lizhi Xing, Jun Guan and Shan Wu. 'Measuring the impact of final demand on global production system based on Markov process'. In: *Physica A: Statistical Mechanics and its Applications* 502 (2018), pp. 148–163 (cit. on p. 3).

[89] M. J. Zaki and M. J. Wagner. *Data Mining and Analysis*. Cambridge University Press, 20th July 2018. 606 pp. ISBN: 0521766338 (cit. on p. 38).

[90] Junying Zhang, Le Wei, Xuerong Feng, Zhen Ma and Yue Wang. 'Pattern Expression Nonnegative Matrix Factorization: Algorithm and Applications to Blind Source Separation'. In: *Computational Intelligence and Neuroscience* 2008 (2008), pp. 1–10 (cit. on p. 26).

A

APPENDIX: SECTORS IN THE OECD DATABASE

We here report all the sectors of the OECD dataset Input-Output Tables.

1. Agriculture, hunting, forestry and fishing
2. Mining and quarrying
3. Food products, beverages and tobacco
4. Textiles, textile products, leather and footwear
5. Wood and products of wood and cork
6. Pulp, paper, paper products, printing and publishing
7. Coke, refined petroleum products and nuclear fuel
8. Chemicals and chemical products
9. Other non-metallic mineral products
10. Rubber and plastics products
11. Basic metals
12. Fabricated metal products
13. Machinery and equipment, nec.
14. Computer, Electronic and optical equipment
15. Electrical machinery and apparatus, nec
16. Motor vehicles, trailers and semi-trailers
17. Other transport equipment
18. Manufacturing nec; recycling
19. Electricity, gas and water supply
20. Construction
21. Wholesale and retail trade; repairs
22. Hotels and restaurants
23. Transport and storage
24. Post and telecommunications
25. Financial intermediation
26. Real estate activities
27. Renting of machinery and equipment
28. Computer and related activities
29. R&D and other business activities
30. Public administration and defence; compulsory social security
31. Education
32. Health and social work
33. Other community, social and personal services